Immersion dans des documents scientifiques et techniques : unités, modèles théoriques et processus

Vanessa Andreani

UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Sciences du Langage – Industries de la Langue**

Arrêté ministériel : 7 août 2006

Présentée par

Vanessa ANDRÉANI

Thèse dirigée par **Thomas LEBARBÉ**

préparée au sein du **Laboratoire LIDILEM (EA 609)**
dans **l'École Doctorale Langues, Littérature et Sciences Humaines**

Immersion dans des documents scientifiques et techniques : unités, modèles théoriques et processus

Thèse soutenue publiquement le **23 septembre 2011**,
devant le jury composé de :

M. Benoît HABERT
Professeur, ENS de Lyon (Président, Rapporteur)
Délégué Information et Connaissance auprès du SG d'EDF R&D
Mme Marie-Paule PÉRY-WOODLEY
Professeur, Université de Toulouse II Le Mirail (Rapporteur)
Mme Frédérique SEGOND
Directrice de Recherche du ParSem Research Group, XRCE (Membre)
M. Loïc MAISONNASSE
Directeur R&D, TKM (Référent Entreprise)
M. Thomas LEBARBÉ
Maître de Conférences Habilité, Université de Grenoble (Directeur)

École Doctorale ED50

Langues, Littérature et Sciences Humaines

Immersion dans des documents scientifiques et techniques : unités, modèles théoriques et processus

THÈSE

présentée et soutenue publiquement le 23 septembre 2011

pour l'obtention du

Doctorat de l'Université de Grenoble

(spécialité sciences du langage)

par

Vanessa Andréani

Composition du jury

Directeur :	Thomas Lebarbé	Maître de conférences HDR, Université de Grenoble
Rapporteurs :	Marie-Paule Péry-Woodley	Professeur, Université de Toulouse II Le Mirail
	Benoît Habert	Professeur, ENS de Lyon et Délégué Information & Connaissance auprès du Secrétaire Général d'EDF R&D
Examinateurs :	Frédérique Segond	Directrice de recherche du Parsing and Semantics research group, Xerox European Research Center
	Loïc Maisonnasse	Directeur R&D, TKM

Laboratoire de Linguistique et Didactique des Langues Étrangères et Maternelles — EA 609

Mis en page avec la classe thloria.

Remerciements

Je tiens à saluer toutes les personnes qui ont contribué, de près ou de loin, à la réalisation de ce projet porté pendant trois ans.

En particulier, je souhaite d'abord sincèrement remercier mon directeur de thèse, Thomas Lebarbé, qui m'a fait confiance, m'a aiguillée, m'a soutenue. Il a su me laisser une grande autonomie dans mes choix (cette formulation ne signifie pas qu'il n'a encadré ce travail que de loin, contrairement à l'usage de ce type d'expressions, paraît-il, dans les remerciements de thèse), tout en étant présent à chaque fois que j'en ai eu besoin. Il a su suggérer avec diplomatie et pédagogie ses idées, toute en écoutant et respectant les miennes. Le dialogue a toujours été une composante essentielle de nos nombreuses réunions.

Je remercie Marie-Paule Péry-Woodley pour m'avoir fait l'honneur d'accepter d'être rapporteur de cette thèse, ainsi que Benoît Habert, pour avoir consenti à jouer le double rôle de rapporteur et de président du jury. Je tiens également à remercier Frédérique Segond, pour avoir accepté d'être examinatrice. Merci enfin à Loïc Maisonnasse, qui non content d'être mon chef, d'avoir relu une bonne partie de mon travail, de m'avoir guidée ces dernières années, a accepté d'être examinateur.

Je souhaite également exprimer ma reconnaissance à Christophe Lecante, qui m'a donné l'opportunité de m'intégrer à la société TKM, entreprise vivante et dynamique. Il m'a fourni là un ample terrain de jeu, et m'a permis d'acquérir une expérience riche d'enseignements.

J'adresse aussi mes remerciements à Thibault Roy, pour avoir initié cette thèse en tant que Directeur R&D de TKM à l'époque, et pour m'avoir mis le pied à l'étrier lorsque je suis arrivée, avec beaucoup de gentillesse et de disponibilité.

Je salue chaleureusement mes collègues de TKM, et en particulier celles qui sont devenues mes copines dans la vraie vie : Delphine, Mélina, Agathe, Cécile, et *last but not least*, Alexia. M'installer à Grenoble valait la peine, rien que pour les rencontrer. J'adresse également le témoignage de ma gratitude éternelle au pôle R&D, pour l'aide apportée durant ces trois ans, les blagues pas toujours fines, et l'immense coup de main et soutien moral en fin de parcours : Alexia (encore elle), Loïc, Hervé, et Julien (*big up !*).

Merci aux stagiaires qui, de passage à TKM, ont croisé ma route pendant quelques mois. Un clin d'œil tout particulier à Cécile, Franck, Maïlys, et bien entendu, mon amie Silvia.

Merci également à Virginie Zampa, pour sa relecture minutieuse, ainsi qu'à mes co-galériennes du laboratoire LIDILEM, qui ont fait de mes journées à la fac de bons moments : Aïcha, Agnès et Auriane.

Je souhaite également manifester ma gratitude à mes « profs » du département de Sciences du langage de l'Université de Toulouse-Le Mirail, pour leur enthousiasme communicatif, leur ouverture d'esprit et leur investissement : ils m'ont transmis le goût de la linguistique, m'ont fait découvrir des domaines d'une grande richesse, et m'ont donné l'envie de poursuivre sur la voie de la recherche. Je pense en particulier à Marie-Paule Péry-Woodley, Ludovic Tanguy et Cécile Fabre pour le TAL, mais également à Patrick M'Pondo-Dicka et Régis Missire du côté de la sémiotique et de la sémantique textuelles. A tous, merci.

Pour leur attention quotidienne et leur confiance en moi inébranlable depuis toujours, je remercie, avec toute mon affection, mes parents. A mes petites sœurs et mon petit frère, Delphine, Laëtitia et Matthias (par ordre d'apparition) : merci d'exister, et souvent, de me faire rire !

Merci à Jessica, mon amie, d'être toujours présente dans ma vie, même de loin en loin, même de temps en temps, et d'être presque aussi enthousiaste que moi à l'idée de finir cette thèse.

Enfin, je souhaite remercier, de tout mon cœur, celui sans qui je ne serais certainement pas allée au bout de cette aventure, et qui a été mon soutien de tous les instants. Merci donc à Romain

pour m'avoir suivie jusqu'à Grenoble *alias* le « Grand Nord », pour avoir géré le quotidien ces derniers mois (années ?), pour sa main tendue, toujours, et quand il l'a fallu, pour ses énergiques mises au point. Je n'ai pas les mots pour lui exprimer toute ma gratitude et mon admiration, mais je sais qu'il saura lire entre les lignes.

De manière générale, je tiens à remercier tous ceux qui ont croisé mon chemin ces trois dernières années, et « avec qui mes rapports furent aussi divers qu'enrichissants ».

Dieu merci, quand on se contente de penser au lieu d'écrire, on a parfaitement le droit de sauter du coq à l'âne, sans s'attirer des remarques désobligeantes. J'aurais dû être dérouleur de pensées plutôt qu'écriveur de bouquins.

Pierre Desproges

Table des matières

Table des figures

Liste des tableaux

Résumé

Cette thèse aborde la problématique de l'accès à l'information scientifique et technique véhiculée par de grands ensembles documentaires. Pour permettre à l'utilisateur de trouver l'information qui lui est pertinente, nous avons œuvré à la définition d'un modèle répondant à l'exigence de souplesse de notre contexte applicatif industriel ; nous postulons pour cela la nécessité de segmenter l'information tirée des documents en plans ontologiques. Le modèle résultant permet une immersion documentaire, et ce grâce à trois types de processus complémentaires : des processus endogènes (exploitant le corpus pour analyser le corpus), exogènes (faisant appel à des ressources externes) et anthropogènes (dans lesquels les compétences de l'utilisateur sont considérées comme ressource) sont combinés. Tous concourent à l'attribution d'une place centrale à l'utilisateur dans le système, en tant qu'agent interprétant de l'information et concepteur de ses connaissances, dès lors qu'il est placé dans un contexte industriel ou spécialisé.

Mots-clés: traitement automatique des langues, ergonomie, représentation des connaissances, ontologies, entités nommées

Abstract

This thesis adresses the issue of accessing scientific and technical information conveyed by large sets of documents. To enable the user to find his own relevant information, we worked on a model meeting the requirement of flexibility imposed by our industrial application context ; to do so, we postulated the necessity of segmenting information from documents into ontological facets. The resulting model enables a documentary immersion, thanks to three types of complementary processes : endogenous processes (exploiting the corpus to analyze the corpus), exogenous processes (using external resources) and anthropogenous ones (in which the user's skills are considered as a resource) are combined. They all contribute to granting the user a fundamental role in the system, as an interpreting agent and as a knowledge creator, provided that he is placed in an industrial or specialised context.

Keywords: natural language processing, ergonomics, knowledge representation, ontologies, named entities

Introduction

La connaissance s'acquiert par l'expérience, tout le reste n'est que de l'information.
Albert Einstein

Les travaux que nous présentons dans ce document s'inscrivent dans le cadre d'un financement CIFRE (Convention Industrielle de Formation par la REcherche), coordonné par l'Association Nationale de la Recherche et de la Technologie. Elle est née d'une collaboration scientifique établie entre la société TKM et le laboratoire LIDILEM de l'Université Stendhal de Grenoble.

TKM est une société de conseil en innovation, et se place dans un contexte industriel très compétitif. Ses analystes fondent leur activité sur l'exploitation de documents électroniques scientifiques et techniques, dans lesquels ils cherchent l'information nécessaire à la réalisation de leurs missions. Or, trouver l'information pertinente dans des masses importantes de documents ne va pas de soi. Le travail que nous présentons est issu du besoin des analystes, relatif à l'optimisation de leur accès à l'information dans de grands ensembles documentaires.

A ce titre, la problématique qui en découle est à la fois appliquée et théorique, et s'exprime par la question suivante : comment accéder efficacement à l'information scientifique et technique véhiculée par de grands ensembles documentaires ?

Un accès efficace est d'abord un accès à l'information pertinente, de manière ciblée, par rapport à toutes les autres informations disponibles. La pertinence, notion relative à un contexte, est donc centrale. La dégager nécessite en premier lieu de modéliser l'information présente dans un ensemble documentaire. De la façon dont elle est modélisée dépend la manière dont l'utilisateur peut par la suite y accéder. Or, un milieu industriel compétitif implique des contraintes métier fortes, et parfois variées. Face à de telles contraintes, la souplesse est une condition *sine qua non* de la qualité de l'accès.

Pour atteindre cet objectif de flexibilité, nous suggérons de ne pas considérer l'information d'un ensemble documentaire donné comme une masse indivisible, mais au contraire de la scinder en autant de catégories raisonnées et exploitées par les utilisateurs. Cette distinction entre les différents types d'informations permet par la suite de les combiner, dans leur dimension calculatoire comme du point de vue de leur visualisation. L'une des composantes principales de ces combinaisons est l'intelligence de l'utilisateur, que nous plaçons au sein même des documents : il est immergé dans l'ensemble documentaire.

Cette immersion passe par la définition de plusieurs types de processus, opposés dans leur principe mais complémentaires dans leur usage. Des processus endogènes, fondés sur les informations fournies directement d'un ensemble documentaire à traiter, sont exploités pour tirer parti des connaissances intrinsèques aux textes. Des méthodes exogènes viennent les compléter, grâce à l'apport de ressources externes. Enfin, des processus anthropogènes modélisent l'intelligence de l'utilisateur comme une ressource supplémentaire, fondamentale là où les deux précédentes atteignent leurs limites.

Notre problématique de l'accès à l'information électronique en général, et de l'immersion documentaire en particulier, relève d'un grand nombre de champs scientifiques, de l'ingénierie des connaissances à la recherche d'information et à sa visualisation, en passant par le traitement automatique des langues ou l'ergonomie. Quel que soit le point de vue adopté, le besoin de modélisation est mis en avant.

L'ingénierie des connaissances modélise les informations, en tant qu'inscriptions numériques permettant d'accéder à la connaissance. Cette modélisation de l'information passe bien souvent par celle de son moyen d'expression, c'est-à-dire la langue : le traitement automatique des langues naturelles, la linguistique, la terminologie sont donc impliqués dans ces travaux.

L'ergonomie cognitive a également sa place dans ce contexte : elle interroge les interactions entre l'humain et la machine. En l'occurrence, elle travaille à dégager des modèles cognitifs liés à l'activité de recherche d'information informatisée, permettant d'optimiser les conditions de cette activité.

La recherche d'information en tant que discipline informatique fournit des concepts et méthodes utilisées pour l'accès documentaire, et y apporte des solutions numériques. Enfin, la visualisation d'information est l'étude de la représentation visuelle d'informations, généralement sur un support informatique. Elle oriente donc les choix relatifs à la représentation visuelle des informations modélisées d'un ensemble documentaire.

En somme, chacun de ces domaines apporte une partie des réponses à la problématique générale, dont le cœur est la modélisation : modélisation des informations, de leur moyen d'expression, de la perception qu'en a l'humain, de leur mode de visualisation. Nos travaux se placent donc dans une perspective fortement pluridisciplinaire, et bénéficient des apports de tout un ensemble de domaines de recherche.

Ce mémoire est divisé en deux parties. La première s'attache à décrire les unités, les modèles théoriques et les processus impliqués dans l'immersion documentaire telle que nous la concevons. La seconde partie de ce document se focalise particulièrement sur l'utilisation des différents types de processus que nous avons présentés, et dont la combinaison permet, aux différents niveaux de granularité, la conception de notre modèle d'immersion documentaire.

Dans la première partie, nous présentons au cours du premier chapitre la notion centrale de pertinence, fortement saillante dans la recherche d'information. Puis, dans le chapitre 2, nous proposons un mode d'accès à l'information pertinente par l'immersion documentaire, qui prend en compte des paramètres inhérents à une situation de recherche d'information du point de vue des humains impliqués dans une telle activité. Cette immersion intègre dans le système de recherche d'information l'utilisateur et ses compétences, et l'incite à se plonger dans les documents. Nous nous attachons au cours du chapitre 3 à décrire la modélisation des documents et des informations qu'ils contiennent, à travers une structure de représentation. Cette modélisation est orientée à la fois sur les documents et sur la façon dont ils sont exploités par les analystes. Le chapitre

4 porte sur les unités linguistiques véhiculant les connaissances contextuelles et thématiques potentielles modélisées dans la structure. Enfin, les trois types de processus complémentaires sont abordés dans le chapitre 5, pour chacun des niveaux de granularité que représentent l'immersion documentaire, au niveau global ; la structure de représentation, au niveau intermédiaire ; et enfin, les unités linguistiques impliquées, au niveau de granularité le plus fin.

La seconde partie est consacrée à l'examen des processus fondant la conception et la construction du système d'immersion en fonction des ressources utilisées. Dans le chapitre 6, les méthodes endogènes (le corpus pour analyser le corpus) sont décrites successivement du point de vue de chaque niveau de granularité. Le chapitre 7 suit le même découpage pour les ressources exogènes (le corpus est traité par des ressources externes). Enfin, le chapitre 8 est consacré selon le même déroulement à la description des processus anthropogènes (les compétences et les connaissances de l'utilisateur sont utilisées voire capitalisées en tant que ressources).

Nous concluons en synthétisant l'ensemble de nos réalisations, théoriques et appliquées. Nous appuyons particulièrement sur l'intérêt de combiner les trois types de processus s'appuyant sur les ressources endogènes, exogènes et anthropogènes. Nous soulignons également l'importance de la place de l'utilisateur dans un système d'immersion tel que nous le définissons : tous nos processus concourent en effet à lui accorder un rôle prégnant. Plus largement, cette place centrale est fondamentale dans les systèmes voués à fournir à l'humain de l'information à partir de documents numériques, dès lors que ce dernier est positionné dans un contexte industriel, ou tout au moins dans une situation où il est spécialiste des informations à traiter.

Première partie

L'immersion documentaire : unités, processus et modèles théoriques en présence

Chapitre 1

Chercher l'information pertinente : pourquoi, pour qui, pour quoi faire ?

Sommaire

Nos travaux portent sur la problématique de l'accès à l'information par l'utilisateur à travers des documents électroniques. Plus précisément, nous nous sommes penchée sur la recherche d'information (RI). Cette recherche est informatisée, au sens où elle est réalisée *via* l'outil informatique, qui fournit une assistance à l'utilisateur : il ne remplace pas intégralement l'humain.

Nous entendons *recherche d'information* comme la traduction de *Information Retrieval*, au sens de [Manning et al., 2008], en tant qu'activité consistant à « *finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)* ».

La recherche d'information fait appel à un certain nombre de notions ou d'objets, comme ceux d'information, d'utilisateur, de documents ou de requête. Cependant, la notion centrale, autour de laquelle s'articulent les processus impliqués dans le recherche d'information, est celle de pertinence. En effet, selon [van Rijsbergen, 1979] :

> « It is this notion [of relevance] which is at the centre of information retrieval. The purpose of an automatic retrieval strategy is to retrieve all the *relevant* documents at the same time retrieving as few of the *non-relevant* as possible. When the characterisation of a document is worked out, it should be such that when the document it represents is relevant to a query, it will enable the document to be retrieved in response to that query. »

Selon l'auteur, l'objectif d'un système de recherche d'information est de présenter des documents pertinents par rapport à une requête (*query*) de l'utilisateur. Le même point de vue est adopté par Saracevic [Saracevic, 1970], qui considère que la pertinence est une propriété du mécanisme interne du système d'information, et qu'elle est le résultat d'une correspondance entre les termes d'une requête et ceux qui ont été attachés aux documents. Ainsi, d'après l'auteur, tous les documents correctement retrouvés par le système par ce mécanisme seront, par définition, pertinents pour l'utilisateur. Puisque la pertinence est, selon [van Rijsbergen, 1979], fondamentale en recherche d'information, il semble logique de s'interroger en premier lieu sur ce que recouvre exactement cette notion de pertinence. Par là même, nous pouvons nous demander si cette pertinence d'un document existe exclusivement entre la « caractérisation » (*characterisation*) d'un document et la requête saisie par l'utilisateur, et si ce qui est pertinent pour un système l'est forcément pour son utilisateur.

1.1 Définition de la pertinence

La pertinence est étudiée depuis plusieurs décennies par des disciplines comme les sciences du langage et la psychologie, en tant que notion impliquée dans la communication. [Grice, 1975] notamment, dans ses travaux en pragmatique, fonde l'une de ses célèbres maximes sur la pertinence. En effet, selon lui, tout acte de communication est fondé sur un principe de coopération

linguistique, qui permet à un échange verbal d'être cohérent. Afin d'atteindre cette cohérence, les interlocuteurs d'un échange doivent respecter quatre maximes fondamentales : la maxime de quantité exige que l'énoncé soit aussi informatif que nécessaire, mais pas plus que nécessaire. La maxime de qualité demande que le locuteur ne mente pas, et ne communique pas un contenu dont la véracité n'est pas certaine. La maxime de manière, quant à elle, exige que le locuteur soit clair et précis, et que le contenu de l'énoncé soit le moins ambigu possible et ordonné. Enfin, et c'est là la maxime qui nous intéresse particulièrement, la maxime de relation exige du locuteur qu'il soit pertinent : « Be relevant » [Grice, 1975].

Grice pose donc la pertinence comme un critère fondamental assurant une communication cohérente entre deux ou plusieurs interlocuteurs. Cependant, rien dans la théorie de l'auteur ne permet de caractériser précisément ce concept.

1.1.1 La théorie de la pertinence de Sperber et Wilson

Du côté de la psychologie, Sperber et Wilson, dans [Sperber & Wilson, 1986], ainsi que dans [Sperber & Wilson, 1995] puis [Wilson & Sperber, 2004], ont élaboré une théorie de la pertinence permettant de préciser et définir ce concept, et surtout de lui attribuer une dimension cognitive. De leur point de vue, la pertinence est avant tout affaire de traitement mental : les processus inférentiels humains seraient guidés par des considérations de pertinence [Van der Henst, 2002]. Selon le Principe Cognitif de Pertinence (*Cognitive Principle of Relevance*), qui sous-tend la théorie de Wilson et Sperber, « *human cognition tends to be geared to the maximisation of relevance* » [Wilson & Sperber, 2004], de manière à augmenter la connaissance le plus efficacement possible. La recherche de pertinence serait donc un trait fondamental de l'esprit humain, exploitable par les locuteurs en situation d'échange. De ce fait, tout énoncé déclenche des attentes de pertinence qui guident l'interlocuteur pour déterminer le sens que le locuteur souhaite communiquer.

La pertinence est donc une propriété potentielle de tout énoncé ou phénomène observable, mais aussi de toute pensée, souvenir et conclusion d'inférence. Du point de vue de la théorie de la pertinence, « *any external stimulus or internal representation which provides an input to cognitive processes may be relevant to an individual at some time* » [Wilson & Sperber, 2004]. Ainsi, tout stimulus (ou pensée, représentation mentale, etc.) est potentiellement pertinent pour une situation donnée, c'est-à-dire pour un individu donné, et à un moment particulier. Le contexte semble donc prépondérant pour la notion de pertinence. Alors que dans un contexte C, un énoncé E est pertinent pour l'individu I, ce même énoncé E n'est plus pertinent pour le même individu I dans un autre contexte C'. Nous faisons ici varier le contexte, mais le même phénomène peut être obtenu, selon [Wilson & Sperber, 2004], en faisant varier l'énoncé, l'individu, ou toute combinaison de ces trois paramètres.

Le contexte s'exprime en partie, pour Sperber et Wilson, dans le cadre du rapport existant

entre effets et efforts qui fixe le degré de pertinence d'une information. Les effets cognitifs - ou contextuels - d'une information sont produits par l'interaction entre une information ancienne, qui représente le contexte, et une information nouvelle, qui est le stimulus. L'implication contextuelle, le premier effet, est l'information déduite du stimulus (la nouvelle information) dans le contexte (l'information ancienne). Ainsi, cette information naît de la conjonction entre information nouvelle et contexte, et n'aurait pu être déduite de la seule information nouvelle, ni du contexte isolé. Le renforcement contextuel, deuxième type d'effet, est le renforcement d'une hypothèse grâce à l'information nouvelle. Enfin, le troisième effet est l'élimination d'une hypothèse, due à un conflit entre information nouvelle et information ancienne. Plus les effets sont importants, plus l'information est pertinente.

Les efforts, quant à eux, représentent le coût cognitif entraîné par le traitement d'une information. Plus l'effort est faible, plus l'information traitée est pertinente.

Pour résumer, et poser ces éléments en termes de règles de la théorie de la pertinence, [Wilson & Sperber, 2004] définissent ainsi la pertinence d'une entrée (*input*) pour un individu :

> « Relevance of an input to an individual :
>
> 1. Other things being equal, the greater the positive cognitive effects achieved by processing an input, the greater the relevance of the input to the individual at that time.
> 2. Other things being equal, the greater the processing effort expended, the lower the relevance of the input to the individual at that time. »

Un principe de moindre effort (*principle of least effort*, [Wilson & Sperber, 2004]) est donc exploité par l'individu pour maximiser la pertinence des énoncés qu'il rencontre. C'est pourquoi, au sein d'un ensemble d'énoncés disponibles, l'individu sélectionnera celui qui est susceptible d'être le plus pertinent pour lui, dans le contexte où il se trouve. Le degré de pertinence d'un énoncé est donc pour [Wilson & Sperber, 2004] plus comparatif que quantitatif. En fonction de la situation, un énoncé donné sera plus pertinent qu'un autre parce qu'il nécessitera moins d'efforts et/ou entraînera un effet positif plus important. Les auteurs citent l'exemple d'un individu demandant l'heure à un passant : si le passant voit à sa montre qu'il est 11h58, il peut choisir de donner l'heure exacte, à la minute près, ou bien d'arrondir à 12h, ou *midi*. Si l'individu qui a posé la question est devant la gare et doit prendre un train à 12h04, alors l'heure exacte serait l'information la plus pertinente, puisqu'il saurait alors qu'il lui reste exactement 6 minutes pour trouver et prendre son train : l'effet cognitif serait plus important que si le passant lui répondait *midi*. En revanche, s'il n'a pas besoin de prendre un train rapidement, *midi* peut être plus pertinent, dans la mesure où il n'a pas besoin de savoir l'heure à la minute près, et que dans ce cas, traiter l'information *11h58* lui coûtera un effort plus grand sans engendrer d'effet positif supplémentaire.

Ce rapport entre coût et bénéfice n'est pas sans rappeler les théories écologiques, et en particulier celle du linguiste Zipf, dans son ouvrage *Human behaviour and the principle of least effort - An introduction to human ecology* [Zipf, 1949]. Zipf a constaté que la longueur des mots tendait à être inversement proportionnelle à leur fréquence d'utilisation dans les textes, et en a déduit la loi du moindre effort. Les humains agiraient à tous les niveaux selon un principe d'économie, visant à limiter au minimum le coût de l'ensemble des processus cognitifs qu'ils mettent en place. Bien que les travaux de Zipf portent avant tout sur le langage du point de vue linguistique, il est bien là question de mesurer les efforts tout en en maximisant les bénéfices, tout comme dans le cadre de la théorie de la pertinence.

La notion de pertinence, du point de vue cognitif, est donc une notion relative, puisqu'un énoncé sera plus ou moins pertinent en fonction du contexte dans lequel il est produit ou transmis. C'est entre autres ce contexte qui déterminera l'effet positif et l'effort engendrés par le traitement de cet énoncé ou, plus largement, d'une information. Il s'ensuit que, si le contexte change, la pertinence d'un énoncé est elle aussi susceptible d'évoluer. Par ailleurs, si l'effet cognitif est fondamental, l'effort produit est lui aussi à prendre en compte dans le calcul et l'évaluation de la pertinence de l'information. Ces principes valent tout autant pour les documents dans le cadre d'une recherche d'information, puisqu'un document peut être considéré comme un énoncé.

1.1.2 La pertinence en recherche d'information

Dans le champ de la recherche d'information, nous l'avons vu au début de ce chapitre, la notion de pertinence est fondamentale. [van Rijsbergen, 1979] la présente comme l'adéquation d'une requête saisie par l'utilisateur d'un système de recherche d'information avec les descripteurs d'un document utilisés par le système. C'est ce que [Mizzaro, 1998] nomme *system relevance* (*pertinence du système*), ou ce que [Vickery, 1959b] ou [Vickery, 1959a], appelle *relevance to a subject*, que nous pourrions traduire approximativement par la *pertinence par rapport à un sujet*. Il s'agit là de ce que le système de recherche d'information juge pertinent. Vickery parle de la pertinence par rapport à un *sujet*, puisque lorsque les descripteurs d'un document et une requête saisie sont en adéquation, la correspondance se fait sur des éléments représentant le sujet traité par un document.

Or, cette vision de la pertinence est quelque peu limitée, et les chercheurs en recherche d'information ont relevé d'autres types de pertinence, qui prennent notamment en compte l'utilisateur, et plus seulement le système de recherche d'information et ses entrées. C'est ainsi qu'à côté de la pertinence par rapport à un sujet, Vickery recense également la pertinence pour l'utilisateur (*user relevance*), qui réfère à ce dont l'utilisateur a besoin lorsqu'il effectue une recherche.

[Mizzaro, 1997] établit une revue de la littérature recensant 160 articles consacrés à la notion de pertinence. Ce qui ressort, entre autres, de cette étude est la diversité des termes employés

lorsque les chercheurs doivent parler de pertinence. Cependant, un grand nombre d'entre eux distingue la pertinence par rapport au système de recherche d'information et celle qui est établie - ou non - par le jugement des utilisateurs. Par la suite, cet état de l'art lui a permis de proposer dans [Mizzaro, 1998] un concept précis et composite de la pertinence en recherche d'information, qui se décline alors en quatre dimensions.

La première dimension concerne les sources de l'information. Elle englobe le document, les descripteurs de ce document, et l'information, c'est-à-dire ce qui sera reçu ou créé par l'utilisateur à la lecture du document.

La deuxième dimension contient quant à elle la représentation du problème de l'utilisateur qui motive sa recherche d'information. Cette dimension se divise elle-même en ce que nous pourrions appeler des paliers, au nombre de quatre. Le palier le plus haut est celui du besoin d'information réel (BIR) de l'utilisateur. Vient ensuite le deuxième palier, c'est-à-dire le besoin d'information perçu (BIP) par l'utilisateur, qui est « *a representation (implicit in the mind of the user) of the problematic situation* » [Mizzaro, 1998]. Ce besoin perçu est donc une représentation mentale du besoin réel, et à ce titre, il en est différent. Le troisième palier est la formulation en une requête (*request*) en langue naturelle (LN) du besoin d'information perçu. Enfin, le quatrième et dernier palier est la formalisation en une requête (*query*), dans un langage formel cette fois, de la requête en langue naturelle.

A ce stade, la combinaison des différents éléments de chacune des deux dimensions permet de caractériser la pertinence dans le cadre d'une recherche d'information. Nous reproduisons dans la figure 1.1 l'espace bidimensionnel représenté dans [Mizzaro, 1998], sur lequel il situe chacune des pertinences potentielles.

L'adéquation entre les descripteurs d'un document et la requête formalisée et saisie par l'utilisateur, c'est-à-dire la requête formelle, correspond à ce que nous avons appelé, à la suite de [Mizzaro, 1998], la pertinence du système. D'un autre côté, l'accord entre l'information reçue par l'utilisateur et le BIR de ce dernier correspond à ce que [Vickery, 1959b] nomme la pertinence de l'utilisateur. Selon Mizzaro, la pertinence « idéale », celle qu'il faut chercher à atteindre, est cette pertinence de l'utilisateur, puisque c'est celle qui sera pour ce dernier la plus satisfaisante. Or, ces deux types de pertinence se situent aux antipodes l'un de l'autre. Les flèches sur la figure représentent le chemin à parcourir, à partir de n'importe quel type de pertinence, pour atteindre la pertinence idéale. De fait, la pertinence du système, souvent celle qui est pratiquée dans les systèmes de recherche d'information comme dans le cas de moteurs de recherche sur le web comme Google ou Yahoo !, est celle qui est la plus éloignée de l'état idéal à atteindre.

La troisième dimension de la pertinence est celle du temps : un document (ou ses descripteurs ou l'information reçue par l'utilisateur) peut être non pertinent pour une requête formelle (ou naturelle, ou pour le besoin d'information réel ou perçu de l'utilisateur) à un moment t et le devenir plus tard, et inversement. Cela peut être dû, par exemple, au changement du besoin

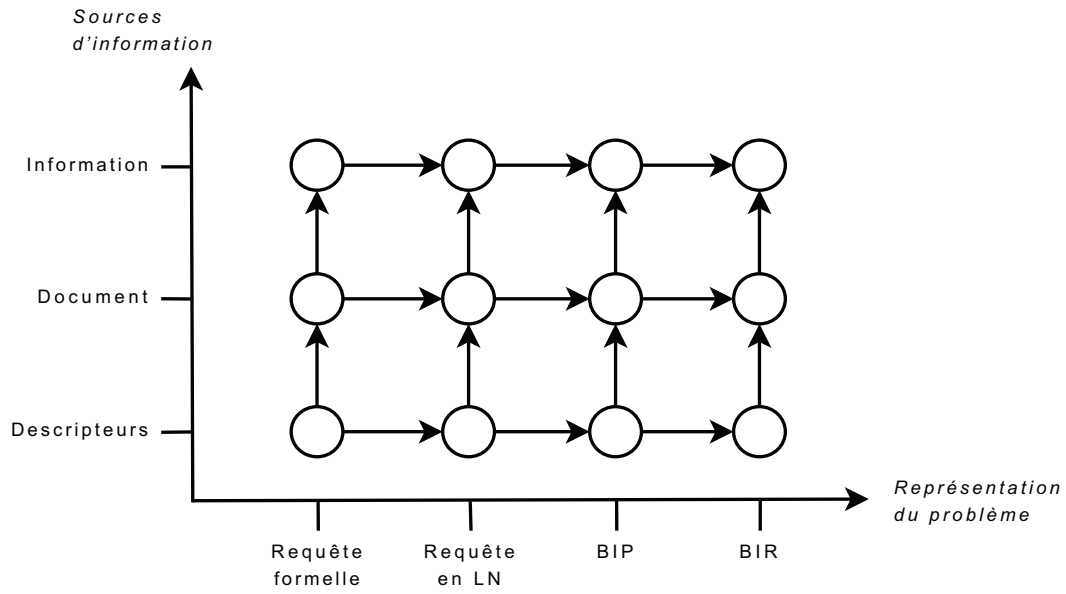

FIGURE 1.1 – La pertinence considérée comme point sur un espace bidimensionnel, selon [Mizzaro, 1998]

réel de l'utilisateur, ou au fait que le niveau de compréhension de ce dernier dans un domaine donné évolue. Cette dimension dynamique a donc une grande importance pour la recherche de la pertinence.

Enfin, **la quatrième dimension** concerne les composants dans lesquels peuvent se décomposer les éléments des deux premières dimensions. En effet, chacun d'entre eux, et par là chacune des pertinences, peuvent être découpés en trois composants distincts :

- Le sujet (*topic*) concerne le domaine thématique d'intérêt de l'utilisateur ;
- La tâche (*task*) concerne l'activité qu'effectuera l'utilisateur avec les documents trouvés, comme par exemple la rédaction d'un état de l'art ou la préparation d'un cours ;
- Le contexte (*context*) inclut tout ce qui ne relève ni de la tâche ni du sujet, mais qui a une influence sur la manière dont la recherche d'information est réalisée et dont les résultats sont évalués. Par exemple, le caractère nouveau d'une information, son caractère compréhensible, ou les paramètres de la situation concrète dans laquelle la recherche s'intègre, etc.

Ainsi, lorsque les descripteurs d'un document sont en adéquation avec une requête formelle par exemple, ils le sont pour un ou plusieurs des composants énumérés ci-dessus. La pertinence la plus complète, et la plus souhaitable, est en toute logique la pertinence de l'utilisateur, c'est-à-dire l'adéquation entre l'information reçue et le besoin d'information réel, et ce pour les trois composants : sujet, tâche et contexte.

La pertinence par rapport au sujet est donc loin d'être la seule qui importe dans une activité

de recherche d'information. Nous pourrions au contraire la qualifier de minimale. Il est primordial, d'après [Mizzaro, 1998], et sans pour autant délaisser le sujet, de s'intéresser à ce qu'il nomme la tâche et le contexte qui entourent la recherche d'information d'un utilisateur.

Bien que la théorie de la pertinence de [Sperber & Wilson, 1986] ne soit à aucun moment citée chez [Mizzaro, 1998], nous pouvons établir certaines correspondances entre les deux. Dans les deux cas, ce que les différents auteurs nomment pertinence est un effet cognitif positif sur l'individu, spécifié comme utilisateur d'un système de recherche d'information chez Mizzaro. En effet, ce dernier met en avant la nécessité de tendre vers une pertinence « idéale », qui ne prend pas seulement en compte la pertinence du système, mais celle de l'utilisateur, qui satisfera son besoin réel d'information. Il formalise partiellement ces effets positifs en fonction du contexte, puisqu'il considère le caractère nouveau ou ancien d'une information comme un élément de la pertinence : si l'information est nouvelle, elle est plus pertinente qu'une ancienne, ce qui revient à dire qu'elle a un effet cognitif positif plus important. De même, le caractère compréhensible ou non d'un document renvoie aux efforts cognitifs à produire pour l'utilisateur : un document incompréhensible pour un utilisateur lui demande trop d'efforts, et est donc un document non pertinent.

Ce que Mizzaro nomme la tâche est en fait inclus dans le contexte de Sperber et Wilson : la pertinence d'un énoncé est relative à la situation dans laquelle se trouve l'interlocuteur (*hearer*), à ce qu'il est en train d'accomplir lorsqu'il reçoit l'information (voir l'exemple du passant à qui un individu demande l'heure, en 1.1.1 page 11).

Ces travaux donnent de très intéressantes pistes pour la mise en place de systèmes de recherche d'information efficaces. Ils restent néanmoins limités en ce qui concerne les possibilités d'évaluation évoquées. En effet, il est difficile d'évaluer précisément la pertinence d'un énoncé, et donc du système de RI qui le fournit. Mizzaro propose cependant des pistes de travail, qui permettraient d'utiliser sa matrice de la pertinence. Selon lui, l'idéal de pertinence, c'est-à-dire la pertinence de l'utilisateur tenant compte de tous les composants de la quatrième dimension, est très complexe à mesurer, en particulier dans le cadre d'une recherche dans un système de RI. L'évaluation semble donc complexe, d'autant plus si les critères auxquels confronter le résultat de la recherche sont imprécis. En conséquence de quoi, [Mizzaro, 1998] suggère de tendre vers un compromis, sans préciser lequel. Il rappelle cependant que les calculs de précision et de rappel restent des mesures efficaces pour tous les types de pertinence qu'il recense.

De même, la théorie de la pertinence de [Wilson & Sperber, 2004] propose peu de méthodes d'évaluation des énoncés. En l'occurrence, l'objectif des auteurs est, en 2004, plus de valider par des expériences en psychologie cognitive la théorie elle-même que d'évaluer des énoncés produits en fonction de leur degré de pertinence.

Cette notion est donc complexe, même dans le contexte de la recherche d'information informatisée, et dépasse largement du cadre que lui donne [van Rijsbergen, 1979]. Elle peut avant

tout être considérée comme une notion relevant de la psychologie, qui implique des processus cognitifs et rentre en jeu dans la communication humaine. Elle est une propriété potentielle de tout énoncé, et par extension, de tout document. Ainsi, dans le cadre de la recherche d'information, bien plus qu'une correspondance entre descripteurs d'un document et requête saisie par l'utilisateur, elle est la conjonction d'un ensemble de paramètres, dépendant de l'utilisateur, de son activité, etc. Ces paramètres doivent donc être pris en compte pour établir le caractère pertinent d'un document, qui est relatif à plusieurs éléments, et en particulier aux composants de [Mizzaro, 1998], à savoir le sujet, la tâche et le contexte. Il est à noter que d'autres travaux, tels que ceux de [Pirolli & Card, 1999] par exemple, adoptent peu ou prou le même point de vue. Pour ces derniers, la pertinence est relative au contexte, à la tâche et à l'individu. Elle n'existe donc pas dans l'absolu, et un document pertinent dans une situation donnée peut ne pas l'être dans une autre. Définir cette situation, à travers ces composants et les autres dimensions de Mizzaro, est donc la première étape pour la mise en place d'un système de recherche d'information efficace. Dans ce qui suit, nous présentons donc brièvement ces trois composants dans notre cadre applicatif.

1.2 Les composants de la pertinence pour un cadre applicatif spécifique : la recherche d'information pour le conseil en innovation

Nous avons vu, en 1.1.1 page 11, que la pertinence d'un énoncé, ou d'un document, est relative à la situation dans laquelle l'information qu'il véhicule est transmise. Or, un système de RI efficace doit être capable de fournir à son utilisateur de l'information pertinente. En conséquence de quoi, le paramètre situationnel doit être pris en considération dans le fonctionnement d'un tel système. Les travaux de [Mizzaro, 1998] contribuent à cette prise en compte, en fournissant des clés pour la formalisation, au moins partielle, de la situation de recherche. C'est pourquoi nous nous fondons sur ces clés, et particulièrement sur la dimension des composants de la pertinence, à savoir le sujet, la tâche et le contexte de l'activité de recherche d'information sur notre terrain applicatif, et ce pour deux autres dimensions de la pertinence : les sources d'information et la représentation du problème de l'utilisateur.

Nos travaux se placent dans le cadre d'une thèse CIFRE, dont le terrain applicatif, la société TKM, est une jeune entreprise innovante de conseil en stratégie de l'innovation. Rappelons que notre objectif est d'élaborer un modèle qui fournisse aux analystes de la société des moyens d'accès à l'information plus performants. Il s'agit donc de concevoir un modèle pour un nouveau système de recherche d'information au sein de documents scientifiques et techniques.

1.2.1 La tâche

Les analystes de TKM sont chargés de réaliser les études qui permettent à la société de fournir des conseils en stratégie de l'innovation. Des acteurs de l'innovation, tels que des jeunes entreprises innovantes, des universités, des incubateurs ou des instituts de recherche privés par exemple peuvent faire appel à la société. Cette dernière propose des types d'études variés, classés en trois grandes catégories :

- **Stratégie en propriété industrielle :** dans le cadre de ces études, les analystes se focalisent particulièrement sur l'aspect *propriété industrielle* de l'innovation. Dans ces cas-là, ils sont amenés à travailler sur les aspects proprement stratégiques présentés par un projet de brevet ou d'exploitation d'une technologie par exemple. Les types d'études correspondants sont : la valorisation d'un portefeuille de brevets, des conseils en liberté d'exploitation, la constitution de *patent landscapes* (traduisible par *cartographie de brevets*). Pour l'année 2010, ces études représentaient 7,37% du nombre total d'études réalisées par TKM.
- **Intelligence économique et technologique :** ici, les analystes se penchent plus sur les technologies elles-mêmes et sur les aspects qui peuvent leur être liés ; ils peuvent réaliser des états de l'art approfondis, des cartographies de compétences autour d'un domaine technologique ou scientifique donné, délivrer des conseils en gestion des connaissances, effectuer des veilles concurrentielles, ou encore fournir un appui stratégique au projet de R&D d'un client. En 2010, les analyses de cette catégorie représentaient 43,16% des études.
- **De l'innovation au marché :** pour ces études, les analystes se concentrent sur l'arrivée d'une innovation dans un marché concurrentiel, de manière à ce qu'elle soit la plus efficace possible en ciblant le marché pertinent. Cette classe couvre les études tactiques de marché, la recherche d'applications pour l'innovation du client, les études de concurrence, le montage de projets, et enfin, la recherche d'acteurs et de partenariats. En 2010, ces études couvrent 44,21% du nombre total d'études TKM [1].

Il est à noter que certains types d'études en englobent d'autres. Par exemple, une étude tactique de marché approfondie comportera le plus souvent un état de l'art poussé, une étude de concurrence, éventuellement une recherche d'applications secondaires, etc.

Ce que [Mizzaro, 1998] appelle la tâche, c'est-à-dire l'activité plus large que les analystes réalisent avec les documents trouvés lors d'une recherche d'information, n'est donc ici pas uniforme : d'un type de mission de conseil à un autre, elle varie fortement. Par conséquent, la pertinence effective d'une information fluctue avec elle. Deux types de missions de conseil ne prescrivent pas les mêmes objectifs, et n'impliquent pas les mêmes besoins. Les études de marché par exemple s'intéressent de près aux divers acteurs d'un domaine, alors que les états de l'art techniques font prévaloir les données techniques et scientifiques avant tout.

1. Les 5,26% restants concernent des analyses diverses ne rentrant pas dans l'une des trois catégories d'études énumérées ici.

Néanmoins, les documents dans lesquels se trouve l'information potentiellement pertinente, eux, sont restreints dans un périmètre défini par la méthode propre à TKM. Celle-ci, appliquée à toute analyse produite, distingue la société de la plupart des autres entreprises de conseil. Elle prescrit à la manière d'une procédure les tâches secondaires permettant d'effectuer la tâche principale, c'est-à-dire la réalisation d'une étude. Par là, elle contraint son mode d'exécution, et influence également la définition des types de documents potentiellement pertinents. C'est pourquoi, afin de contextualiser nos travaux, nous livrons en plus de la description globale de la tâche principale, celle des tâches secondaires qui en découlent, et qui conditionnent finalement l'activité des analystes.

Alors que la plupart des sociétés de conseil réalisent leurs analyses à partir d'entretiens avec des experts d'un domaine, TKM quant à elle fonde ses études et ses conclusions sur l'analyse de documents scientifiques et techniques, et plus précisément sur des articles scientifiques et des brevets d'invention. Peuvent également être exploitées, dans une moindre mesure, la presse économique, des bases de données d'entreprises telles que Kompass, ou des pages web. Cette méthode est appliquée dans une volonté d'exhaustivité et d'objectivité, qui n'est pas envisageable par le seul moyen d'entretiens avec des experts. Les articles scientifiques et brevets sont exploités majoritairement parce qu'ils offrent en outre une crédibilité forte : ils sont issus de sources officielles et fiables. Or, travailler à partir de ces documents nécessite en premier lieu d'y accéder, et plus précisément d'accéder aux documents pertinents pour une étude donnée. Pour cela les analystes, dans 98% des études selon les estimations des analystes seniors, procèdent donc à une recherche d'information, qui porte pour 90% d'entre elles sur des brevets et articles scientifiques.

Nous représentons sur la figure 1.2 page suivante la méthode d'analyse TKM telle que nous la formalisons graphiquement, qui prescrit les tâches secondaires nécessaires à la réalisation d'une mission impliquant l'exploitation d'articles scientifiques et de brevets.

Une première phase de recherche, ou plutôt de récupération, correspondant à une tâche secondaire, consiste à collecter de telles publications à partir de bases de données en ligne dédiées [2]. Celles-ci référencent plusieurs dizaines de milliers de titres (revues scientifiques, actes de conférences, etc.) pour les articles scientifiques. Les bases de brevets recensent quant à elles les brevets d'inventions déposés dans de nombreux offices dans le monde. Par exemple, un analyste peut avoir à réaliser un état de l'art technique sur les biocarburants de dernière génération. Dans ce cas, après définition précise de la mission avec le client, et des mots-clés qui lui sont rattachés, l'analyste commence par consulter ces bases de données de manière à récupérer les documents correspondant aux exigences de l'exécution de la mission (phase 1 sur la figure 1.2). Pour ce faire, il construit et projette sur la base une ou plusieurs *équations de recherche*, c'est-à-dire des re-

2. Des fournisseurs d'accès à des bases de données telles que Questel (`http://www.qpat.com/index.htm`) ou Thomson Innovation (`http://thomsonreuters.com/`) pour les brevets, ou Scopus (`http://www.scopus.com/home.url`) pour les articles scientifiques, proposent des abonnements pour accéder aux documents et surtout pour les télécharger, afin de les utiliser par la suite pour des recherches et/ou différents traitements.

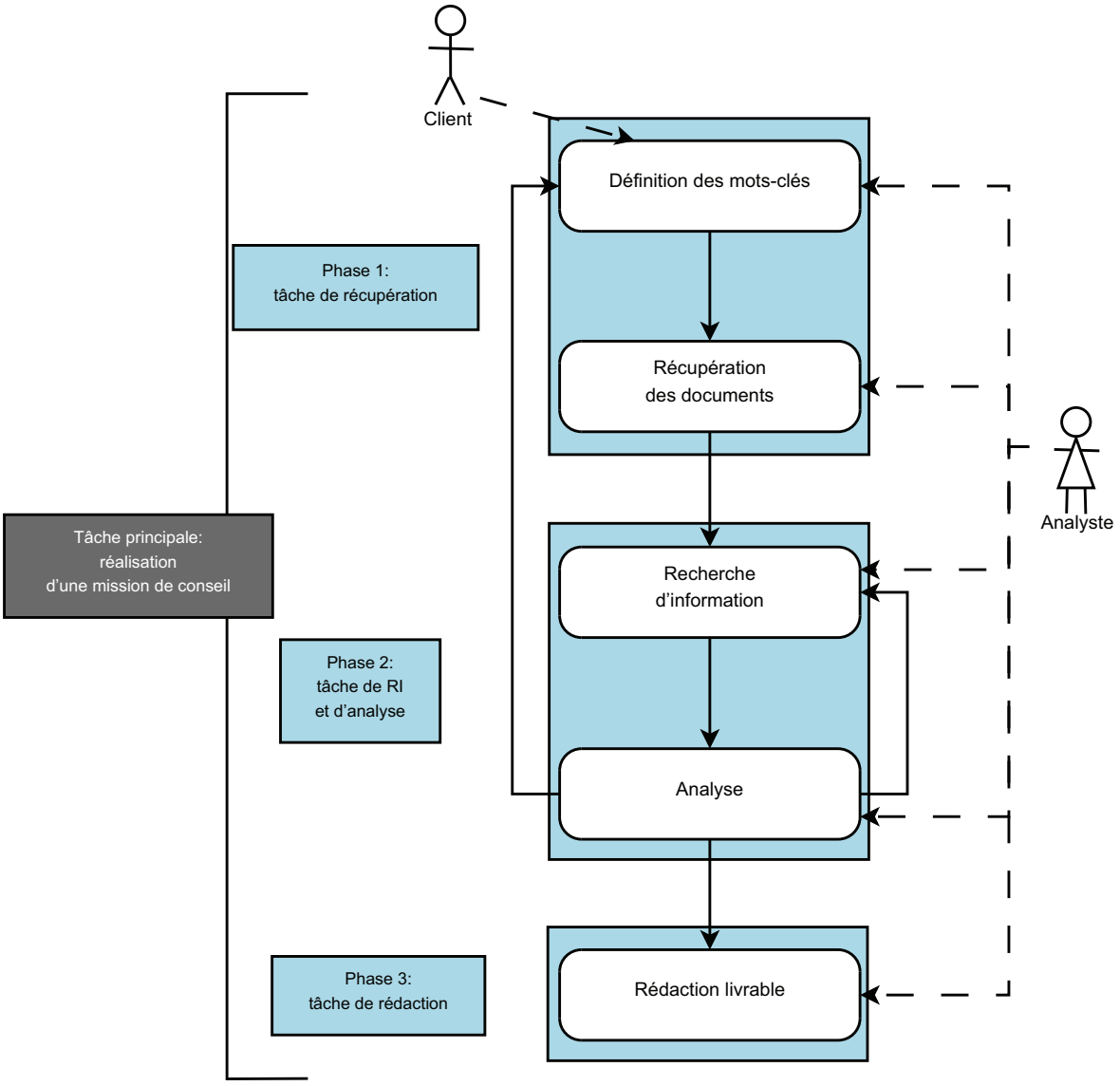

FIGURE 1.2 – Tâches secondaires prescrites par la méthode TKM et déroulement global de l'activité, pour la réalisation de la tâche principale

quêtes booléennes exprimées dans un langage propre à chaque base de données et fondées sur les mots-clés, *via* une interface spécifique. Ces équations peuvent atteindre une grande complexité, de manière à cibler le mieux possible l'ensemble des documents *a priori* pertinents.

La deuxième tâche secondaire (phase 2 sur la figure 1.2) correspond à la phase de recherche d'information à proprement parler couplée à l'analyse, à partir de la collection constituée à l'étape précédente. Elle est réalisée sur une plateforme conçue en interne par le service de recherche et développement de TKM, qui permet aux utilisateurs de trouver l'information dont ils ont besoin pour mener leur analyse.

Les documents récupérés sur les bases de données en ligne y sont intégrés en tant qu'éléments d'un ensemble documentaire propre à une mission. Diverses recherches peuvent alors être effectuées grâce aux fonctionnalités de la plateforme. Par exemple, une recherche peut être effectuée par date de publication des documents, ou encore par l'organisation au nom de laquelle un document est publié. Il existe également des possibilités de recherche par mots-clés, à l'aide d'opérateurs booléens. Ces recherches peuvent donner lieu à des visualisations par divers graphiques, tels que des camemberts, nuages de mots, etc. Par ailleurs, l'utilisateur peut également accéder, en parallèle, à la liste des documents correspondant aux critères de recherche. Cette liste est une liste plate, semblable à celle présentée par Google lors du renvoi des résultats. Cela permet à l'utilisateur de mener sa recherche, et lui fournit les informations nécessaires à son analyse. Cependant, la faiblesse d'un tel système réside dans les modes d'accès relativement limités, et à des croisements de critères de recherche rendant la lecture des résultats parfois difficile. Par ailleurs, les visions globales offertes par les graphes restent parfois insuffisantes, dans le sens où chacun des graphes est peu dynamique, et qu'aucun lien ne peut être établi entre deux ou plusieurs graphes.

La recherche d'information et l'analyse qui en résulte s'effectuent de manière itérative, avec plusieurs allers-retours de la recherche vers l'analyse de l'information trouvée. C'est pourquoi nous les avons intégrées au sein d'une même tâche secondaire. De plus, une analyse peut mener à une redéfinition des mots-clés, toujours en interaction avec le client, afin de cibler d'autres documents supposés pertinents.

Une fois les analyses menées à bien, les analystes rédigent dans une dernière phase le livrable à remettre au client, sous la forme d'un rapport écrit, ou *a minima*, d'une présentation (phase 3 sur la figure 1.2 page ci-contre).

La tâche dans laquelle s'inscrit la recherche d'information, et qui permet de définir partiellement la pertinence des informations, se caractérise donc par deux grands points :

– la diversité des tâches principales de conseil ;

– la constance de la méthode employée, et donc des types de tâches secondaires.

La pertinence d'une information varie donc beaucoup d'une tâche à l'autre, mais son type ainsi que son support sont circonscrits à l'intérieur d'un périmètre établi par la méthode TKM.

1.2.2 Le sujet

Le sujet est lui aussi, d'après [Mizzaro, 1998], partie prenante de la pertinence d'une information. Dans le cadre du conseil en stratégie de l'innovation, tous les domaines concernés par l'innovation sont, par définition, susceptibles d'être abordés. Dans les faits, bien que certains domaines soient plus représentés que d'autres dans les études effectuées dans la société, un grand nombre de disciplines et sous-disciplines scientifiques et techniques sont traités, telles que la pharmacologie, la biologie, la mécanique, la physique des matériaux, l'optique, les technologies de l'information et de la communication pour l'enseignement, etc.

Cependant, pour une étude particulière, le sujet est généralement relativement circonscrit. Par conséquent, un analyste a, la plupart du temps, un seul sujet à traiter à la fois, ou en tout cas un ensemble restreint de sujets proches. C'est pourquoi, à l'issue de la phase préliminaire de récupération de l'information, la collection de documents dans laquelle la recherche d'information est effectuée est, dans la grande majorité des cas, thématiquement homogène.

Toutefois, au sein même d'une collection de documents, il peut bien entendu exister plusieurs sous-sujets. Le traitement de ces variations autour d'un même sujet dominant est à prendre en considération pour la pertinence. Par exemple, une étude de positionnement pour les universités technologiques de Compiègne et de Troyes portait globalement sur « l'Homme équipé biologiquement et intellectuellement ». Elle a nécessité la prise en compte d'un certain nombre de domaines liés. Nous présentons en figure 1.3 les sous-thématiques traitées pour cette étude, telles qu'elles ont été schématisées par l'analyste dans le document constituant le livrable.

Les sous-sujets sont ici connexes, et l'ensemble reste cohérent étant donné le sujet de la mission globale. Néanmoins, dans les faits, il existe une différence notable entre, par exemple, la biomécanique et l'ingénierie des connaissances. Cette disparité est atténuée par le fait que dans ces cas-là, les différents sous-thèmes sont traités dans des sous-ensembles documentaires différents, chaque sous-ensemble devenant ainsi plus homogène. Pour finir, il est à noter qu'une étude aussi large du point de vue du thème reste marginale, et que la plupart du temps, les missions sont plus ciblées.

1.2.3 Le contexte

Selon [Mizzaro, 1998], le contexte concerne tout ce qui ne relève pas du sujet ni de la tâche, mais qui a malgré tout une influence sur la recherche d'information, et donc sur la pertinence d'un document ou d'une information. Bien que Mizzaro dise lui-même que ce contexte est difficile à formaliser, il donne quelques pistes de modélisation. Il cite en particulier les contraintes de temps qui entourent une recherche d'information. Or, le paramètre temporel est prépondérant pour les analystes de TKM. Ces derniers ont des contraintes de temps souvent très fortes, dues au caractère très compétitif du milieu du conseil.

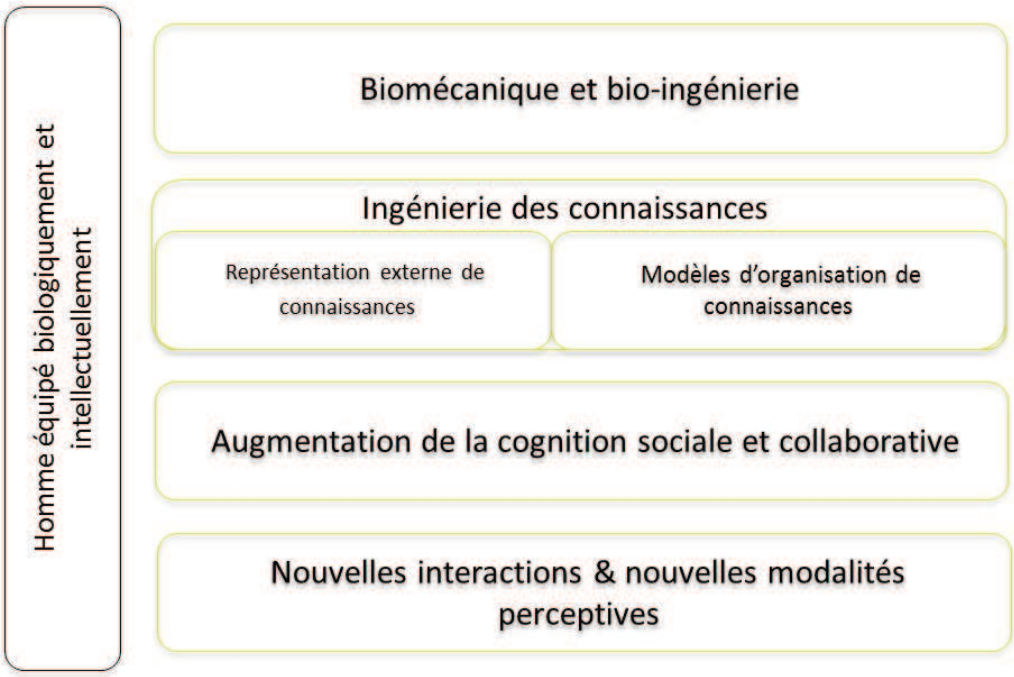

FIGURE 1.3 – Les différentes sous-thématiques abordées pour l'étude « Homme équipé biologiquement et intellectuellement » (figure tirée du livrable réalisé par TKM)

D'autre part, le caractère compréhensible d'un document est un paramètre cité par Mizzaro. Dans notre cadre applicatif, les utilisateurs sont des experts dans leur domaine, mais sont parfois amenés à travailler sur des champs de recherche qu'ils ne connaissent pas bien. Ainsi, si l'utilisateur traite pour une étude son sujet de prédilection, peu de documents seront pour lui peu compréhensibles, donc peu pertinents. En revanche, s'il doit se pencher sur un sujet qu'il ne maîtrise pas, ce paramètre rentre en jeu de manière prépondérante, et un document comportant beaucoup de détails techniques peut ne pas être pertinent au début d'une étude, lors de la phase d'appréhension globale du sujet ; dans cette situation, une article établissant un état de l'art général pourra s'avérer plus adéquat.

De même, le caractère nouveau ou ancien d'une information joue sur sa pertinence : si l'utilisateur a déjà pris connaissance d'une information, la lui présenter à nouveau a peu de chances d'être pertinent. Au contraire, lui présenter une information nouvelle, ou éventuellement, lui présenter une même information différemment, peut se révéler cognitivement intéressant.

Après ce rapide passage en revue des éléments rentrant dans les composants de la pertinence de [Mizzaro, 1998], nous avons déjà pu voir que des paramètres extérieurs au système d'information lui-même étaient à intégrer dans l'évaluation de la pertinence d'un document ou d'une

information. Ainsi, les descripteurs d'un document et la requête saisie par l'utilisateur sont loin de pouvoir rendre compte seuls de la pertinence d'une information pour les besoins d'un utilisateur.

A ce sujet, [Dervin & Nilan, 1986] notent un glissement de la prise en compte de la pertinence. L'information a longtemps été considérée comme objective, et les utilisateurs comme des processeurs d'entrées et de sorties. Or, cette perspective évolue aujourd'hui vers une conception de l'information et de la pertinence où les utilisateurs sont libres de créer ce qu'ils veulent à partir de systèmes et de situations. Pour ces auteurs, ce sont les utilisateurs qui construisent le sens. Pour respecter cette optique, l'ensemble des dimensions de la pertinence est à considérer dans la conception d'un système de recherche d'information : la source d'information, la représentation du problème par l'utilisateur et le caractère dynamique de la pertinence jouent un rôle qui ne doit pas être négligé. D'autre part, la notion de contexte chez [Wilson & Sperber, 2004], ou les notions conjointes de tâche, sujet et contexte chez [Mizzaro, 1998], sont fondamentales. Un système de recherche d'information, pour être efficace, doit par conséquent tenir compte de manière optimale de ces éléments ; pour cela, la conception doit considérer l'utilisateur dans son contexte, à travers ses caractéristiques, les tâches qu'il doit mener et l'activité qu'il effectue.

La pertinence est une notion complexe, relative, évolutive, et non réductible à une correspondance entre mots-clés. Dans notre cadre, et reprenant le titre de ce chapitre, il est donc crucial de se demander pour qui l'information doit être pertinente, et pour quoi faire. Si la caractérisation de documents par des mots-clés peut s'avérer nécessaire, elle n'est en aucun cas suffisante à la prise en compte des besoins d'utilisateurs pour l'accès à l'information scientifique et technique. La navigation au sein de collections de documents, ou ensembles documentaires, doit être mise en place avec l'objectif de respecter la diversité des besoins en présence, cette diversité devant s'exprimer autrement qu'à travers une liste de descripteurs rattachés aux documents.

Chapitre 2

L'accès à l'information pertinente par l'immersion documentaire

Sommaire

2.1 La navigation pour la recherche d'information

Alors qu'Internet devient l'un des principaux modes d'accès à l'information, que de plus en plus de données sont disponibles au format électronique, la recherche d'information (RI), en particulier sur le Web, passe de plus en plus par des moteurs de recherche tels que Google, Yahoo ! ou Voila. Les résultats correspondant à une requête prennent la plupart du temps la forme de listes de documents, ordonnés par ordre de « pertinence », cette dernière étant calculée à partir de critères pour le moins obscurs [3]. Quoi qu'il en soit, la liste de résultats est une liste énumérative. Paradoxalement - en apparence - ces moteurs de recherche, largement utilisés, ne satisfont pas forcément leurs utilisateurs. [Véronis, 2006], qui a établi un classement de cinq moteurs de recherche, montre qu'aucun d'entre eux n'atteint la moyenne en termes de satisfaction des utilisateurs, et plus précisément de « pertinence perçue » des résultats. Google et Yahoo ! arrivent en tête du classement avec une note de 2,3 sur 5. La note la plus basse est attribuée au moteur Voila, avec 1,2 sur 5. Véronis explique ces scores très bas par une proportion importante de résultats « hors thème », c'est-à-dire qui ne correspondent pas à la requête, et par la présence de liens commerciaux, pas toujours pertinents [4].

Pour notre part, nous posons l'hypothèse selon laquelle ces problèmes viennent également de la présentation énumérative des résultats, dans laquelle il peut être difficile de se déplacer, et de la faible prise en compte des besoins des utilisateurs pour l'accès à l'information. En effet, le public de ces moteurs de recherche, très large, peut avoir de multiples besoins, dans de multiples contextes. Google a certes mis en place quelques fonctionnalités, telles que la « roue magique », le classement chronologique, les suggestions de requête (voir les suggestions proposées par Google pour le mot-clé *andreani* en figure 2.1 page suivante) ou l'accès par types de documents définis à l'avance. Mais ces fonctionnalités, ainsi que leurs catégories, restent fondées sur des choix arbitraires des concepteurs et/ou sur des statistiques des requêtes les plus saisies, ne laissant que peu de place aux utilisateurs et aux dimensions de la pertinence qui s'appliquent pour eux lors d'une recherche (voir 1.1.2 page 13).

Ces résultats médiocres n'empêchent pourtant pas une utilisation massive, surtout en ce qui concerne Google qui enregistre plus de 91% du trafic web en décembre 2010 [Atinternet, 2011]. Puisque pour répondre au plus grand nombre, l'outil est mal adapté aux besoins individuels, il revient finalement à l'utilisateur de s'adapter à un outil qui ne le satisfait pas. En conséquence de quoi, l'utilisateur en vient à être formaté pour l'utilisation des moteurs de recherche tels que nous les connaissons aujourd'hui.

3. Voir à ce sujet le poisson d'avril de la société Google, sur le Pigeon Ranking : `http://www.google.com/technology/pigeonrank.html`

4. L'auteur, qui a ici pour objectif d'évaluer ces moteurs vis-à-vis d'un « public de base », ne s'attache pas à la formulation plus ou moins « bonne » des requêtes : il part du principe que le grand public doit avoir accès à ce qu'il cherche sans avoir besoin d'une expertise en recherche d'information.

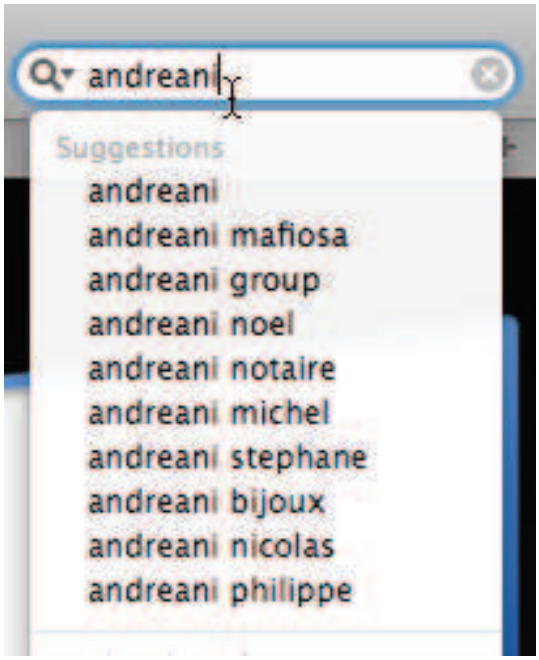

FIGURE 2.1 – Suggestions de requêtes fournies par Google pour le mot-clé *andreani*

La RI « courante » par moteurs de recherche, comme toute utilisation d'« artefacts cognitifs » [Merzeau, 2010], contraint donc le parcours de l'utilisateur (*ibid.*), là où il conviendrait de le libérer. En effet, le meilleur moyen pour l'utilisateur d'un système de RI d'obtenir son information pertinente est de lui permettre plus de liberté d'action, tout en lui faisant apparaître des liens de diverses natures entre documents de l'ensemble des résultats présentés. En somme, une navigation documentaire, comme une navigation heuristique, paraît plus pertinente que l'utilisation des moteurs de recherche tels qu'ils existent aujourd'hui. La navigation heuristique a lieu au sein d'informations ou documents présentés en réseau. L'utilisateur peut alors naviguer dans les informations, voire les explorer, à travers la sélection de nœuds et/ou de liens. Elle va donc au-delà d'une interrogation « classique » de moteurs de recherche, mais est aussi plus riche que la navigation par des liens hypertextes présents dans des documents. La navigation heuristique nous paraît donc la plus adaptée pour rendre à l'utilisateur une certaine liberté d'action.

Plusieurs disciplines s'intéressent à la navigation, chacune en fonction de son point de vue. L'informatique, dans la mesure où la navigation documentaire est électronique, et où la recherche d'information passe de plus en plus par des documents numériques, est en première ligne, en particulier du point de vue technique. Mais ce n'est pas le seul champ scientifique se penchant sur la question. Le traitement automatique des langues et les sciences du langage, l'ergonomie cognitive et l'interaction humain-machine, l'ingénierie des connaissances en tant que branche de l'intelligence artificielle et des sciences de l'information, entre autres, apportent tous une contribution importante. Il convient donc de s'intéresser aux apports et aux points de vue de

ces différentes sciences concernant la navigation, avant de définir les paramètres d'un système adapté à une situation professionnelle donnée.

Alors que l'ergonomie s'intéresse à la recherche d'information et à la navigation en tant qu'activité cognitive et interaction entre humain et machine, l'ingénierie des connaissances et les sciences de l'information se penchent sur leur aspect technique, en tant que support informatique pouvant faire appel à des connaissances, mais aussi sur les modèles qui sous-tendent de tels systèmes. Le TAL et la linguistique considèrent quant à eux la recherche d'information et la navigation comme une activité impliquant la langue en tant que système cohérent, matériau des documents mais aussi mode d'interrogation, en particulier dans la recherche ou l'extraction d'information textuelle. Par exemple, un grand nombre de travaux en TAL portent entre autres sur l'extension de requêtes d'utilisateurs [Bouillon et al., 2000], [Moreau & Claveau, 2006], l'indexation de termes complexes [Boulaknadel, 2008], l'intégration d'analyseurs syntaxiques [Roy, 2007], [Jacquemin & Zweigenbaum, 2000], l'indexation, l'analyse sémantique [Valette & Slodzian, 2008], [Bernhard & Ligozat, 2011], etc.

Nous verrons que les différents domaines de recherche impliqués dans le champ scientifique de la recherche d'information sont loin d'être hermétiques les uns aux autres. Nous nous attacherons ici à décrire quelques apports de l'ergonomie et de l'ingénierie des connaissances à la recherche d'information et à la navigation. Nous reviendrons plus particulièrement sur les apports de la linguistique et du TAL dans le chapitre 4.

2.2 La recherche d'information en tant qu'activité cognitive : le point de vue de l'ergonomie

La navigation est un moyen de réaliser une tâche de recherche d'information (RI). Elle implique une interaction entre un utilisateur humain et un dispositif informatique, qui doit permettre au premier de consulter des documents, quelle que soit leur forme, *via* le second.

Cette tâche de RI est une tâche secondaire, qui se positionne par rapport à une tâche principale dont la réalisation nécessite des connaissances. Classiquement, nous recherchons en effet de l'information dans un but plus large que la recherche elle-même : une recherche d'information s'effectue rarement pour elle-même, sans finalité et « dans le vide ». Les tâches principales à l'origine d'un besoin d'information sont extrêmement variées, et peuvent aller du besoin d'un numéro de téléphone pour fixer un rendez-vous avec un médecin à la préparation d'un voyage à l'étranger, en passant par la rédaction d'un document technique par exemple.

La tâche principale comme la tâche secondaire qu'est la recherche d'information relèvent bien souvent de ce que les psychologues et/ou ergonomes nomment la résolution de problèmes mal définis [Simon, 1973]. Un problème mal défini est un problème dont les données sont peu claires, partiellement déterminées par l'individu devant résoudre le problème, avec un but flou

[Falzon, 2005]. Il n'existe pas de solution unique à ce type de problème. Par exemple, un travail de conception de dispositifs techniques et organisationnels est un problème mal défini (*ibid.*), de même qu'un travail de thèse de doctorat. De ce point de vue, la recherche d'information est bien un problème mal défini, puisqu'il existe rarement une solution unique pour arriver à l'information recherchée, et permettant d'exécuter la tâche principale. Par ailleurs, dans bien des cas, l'utilisateur d'un système de recherche d'information ne sait pas avec exactitude ce qu'il cherche, et son but est donc flou. Il en va souvent de même pour la tâche principale ; dans le cas d'une tâche de rédaction d'un document technique ou scientifique par exemple, il n'y a pas de résultat unique à atteindre, mais tout au plus une représentation de ce qui est jugé satisfaisant vers laquelle tendre, en fonction de certains critères, tels que la rigueur scientifique, la qualité de la rédaction, etc. Quoi qu'il en soit, la solution à un « problème principal » passe souvent par la résolution d'un problème secondaire de recherche d'information. Cette dernière vise l'acquisition de connaissances complétant celles de l'individu réalisant la tâche principale. Nous appelons *connaissances* un ensemble de structures mentales cohérentes et non contradictoires, élaborées à partir du réel [von Glasersfeld, 1994].

Cette acquisition de connaissances se fait très souvent par le biais de sources externes, dans deux types de mémoires. Par *mémoire*, nous entendons la faculté ou le support permettant « [l']encodage, [le] stockage, et [la] récupération des représentations mentales » [5], c'est-à-dire des connaissances. Les mémoires externes artificielles sont des documents, quelle que soit leur forme, et les mémoires externes naturelles sont celles de tiers, auxquelles l'accès se fait par des entretiens, des interactions ou des dialogues. Dans le cadre des mémoires externes artificielles, la recherche d'information fait donc appel à des documents, qui viennent compléter les mémoires naturelles humaines [Tricot, 2003]. De fait, pour assister correctement la mémoire naturelle de l'utilisateur, un système de recherche d'information doit avoir été conçu en prenant en compte les besoins spécifiques de cet utilisateur dans son contexte, et en fonction du but rattaché à sa tâche principale.

Pour cela, il est important de considérer à la fois la tâche et l'activité de l'utilisateur, ou du groupe d'utilisateurs. Ces deux concepts issus de l'ergonomie sont fondamentaux dans cette discipline. L'ergonomie est en effet « la discipline scientifique qui s'occupe de la compréhension des interactions entre les hommes et les autres éléments d'un système » [IEA, 2010], traduit par [SELF, 2008]. Elle est aussi « la profession qui applique les théories, les principes, les données, et les méthodes pour concevoir dans le but d'optimiser le bien-être des hommes et la performance du système dans son ensemble » (*ibid.*). Or, pour rendre possible une conception optimisant les performances et le bien-être des hommes, l'ergonomie s'attache entre autres à décrire les tâches et activités qu'ils sont amenés à réaliser dans leur situation de travail. Puisque la recherche

5. Article de l'encyclopédie collaborative en ligne Wikipédia pour la vedette *mémoire* (`http://fr.wikipedia.org/wiki/M\%C3\%A9moire_\%28psychologie\%29`)

d'information fait appel en priorité à des processus cognitifs, c'est l'ergonomie cognitive plus précisément qui en fait l'étude. L'ergonomie cognitive « s'intéresse aux processus mentaux (tels que la perception, la mémoire, les raisonnements et les réponses motrices) influant sur l'interaction entre les hommes et les autres éléments des systèmes (e.g. étude de la charge de travail mental, de la prise de décision, de l'interaction homme-machine et de la fiabilité humaine) » (*ibid.*).

La machine interagissant avec l'humain dans une activité de recherche d'information est, dans notre contexte applicatif, un ordinateur, et ce sont les processus cognitifs en jeu lors de cette interaction qui font l'objet de l'étude ergonomique. Les informations tirées de cette étude doivent être considérées dans la conception d'un système de navigation, de manière à ce que ce dernier soit efficace.

Selon [Leplat & Hoc, 1983], qui utilisent les notions de tâche et d'activité, et surtout qui les distinguent clairement, « la tâche indique ce qui est à faire, l'activité ce qui se fait. La notion de tâche véhicule avec elle l'idée de prescription, sinon d'obligation. La notion d'activité renvoie elle, à ce qui est mis en jeu par le sujet pour exécuter ces prescriptions, pour remplir ces obligations ». Alors que la tâche fait référence à ce qui doit être fait, ce qui est en quelque sorte « demandé », l'activité concerne tout ce qui est effectivement réalisé pour exécuter la tâche. Même si ces deux notions sont très distinctes, elles sont intimement liées de fait : la tâche et ses contraintes et caractéristiques déterminent les besoins de l'individu pour sa réalisation, et influencent donc fortement leur activité.

De ce point de vue, le fait que la navigation soit une tâche secondaire qui s'intègre à une tâche principale est une caractéristique importante : elle implique un coût cognitif relativement lourd de la navigation, qui doit être menée de front avec la tâche principale. Un système de navigation s'intègre donc inévitablement à un contexte large dans lequel se déroule l'activité répondant à la tâche principale. Cela implique directement que l'efficacité d'un système de navigation est relative à son contexte d'utilisation [6].

2.2.1 L'étude de l'activité de navigation en ergonomie

D'après [Tricot, 2003], la première phase d'une recherche d'information en tant qu'activité, en quelque sorte une étape préliminaire, est la prise de conscience du besoin d'information. Pour cet auteur, il s'agit d'un besoin de réduction d'incertitude : celle-ci est la connaissance explicite,

6. Nous rappelons que des moteurs de recherche web tels que Google, Yahoo ! ou autres se distinguent de cette approche, puisque leurs utilisateurs potentiels s'inscrivent dans un public si large qu'il est impossible d'établir des contextes d'utilisation, même approximatifs. En conséquence de quoi, il revient très souvent à l'utilisateur de s'adapter à l'outil, et non l'inverse. La rançon du succès de ce type de moteurs de recherche, et l'ironie de cette situation, est que les utilisateurs sont en quelque sorte formatés par ces outils, alors même que certaines études de satisfaction indiquent que la majorité des utilisateurs ne sont pas satisfaits par les résultats obtenus (voir section 2.1 page 26), confirmant donc l'aspect contextuel d'une recherche d'information efficace. Voir au sujet du niveau d'(in)satisfaction des utilisateurs [Véronis, 2006].

ou prise de conscience, d'un manque de connaissances. En somme, l'individu effectuant une tâche principale doit avoir suffisamment de connaissances pour prendre conscience du fait qu'il n'en a pas assez pour réaliser cette tâche. Tricot classe les différents besoins d'information en six catégories générales, dont les trois premières ont une source interne à l'individu, et les trois dernières une source externe. Nous reproduisons ici cette typologie, en numérotant les types pour plus de lisibilité :

«

1. rechercher une connaissance que l'on n'a pas ;
2. rechercher une confirmation d'une connaissance que l'on a ;
3. rechercher une connaissance plus complète que celle que l'on a (mais aussi : un exemple, une illustration, un contre-exemple, etc.) ;
4. rechercher pour être conforme aux buts, aux contraintes, aux attentes de la situation ;
5. rechercher des indications sur la forme de la connaissance à utiliser dans la situation ;
6. rechercher parce que l'on a détecté un marqueur de pertinence dans la situation (ostension, mise en exergue visuelle, sonore, etc.).

»

La nature du besoin, qui peut recouper plusieurs catégories et mêler des besoins internes et externes, doit être prise en compte lors de la conception d'un système de navigation pour la recherche d'information. Cela est d'autant plus important que ce besoin est par la suite transformé par l'individu en une représentation mentale de son but de recherche. Cette représentation est divisée en une composante conceptuelle, revenant pour l'individu à se demander ce qu'il va chercher, et une composante procédurale, qui concerne la manière dont la recherche va s'effectuer. La structure du système de navigation rentre donc directement en compte dans la définition du but de la recherche, et plus précisément dans sa composante procédurale. [Tricot, 1993] définit le but de recherche d'information comme l'interaction entre la représentation mentale du besoin d'information, qui peut être précise ou floue, et la localisation de l'information pertinente, la cible, qui peut être unique et localisée ou multiple et distribuée. Par la combinaison de ces caractéristiques, il obtient quatre types de buts, représentés sur la figure 2.1.

Ce but n'est pas fixé une fois pour toutes au début de la recherche : il peut évoluer au cours de la navigation et au fil de la consultation des résultats. Ainsi, la gestion continue du but devient une composante majeure de l'activité de recherche d'information.

A partir de cette représentation du but de recherche, une requête est formulée, qui dépend donc d'une part du contenu recherché, et d'autre part des contraintes imposées par le système

Localisation de la cible	Représentation du besoin informationnel : Précise	Représentation du besoin informationnel : Floue
Unique, localisée	Chercher un renseignement	Explorer
Multiple, distribuée	Collecter	Butiner

TABLE 2.1 – Les quatre buts de recherche d'information d'après [Tricot, 1993]

d'information. En fonction du système choisi, il sera possible de formuler explicitement une requête, ou bien de consulter un ensemble de titres ou de mots-clés, comme c'est le cas pour les hypertextes ou les index, tables des matières, etc. Il est à mentionner que dans un grand nombre de systèmes de navigation, les deux accès sont possibles.

Un grand nombre de travaux ont tenté de modéliser l'activité en situation de recherche d'information. Les théories de la pertinence telles que [Mizzaro, 1998] permettent d'appréhender les caractéristiques du contexte de la recherche, mais pas vraiment l'activité en elle-même.

Des approches descriptives de la RI en séquence, ou encore cycliques, permettent d'aborder plus précisément cette activité. Parmi elles, [Rouet & Tricot, 1996] et [Rouet & Tricot, 1998] décrivent cette dernière en articulant la compréhension du contenu des documents d'une part et le processus de RI comme résolution de problème d'autre part [7]. L'activité est particularisée en fonction de la tâche et de l'environnement. Le cycle Evaluation - Sélection - Traitement (EST) est représenté dans la figure 2.2, adaptée de celle de [Rouet & Tricot, 1996].

Il est composé de trois modules principaux :

1. **l'évaluation** consiste à comparer la représentation du but avec l'état actuel de la solution. Lors de la première itération du cycle, l'état de la solution est nul, puisque c'est ce qui motive la recherche d'information ; par la suite, c'est la mesure de l'écart entre la représentation du but et la représentation du contenu du résultat de recherche qui est évalué. Lorsque les deux représentations sont suffisamment proches, la recherche d'information s'arrête ;
2. **la sélection** de l'information, ou plus précisément des catégories d'informations, passe par le calcul de la valeur d'intérêt de chacune de ces catégories : la ou les catégories d'informations qui paraissent les plus pertinentes sont sélectionnées ;
3. **le traitement** de l'information sélectionnée consiste pour l'utilisateur à examiner une unité

7. Voir aussi [Rouet & Tricot, 1995].

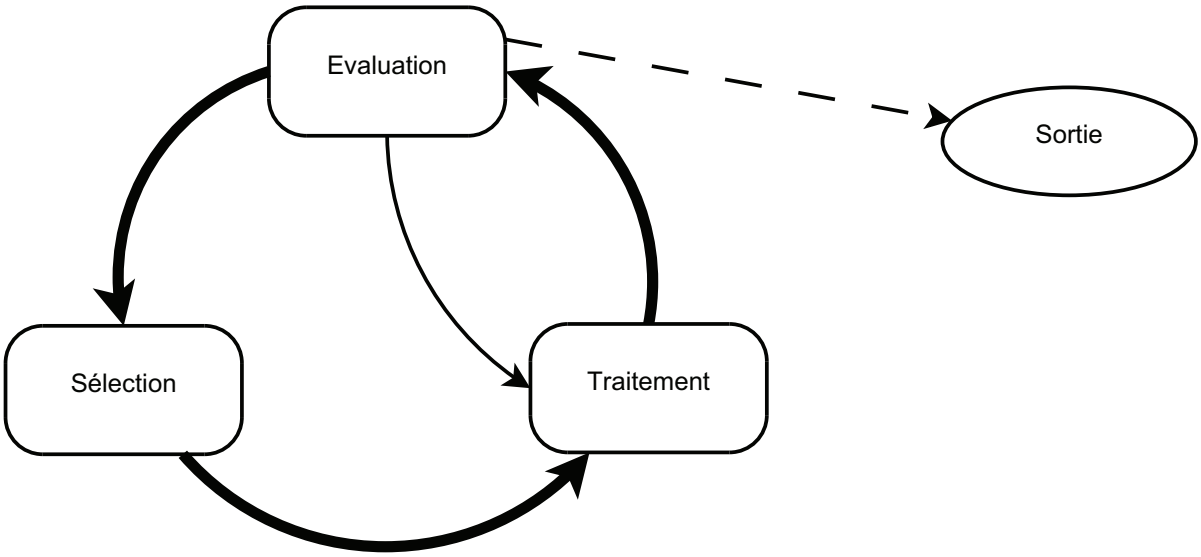

FIGURE 2.2 – Représentation du cycle EST de recherche d'information adaptée d'après [Rouet & Tricot, 1996]

de contenu : il doit la comprendre, puis évaluer la pertinence de l'information par rapport à son objectif.

La sélection d'un document se fonde sur le traitement d'éléments limités du document, comme le titre, les mots-clés ou des indicateurs para-linguistiques. Le document choisi peut alors être traité plus en profondeur. L'individu peut également revoir sa stratégie de recherche si les résultats obtenus sont trop loin des résultats attendus. L'évolution de la représentation du but, avec son pendant négatif qu'est le risque de perdre ce but, fait du maintien de sa représentation dans un état stable la principale difficulté du module d'évaluation. Ce risque est plus important dans le cadre de systèmes hypertextes. Une surcharge cognitive importante peut en effet être la conséquence de la complexité des chemins parcourus depuis le début de la navigation, et l'utilisateur peut ne pas se rappeler des documents précédemment consultés ni du but initial de sa recherche. Cependant, cette évolution du but n'est pas entièrement négative : si le but est flou, la modification de sa représentation peut alors consister à le spécifier. Par ailleurs, [Ericsson & Kintsch, 1995] montrent que cette déformation du but est beaucoup plus faible quand l'individu est expert, à la fois du contenu et de l'utilisation du système de navigation.

Les approches psycho-ergonomiques de la modélisation de l'activité de recherche d'information décrivent quant à elles les composantes de l'activité, et non son déroulement. Enfin, les approches écologiques telles que la théorie de la fouille d'information de [Pirolli & Card, 1999] postulent que l'activité de recherche d'information chez le sujet humain obéit à un principe de maximisation du rapport bénéfice/coût, qui représente la valeur de cette information. Le coût relève pour l'utilisateur du temps d'accès à l'information, du coût de la fouille et du traitement.

Ainsi, si le coût est trop élevé par rapport au bénéfice que l'individu pense en tirer, la solution de recherche d'information ne sera pas utilisée. Cette approche rappelle la loi du moindre effort posée par [Zipf, 1949], puis reprise par [Mann, 1987] dans le contexte de la recherche d'information. Dans ce cadre, la loi du moindre effort tend à montrer que pour effectuer une recherche d'information, dans des documents papiers comme dans des documents électroniques, un individu aura tendance à adopter la méthode de recherche la plus pratique, préférentiellement qu'il connaît déjà, et qui lui coûtera le moins par rapport au bénéfice retiré.

2.2.2 Les principales caractéristiques de la tâche influençant l'activité de recherche d'information

Les modélisations de l'activité que nous avons exposées, telles que l'EST, les approches psycho-ergonomiques ou les approches écologiques, ont un caractère assez général, qui vaut pour les grandes « composantes » de l'activité de recherche d'information. En réalité, la navigation dépend grandement de certaines caractéristiques propres à une situation donnée, comme l'implémentation du but de recherche, le contexte dans lequel elle s'effectue, en somme, des caractéristiques de la tâche.

Nous ne détaillons pas ici l'ensemble des principes et résultats de la recherche ergonomique, mais exposons plutôt ceux qui ont une influence directe sur la conception d'un système de navigation performant dans un contexte industriel.

La nature du besoin d'information en premier lieu, découlant directement de la nature et de l'objectif de la tâche principale, est prépondérante. Ce besoin a un impact fort sur l'information qui est recherchée, et sur la manière dont la recherche s'effectue. Le système de navigation doit donc être conçu avec la perspective des tâches principales que les utilisateurs ont à réaliser, qui influencent leur activité et dans lesquelles intervient la tâche secondaire de recherche d'information.

Le fait de rechercher de l'information dans des documents issus de différentes sources, notamment, est une caractéristique qui influence fortement l'activité de navigation. Il a été démontré, entre autres par [Perfetti et al., 1999], que des traitements cognitifs spécifiques étaient mis en place par les utilisateurs de systèmes traitant de tels ensembles documentaires, qualifiés de *multisources*. Des études tendent à montrer que le résultat de ces traitements serait l'élaboration d'une représentation multidocumentaire (*ibid.*), différant sensiblement de la représentation d'un document unique. Le degré d'expertise joue un rôle important dans ces opérations et cette représentation. D'après [Wineburg, 1991], [Wineburg, 1994] et [Rouet, 2000], les individus experts dans la lecture des documents multisources procèdent donc à différentes actions cognitives :

– une indexation des documents, en fonction du type de texte, de son auteur, de la date de publication, etc.

– une comparaison des textes, mettant au jour des contradictions, des corroborations, etc.
– une contextualisation du contenu, en fonction des lieux, des temps, des conditions auxquels il est fait référence.

Plus récemment, [Roselli, 2010] a démontré que des étudiants, cette fois sur Internet, et face à des documents multi-sources, mettaient en place des stratégies de lecture partielles, de déplacements fréquents d'un document à un autre et de recoupements entre ces documents pour appréhender un vaste ensemble de textes.

La représentation multi-documentaire résultant de ces traitements est composée de deux éléments, d'après [Perfetti et al., 1999], cité par [Tricot, 2003] :

– un modèle intertexte dans lequel les différentes sources d'information sont représentées, ainsi que certains éléments de contenu et les relations intersources ;
– un modèle intersituation dans lequel les différentes situations proposées par les documents sont représentées.

Cette représentation multi-documentaire n'est pas sans rappeler les notions linguistiques d'intertextualité et d'architextualité en sémantique interprétative chez [Rastier, 2001]. Selon lui, la signification est affaire de contextualisation, quel que soit le niveau, du mot à l'ensemble documentaire. Il pose en effet par ces deux notions l'importance de la mise en contexte d'un texte donné, ou d'un passage d'un texte donné, dans un ensemble documentaire déterminé. Cette contextualisation influence la signification de ce texte ou passage, et peut modifier son interprétation. Le principe d'intertextualité se place au niveau de passages de textes différents, et implique qu'ils sélectionnent réciproquement des éléments de signification lorsqu'ils sont mis côte à côte. Le principe d'architextualité considère un texte dans sa globalité, et pose à la fois que son sens est influencé par l'ensemble du corpus ou ensemble documentaire dans lequel il se trouve, et qu'il influence le sens de chacun des autres textes composant ce même ensemble documentaire. Ainsi, cette relation d'influence est réciproque : non seulement le global influence le local, mais le local lui-même influence le global.

A la lumière de ces principes sémantiques, il paraît évident que la représentation multi-documentaire construite par un lecteur à partir d'un ensemble de documents est influencée par le fait même qu'il soit face à un ensemble, et non à un texte isolé. Les stratégies de lecture mises en place que nous avons énumérées ci-dessus sont la conséquence directe de la prise en compte de l'ensemble documentaire en tant qu'unité plus ou moins cohérente ; la représentation qui en résulte est influencée par cette relation entretenue entre local et global dans l'ensemble de documents. En allant plus loin, nous pouvons intégrer l'utilisateur à cette influence réciproque entre ensemble documentaire et document. En effet, pour peu que l'utilisateur soit celui qui a constitué l'ensemble documentaire, ses propres choix et son point de vue peuvent influencer la signification qui se dégage à la fois de chaque document et de leur ensemble. De même, il influence cette signification en tant qu'interprétant de l'ensemble.

Un système de recherche d'information « classique » - moteur de recherche ou système de navigation - est un système où l'ensemble documentaire et les documents s'influencent mutuellement, avec une interaction de l'utilisateur en tant qu'agent externe. Or, nous proposons d'intégrer l'utilisateur en tant qu'il influence lui aussi la signification des documents. Nous concevons alors un système de recherche d'information à trois éléments fondamentaux, chacun exerçant sur les autres une influence conditionnant l'interprétation. Nous représentons dans la figure 2.3 un système « standard » (à gauche), ainsi que notre proposition de modèle idéal (à droite).

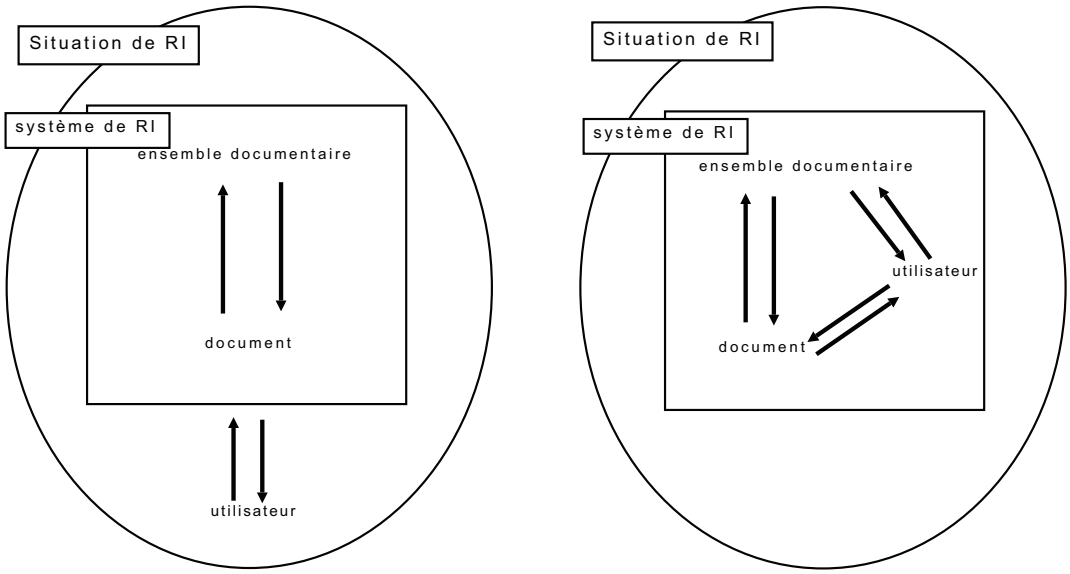

FIGURE 2.3 – Deux manières d'envisager l'utilisateur en situation de recherche d'information (RI) : l'utilisateur comme agent externe du système (à gauche), ou l'utilisateur comme composante interne du système (à droite)

Sur le schéma de gauche, l'utilisateur est à l'extérieur du système d'information, et l'influence de ce dernier sur l'ensemble documentaire et les documents est limitée, du fait même que l'utilisateur est un élément externe au système. En revanche, dans la figure de droite, l'utilisateur est entièrement intégré au système, et en devient une composante. De ce fait, l'influence qu'il peut avoir sur l'ensemble documentaire et les documents, et celle que ces derniers peuvent avoir sur lui, sont elles aussi prises en compte et intégrées comme faisant partie du fonctionnement du système. L'interaction est alors plus forte entre les trois éléments.

Ce type de représentation multi-documentaire soulève par ailleurs des questions concernant le but et le point de vue des individus impliqués dans une situation de communication par le biais de documents. Les courants travaillant sur les documents multimédias , comme [Schnotz, 2001] par exemple, considèrent que le but de l'auteur d'un document est que le sujet comprenant - le « récepteur » - puisse procéder à l'élaboration d'une représentation cohérente du document en tant qu'unité. A l'inverse, pour certains auteurs travaillant plus précisément sur des documents

multisources [Perfetti et al., 1999], le but est « l'élaboration par le récepteur (le sujet comprenant) d'une représentation multidocumentaire (pas forcément cohérente du point de vue des auteurs), intégrant un modèle intersituation et un modèle intertexte » [Tricot, 2003].

Les deux cas de figure trouvent une place au sein du célèbre schéma de la communication de [Jakobson, 1963] [8]. Pourtant, si dans le premier cas, le modèle est orienté par l'émetteur du message dans une situation de communication, dans le second, c'est avant tout le récepteur qui est considéré. Ces deux types de modèles, c'est-à-dire les modèles orientés émetteur et les modèles orientés récepteur, ne nous paraissent cependant pas incompatibles. En effet, il est courant qu'un document voué par son émetteur à un objectif de communication donné soit détourné de son but initial et utilisé à d'autres fins. C'est d'ailleurs ce que fait souvent « l'usager numérique » [Merzeau, 2010], qui détourne outils et documents numériques à d'autres fins que celles qui ont été prescrites par les concepteurs ou auteurs (*ibid.*). Dans ce cas, le document lui-même peut avoir été construit par l'émetteur dans un objectif de cohérence, tout en faisant l'objet d'une interprétation en tant qu'élément d'une représentation multi-documentaire, elle aussi cohérente, par le récepteur.

Un système de navigation efficace doit donc, dans la mesure du possible, rendre compte à la fois d'un document en tant qu'unité, et en tant que partie d'un ensemble partageant avec d'autres documents des éléments communs ou au contraire dissemblables. De cette manière, les objectifs de communication de l'émetteur comme du récepteur peuvent être respectés.

Du point de vue de l'ergonomie, un système de navigation doit donc prendre en compte un certain nombre de paramètres touchant à l'activité cognitive des utilisateurs pour être réellement efficace dans son contexte d'utilisation. Dans le contexte de nos travaux, ces paramètres sont de trois ordres :

1. la tâche principale : en tant que tâche secondaire, la recherche d'information s'intègre à une tâche principale, qui oriente le besoin d'information et conditionne l'activité. Il convient donc de la définir pour concevoir le modèle du système ;

2. le contexte de réalisation de la tâche : il s'exprime dans des critères tels que le temps imparti, le degré d'expertise des utilisateurs, les outils à disposition, etc. ;

3. le caractère multi-sources des ensembles documentaires traités : il est un facteur capital dans le déroulement de l'activité de recherche d'information des utilisateurs d'un système, particulièrement en ce qui concerne la présentation de l'information, du général au particulier, du global au local, de l'ensemble documentaire au document.

8. Le schéma de la communication de Jakobson est présenté en figure 3.1 page 72.

2.2.3 Quel système pour quelles caractéristiques ?

Nous avons déterminé dans la sous-section précédente les paramètres ergonomiques qui nous paraissent les plus pertinents dans notre contexte applicatif : l'objectif de la tâche principale, le contexte applicatif, et enfin, le caractère multi-sources des documents à examiner.

Nous avons par ailleurs présenté, dans le chapitre 1, certaines caractéristiques de la tâche dévolue aux utilisateurs de TKM et de l'activité qui permet de l'accomplir, à travers le prisme des composants de la pertinence de [Mizzaro, 1998]. Il s'agit à présent de caractériser l'activité des futurs utilisateurs à la lumière des paramètres ergonomiques listés. Ainsi, certains aspects déjà énoncés dans le chapitre 1 sont ici repris, et considérés cette fois du point de vue de l'ergonomie.

2.2.3.1 Définition des contraintes de la tâche influençant la conception d'un système d'immersion

Nous avons établi, dans la section 2.1 page 26 de ce chapitre, que la navigation documentaire, préférentiellement heuristique, est pour notre contexte applicatif la meilleure façon de réaliser une recherche d'information. Par ailleurs, nous avons vu, dans la sous-section 2.2.2 page 34 (et plus spécifiquement à travers la figure 2.3 page 36), qu'inclure l'utilisateur au sein du système de RI, plutôt que de le laisser en marge de son fonctionnement, permet d'enrichir l'interaction entre ses composantes. De cette façon, la compétence de l'utilisateur est intégrée à la machine et aux connaissances internes aux documents. A cette fin, le système conçu dans le cadre de nos travaux permet à l'utilisateur de *s'immerger* dans l'ensemble documentaire qu'il étudie, et l'incite à se plonger dans les documents. Nous nommons donc notre modèle *système d'immersion documentaire*. Examiner les contraintes de la tâche et le déroulement d'une activité de recherche d'information permet d'orienter la conception d'un tel modèle.

En premier lieu, la tâche de recherche d'information peut se caractériser par rapport à la tâche principale qu'elle sert. Ainsi que nous l'avons expliqué en début de section, de nombreuses activités nécessitent une recherche d'information pour la résolution de problèmes. En ce qui concerne les activités qui se déroulent sur notre terrain applicatif, nous avons vu dans la section 1.2 page 17 qu'elles se situent dans un contexte professionnel et industriel. Elles relèvent toutes du conseil et du service en stratégie de l'innovation, qui recouvre, nous l'avons expliqué, un grand nombre de possibilités, formant autant de tâches principales : étude de marché, état de l'art technologique, recherche de partenariats, etc.

La recherche d'information intervient donc avec des objectifs globaux extrêmement variés. Cette variété doit être prise en compte dans la conception du système de navigation, en tant que contrainte forte. En effet, une variété de tâches entraîne une variété de besoins d'information, et conséquemment de buts d'information.

Malgré cette variété, il existe des points communs à l'ensemble des activités en présence, sauf

rare exception. Le premier est que toute mission, quelle que soit sa nature, donne lieu à la remise d'un livrable au client. Ce livrable contient le résultat du travail effectué par les analystes de la société, ainsi que leurs recommandations quant à la question posée qui a initié la mission. Ces recommandations sont justifiées par les recherches qui ont été menées et les éléments objectifs qui en ont été tirés. Les livrables ne contiennent pas les documents eux-mêmes, mais les références de certains d'entre eux ainsi que certaines informations qui en ont été tirées.

Le deuxième point partagé réside dans la méthode propre à la société, fondée sur une analyse de documents scientifiques et techniques (voir la section 1.2 page 17).

Le recours à ces sources documentaires permet aux analystes de mener des analyse pointues concernant l'existant dans un domaine ou sous-domaine scientifique ou technologique donné. La tâche de recherche d'information venant en support de la tâche principale est donc primordiale, et quasiment incontournable dans l'activité des analystes, puisqu'elle représente l'une des valeurs ajoutées de la société. L'utilisation quasi-exclusive de ce type de documents a été choisie pour résoudre les problèmes mal définis, en particulier, ainsi que nous l'avons mentionné, en raison de la crédibilité qu'elle apporte. En effet, selon [Cho, 2001], la crédibilité des sources est le facteur qui affecte le plus l'activité de recherche d'information, avant les facteurs liés au degré de connaissances antérieures de l'individu et les besoins de recherche d'information.

Enfin, le document, en tant qu'objet unitaire, a une importance prépondérante, au-delà des informations détaillées qu'il véhicule. Bien que ces dernières aient également un grand intérêt dans l'ensemble des missions, le fait même qu'un document donné ait été produit est une information en soi. En effet, l'existence d'un document comme un brevet ou un article scientifique prouve un lien de propriété entre un contenu et son auteur, mais aussi le simple fait qu'un contenu donné a été traité. Ces informations sont au cœur de l'activité des analystes de TKM. Dans notre cadre applicatif, le document est donc à considérer à la fois comme contenu et comme support. Cet aspect du document est pris en considération dans le modèle du système d'immersion, ce qui revient à dire que le contenu du document ne peut être considéré de manière isolée, sans son contexte de production, qu'il s'agisse de contexte linguistique ou de la forme documentaire de départ.

Ainsi, alors que les points communs à toutes les missions ont tendance à faire naître un besoin d'information systématique, les divers objectifs des tâches à mener par les analystes entraînent une grande diversité dans ces besoins, et dans les buts de recherche qui en découlent.

Dans la plupart des cas, le besoin d'information né de la tâche principale est déjà exprimé par le client au début d'une mission. Ce besoin n'est certes pas toujours très précis, et c'est alors à l'analyste de le dégager plus précisément en collaboration avec celui-ci. Il peut être amené à évoluer au cours de l'étude, faisant évoluer du même coup le but de recherche d'information de l'analyste. L'identification des mots-clés lors de ces échanges fait partie intégrante de la méthode de la société, et permet d'établir une « photo avant étude ». Ces mots-clés constituent le point

de départ de la recherche, permettant de la démarrer et de gagner du temps, même si la liste peut être enrichie par la suite à l'initiative des analystes, mais toujours en accord avec le client.

Passée cette étape de définition, puis celle de récupération des documents (toutes deux incluses dans la phase 1 sur la figure 1.2 page 20), l'analyste doit donc aller chercher l'information correspondante qui lui permettra de réaliser son étude. Si le matériau premier est très souvent le même, c'est-à-dire constitué d'articles scientifiques et/ou de brevets d'invention, il ne s'intéresse pas toujours aux mêmes informations. Nous considérons en effet que les documents véhiculent des types d'informations distincts, qu'il est possible d'organiser intuitivement en catégories. Par exemple, un état de l'art engendre une focalisation sur les technologies elles-mêmes, et donc sur les thèmes précis développés dans les documents ; en revanche, une étude de marché s'intéresse surtout aux organisations publiant ces documents, c'est-à-dire à un autre type d'information. Enfin, si l'objectif de la mission est amené à être modifié au cours de l'étude, les besoins d'information évoluent de même. Ainsi, l'analyste utilise des documents de même type de manière très différente. Par conséquent, le système de recherche d'information doit pouvoir fournir différentes manières d'accéder à un même document et/ou à un même ensemble documentaire, en fonction du besoin de l'utilisateur pour une mission donnée.

Afin de caractériser les besoins des analystes, nous pouvons rappeler la typologie des besoins de [Tricot, 2003] évoquée en partie 2.2.1 page 30 :

«

1. rechercher une connaissance que l'on n'a pas ;
2. rechercher une confirmation d'une connaissance que l'on a ;
3. rechercher une connaissance plus complète que celle que l'on a (mais aussi : un exemple, une illustration, un contre-exemple, etc.) ;
4. rechercher pour être conforme aux buts, aux contraintes, aux attentes de la situation ;
5. rechercher des indications sur la forme de la connaissance à utiliser dans la situation ;
6. rechercher parce que l'on a détecté un marqueur de pertinence dans la situation (ostension, mise en exergue visuelle, sonore, etc.).

»

Rappelons que les trois premiers types de besoins ont une source interne à l'individu, tandis que les trois derniers ont une source externe. Les besoins d'information des analystes de TKM relèvent surtout des quatre premières catégories, c'est-à-dire de l'ensemble des sources internes, et de la première des sources externes. En effet, les types 5 et 6 sont peu pertinents dans notre contexte applicatif : la forme de la connaissance à utiliser dans la situation est déjà prédéterminée lors de l'expression des besoins du client pour lequel la mission est réalisée, ainsi que nous

l'avons mentionné ci-dessus. Une étude de marché implique un besoin d'information prioritaire sur les acteurs d'un domaine, qu'il s'agisse de concurrents ou de clients potentiels ; un état de l'art nécessite une connaissance sur les technologies d'un domaine. Par ailleurs, le fait même d'être uniquement en présence de documents textuels aux genres connus que sont les articles scientifiques et les brevets oriente et contraint déjà les formes possibles de connaissances. D'autre part, dans ces genres de documents, les éléments que [Tricot, 2003] appelle des marqueurs de pertinence sont facilement identifiables, et sont limités. Ils se situent généralement dans la structure explicite des documents, tels que les titres, les mots-clés fournis par l'auteur, etc. Au-delà de ces éléments, il existe peu de marqueurs plus « libres », en raison de la rigidité des deux genres.

Le type de besoin externe qui nous intéresse ici, à savoir celui de rechercher pour être conforme aux buts, contraintes et attentes (type 4), découle directement de l'objectif de mission des analystes, et du fait que nous nous situons dans un contexte industriel. Le milieu professionnel implique inévitablement des contraintes, qu'un individu se doit de respecter pour répondre aux exigences de son poste. En l'occurrence, une mission implique nécessairement un but, plus ou moins précis, et des contraintes et attentes pouvant varier. Par exemple, il peut exister une contrainte temporelle forte si une mission doit être terminée pour le dépôt d'un dossier pour un concours relatif à l'innovation. Ou encore, il peut exister une attente d'exhaustivité concernant un état de l'art en vue d'un dépôt de brevet d'invention par le commanditaire.

Quant aux trois types de besoins internes (types 1 à 3), ils s'articulent autour de ces exigences externes : il est tout d'abord très courant, pour ne pas dire systématique, qu'un analyste ne possède pas en mémoire interne l'ensemble des connaissances suffisantes à la réalisation de sa mission. Il doit donc rechercher une connaissance qu'il n'a pas (type 1). Par ailleurs, il peut aussi avoir une connaissance, mais avoir besoin d'une confirmation, servant à justifier ses arguments par des preuves tangibles exprimées dans les documents (type 2). Enfin, il peut posséder la connaissance, mais avoir besoin de l'illustrer ou de l'enrichir (type 3).

Les types de besoins des analystes sont donc relativement larges, et finalement, c'est la source de besoin externe à l'individu, à savoir la conformité aux objectifs exprimés, qui fait naître les besoins internes. La palette à couvrir par un système de navigation de ce point de vue est donc large.

Une fois que le besoin d'information a été déterminé, le but qui en découle peut être défini, nous l'avons exposé (partie 2.2.1 page 30), par l'interaction entre la représentation de ce besoin et la cible visée. Nous rappelons dans la table 2.2 les quatre types de buts qui ont été définis par [Tricot, 1993] (présentés une première fois dans la table 2.1 page 31).

Là encore, en fonction de la mission, mais aussi de son stade d'avancement, toutes ces configurations peuvent exister pour les analystes de TKM. Au début d'une étude, l'analyste s'attachera le plus souvent à butiner parmi les données (cellule en bas à droite du tableau 2.2), de manière à prendre connaissance d'un domaine ou sous-domaine dans sa globalité. Par la suite, il pourra

Localisation de la cible	Représentation du besoin informationnel : Précise	Représentation du besoin informationnel : Floue
Unique, localisée	Chercher un renseignement	Explorer
Multiple, distribuée	Collecter	Butiner

TABLE 2.2 – Rappel des quatre buts de recherche d'information d'après [Tricot, 1993]

être amené à collecter des informations (cellule en bas à gauche du tableau), une fois que la représentation du besoin est plus précise. A d'autres étapes plus avancées, l'analyste peut ne pas avoir de représentation précise de son besoin d'information, qui concerne alors généralement un point restreint de son étude. Dans ce cas, il procèdera à une exploration des documents à sa disposition (cellule en haut à droite du tableau). Enfin, généralement dans une phase finale de sa mission, l'analyste peut avoir besoin de rechercher quelques renseignements précis (cellule en haut à gauche du tableau), venant compléter les résultats précédents.

Ainsi, ici encore, ce qui se dégage est la diversité des buts de recherche des analystes de TKM, en fonction de la spécificité d'une mission, ou de ses différents stades d'avancement. Ce que nous avions pressenti de manière intuitive se révèle donc à l'aide de ces outils ergonomiques : la plus grande contrainte métier influençant la conception d'un système de navigation dans notre cadre applicatif vient de la diversité des missions et objectifs, engendrant une exigence de souplesse prépondérante.

Le contexte applicatif est constitué d'autres caractéristiques, qui concernent plus le contexte général que les spécificités d'une mission donnée. En premier lieu, nous avons vu dans la section 1.2 que le temps imparti pour les missions est souvent relativement court. Or, il a été démontré par [Proctor, 2001] que la pression temporelle, en d'autres termes la conscience d'un délai restreint, a une influence négative sur une tâche de recherche d'information. A ce titre, il est important que le système de navigation conçu facilite au maximum la tâche de recherche, de façon à donner à l'utilisateur la possibilité de trouver l'information de façon plus immédiate.

Le degré d'expertise des utilisateurs joue lui aussi un rôle dans l'activité de recherche d'information. Il existe deux types d'expertises : l'expertise conceptuelle d'une part, c'est-à-dire celle du domaine traité, par exemple les biocarburants, et l'expertise procédurale, en l'occurrence celle du système de recherche d'information et des caractéristiques de l'information à traiter, comme le caractère multisources des documents par exemple. Les analystes de TKM sont tous des experts

de la recherche dans des documents multisources. Par ailleurs, ils sont experts dans la recherche d'information avec les outils qu'ils ont actuellement à disposition. Enfin, ils connaissent un grand nombre de domaines techniques et scientifiques, cette connaissance étant le fruit de leurs études, mais aussi de l'expérience acquise au fil de leurs missions. Ils ne sont pas à proprement parler des experts dans chacun de ces domaines, mais ont suffisamment de connaissances pour savoir où et comment rechercher l'information qui leur manque. Dans ces cas-là, leur expertise procédurale pallie en quelque sorte leur manque de connaissances conceptuelles.

Dans tous les cas, nos travaux se placent dans un contexte où les utilisateurs des systèmes de recherche d'information sont formés aux outils à disposition, auxquels ils sont habitués, voire « formatés ». Le public de ces outils est donc restreint et formé, et au bout de quelques semaines d'utilisation, il devient expert.

Comme nous venons de le mentionner, les utilisateurs traitent dans la grande majorité des cas des documents multisources, c'est-à-dire des ensembles documentaires dont les unités émanent de différentes sources. Nous avons vu en 2.2.2 page 34 que la lecture de tels ensembles menait à une représentation multi-documentaire, à partir de stratégies de lecture spécifiques. Le modèle intertexte et le modèle intersituation qui composent cette représentation sont, d'après [Britt et al., 1999], le résultat d'une indexation sélective de certains contenus à certaines sources, caractéristique des lecteurs experts de documents multisources. Par conséquent, il est important de permettre à l'utilisateur du système de navigation d'accéder à l'information de manière adéquate du point de vue de la représentation multi-documentaire qu'il est en train de construire. Il est donc nécessaire de lui donner accès à une représentation globale, mais sélective, des documents qui composent l'ensemble qu'il traite, avant de lui donner la possibilité d'aller dans le détail d'un document particulier qu'il aura identifié.

Le dernier élément du contexte que nous présentons, c'est-à-dire les outils utilisés actuellement, ont une influence forte sur l'activité. Ainsi que nous l'avons rapidement mentionné au début de cette section, dans une certaine mesure, les utilisateurs adaptent leur activité en fonction des outils de recherche qu'ils ont à leur disposition, alors qu'il paraîtrait plus pertinent et efficace que l'outil s'adapte à l'utilisateur et à ses usages. La démarche de recherche d'information intervient, nous l'avons expliqué dans la section 1.2 page 17, après la phase de récupération des documents. Cette dernière permet, à partir d'une liste de mots-clés définis avec le client, de rassembler des documents à partir de bases de données en ligne. L'analyste est donc également le concepteur de l'ensemble documentaire qu'il va étudier.

La phase de recherche d'information à proprement parler s'effectue grâce à la plateforme interne de TKM. Si cette plateforme de recherche se révèle efficace dans des cas où le besoin d'information est relativement précis, elle peut être insuffisante dans les situations où l'utilisateur a un but d'information flou, et/ou que la cible visée par la recherche est disséminée dans plusieurs documents. Dans ce cas, il serait souhaitable que les visualisations existantes puissent

être connectées entre elles, mais aussi que l'accès aux documents soit possible directement à partir de ces vues, et non dans une liste énumérative.

Finalement, le système de navigation doit permettre à l'utilisateur de « glisser » du niveau global au niveau local, de l'ensemble documentaire au document, puis de faire le parcours inverse, éventuellement en modifiant certains critères de recherche, au sein d'une même visualisation. Il peut dans ce cas accéder par le biais d'un parcours interprétatif qui lui est propre aux informations qu'il juge comme répondant à son besoin, plutôt qu'être mis face à des résultats construits *a priori*, graphiques ou textuels, qu'il doit considérer comme « finis » et définitifs. Ce lien à conserver entre global et local pour l'accès aux ensembles documentaires est entre autres défendu par [Roy, 2007]. Nous allons plus loin en envisageant de restituer, outre cette influence entre le niveau global et le niveau local, l'influence réciproque existant entre l'utilisateur et le niveau global, et celle entre l'utilisateur et le niveau local. Dans ce cas, l'utilisateur n'est plus un « simple » utilisateur, mais un concepteur de *sa* propre information pertinente. Pour cela, l'utilisateur-concepteur doit être placé au centre des données, immergé au sein de l'ensemble documentaire et en possession de son libre arbitre, de manière à ce qu'il puisse agir directement dessus grâce à son libre arbitre (voir schémas en 2.2.2 page 34).

Il n'est actuellement pas envisageable d'intervenir sur ce que nous nommons l'outil de récupération de l'information utilisé dans l'exécution d'une mission, c'est-à-dire sur les interfaces de recherche fournies par les bases de données scientifiques et techniques payantes. En effet, elles représentent pour l'instant le seul accès possible en masse aux documents scientifiques et techniques, et il est donc obligatoire de se plier à leurs structure et possibilités dans l'activité de collecte. Par contre, une fois l'ensemble documentaire constitué, il est possible de travailler à un système d'immersion au sein de cet ensemble, adaptable aux besoins des utilisateurs et qui permette un meilleur accès à l'information pertinente. Dans la figure 8.9 page 350 (formalisée à partir de la figure 1.2 page 20), nous situons la place de ce système d'immersion (zone cerclée sur la figure), par rapport aux grandes étapes de l'activité d'un analyste, prescrite par des sous-tâches rattachées à une tâche principale de réalisation d'une étude.

Nous rappelons que les deux premières phases peuvent être itératives, puisque l'analyse peut mener à une nouvelle recherche d'information, voire, moins couramment, à une redéfinition des mots-clés. Enfin, à l'issue de ces étapes, l'analyste rédige le livrable qui sera remis au client.

Pour qu'il soit réellement pertinent, notre système d'immersion doit être à la fois utile, utilisable et acceptable. Cette caractérisation en termes d'utilité, d'utilisabilité et d'acceptabilité est relative à un contexte de recherche d'information donné, et doit donc être envisagée comme telle, par rapport aux caractéristiques contextuelles que nous venons d'énoncer.

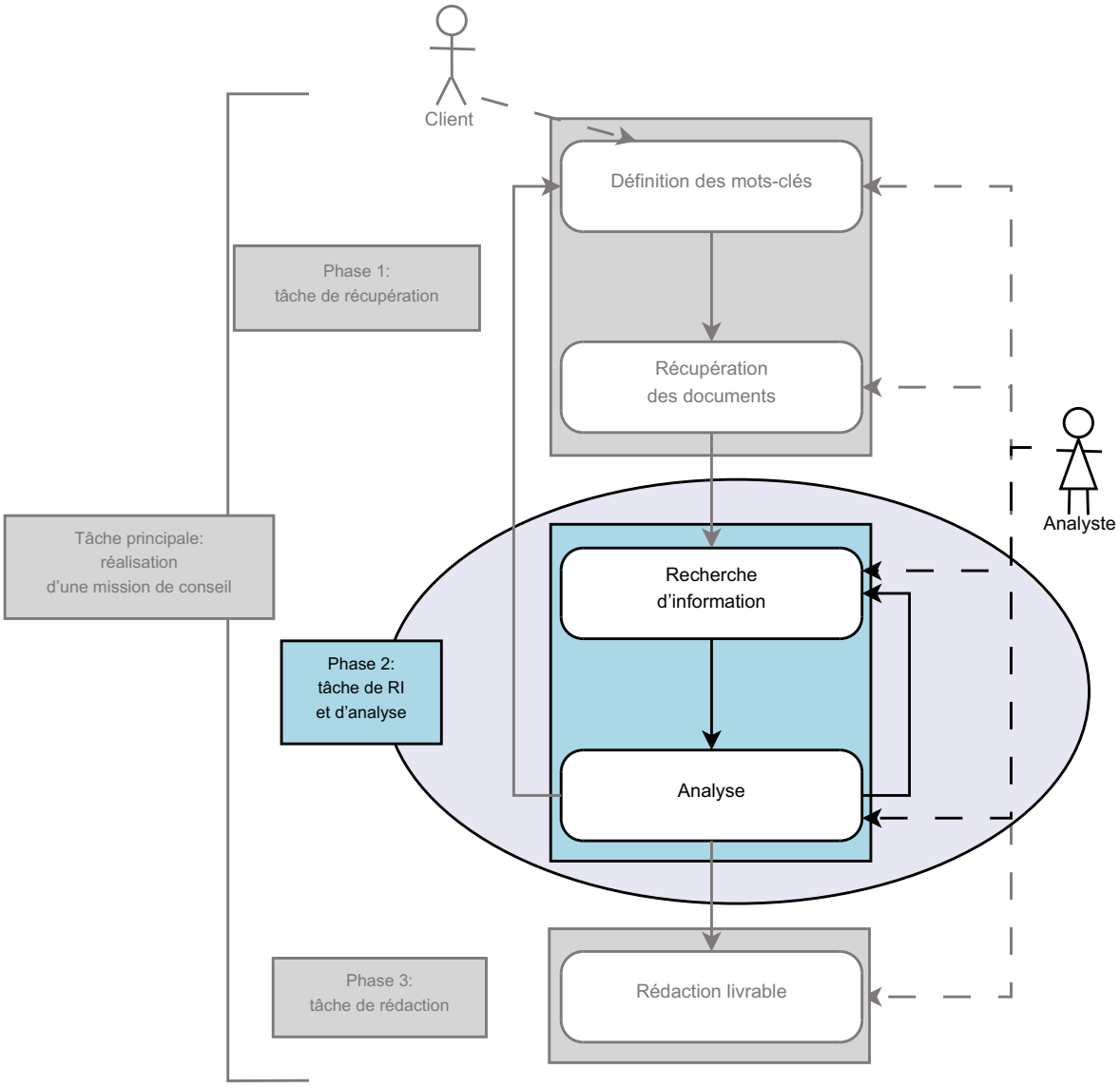

FIGURE 2.4 – Intervention de notre système d'immersion dans le déroulement par étapes de l'activité de l'analyste, pour une tâche de réalisation d'étude

2.2.3.2 La caractérisation ergonomique de la qualité d'un outil : utilité, utilisabilité, acceptabilité

Nous présentons succinctement les notions d'utilité, d'utilisabilité et d'acceptabilité. Elles sont les trois aspects fondamentaux qui permettent de déterminer la qualité d'un outil en termes d'ergonomie, et représentent pour nous des points d'appui pour la conception qui soit en adéquation avec la fonction qui lui est assignée. Nous considérons donc que ces trois éléments, typiquement utilisés comme critères d'évaluation d'un outil déjà terminé, constituent dans notre cas un guide de bonnes pratiques à exploiter pendant la conception. En effet, de notre point de vue, respecter ces points permet d'obtenir un système efficace et adapté à une situation donnée.

L'utilité est, d'après [Tricot, 2001], l'adéquation entre les buts de l'utilisateur et la finalité du document électronique ou logiciel à concevoir. Cette notion étant fortement liée à celle de pertinence (cf. chapitre 1 page 9), Tricot précise sa définition, en ajoutant que cette adéquation entre finalité du document ou logiciel et but de l'utilisateur vaut « pour un domaine, une exploitation et un environnement donnés » (*ibid.*). En somme, un système de recherche d'information utile est un système capable de répondre aux buts de recherche d'un utilisateur en lui apportant les réponses qu'il cherche. L'utilité se mesure comme la possibilité d'atteindre un but visé à l'aide du système (*ibid.*). En l'occurrence, il s'agit de la possibilité de trouver l'information pertinente à l'aide du système de recherche d'information.

L'utilisabilité, comme son nom l'indique, concerne quant à elle l'utilisation de l'objet à concevoir. Cette utilisabilité peut se mesurer, selon [Nielsen, 1993], grâce à cinq critères :

1. **l'efficience** est le fait d'atteindre sans perdre de temps le but qui a été fixé ;
2. **l'apprenabilité** est la facilité d'apprentissage du fonctionnement du système par l'utilisateur ;
3. **la mémorisation** concerne le fait que l'utilisateur parvienne à retenir facilement ce qu'il a fait et comment le système fonctionne ;
4. **la fiabilité** mesure le degré de prévention ou de gestion des erreurs par le système ;
5. enfin, **la satisfaction** subjective de l'utilisateur a à voir avec le fait que le système soit agréable à utiliser.

Ces cinq critères, plus facilement mesurables que la seule notion d'utilisabilité, permettent d'objectiver cette dernière, de façon à rendre compte par des observables de certaines qualités (ou au contraire de défaillances) du système évalué en cours de conception.

Enfin, **l'acceptabilité** est une notion double, et recouvre les idées d'acceptabilité sociale d'une part et d'acceptabilité pratique d'autre part. L'acceptabilité sociale concerne les valeurs des utilisateurs d'un système, et par là leurs opinions ou attitudes concernant le système à

concevoir [9]. L'aspect pratique a plutôt à voir avec des critères tels que le coût du système, son support, sa compatibilité avec les systèmes existants, etc. [Nielsen, 1993]. Si dans notre cas, le pendant social de ce critère ne rentre pas vraiment en jeu, son aspect pratique en revanche est très important dans un milieu industriel où des infrastructures et procédés sont déjà en place et bien ancrés dans les fonctionnements. Par exemple, un système qui nécessiterait de changer totalement le mode de stockage des données servant de matériau à la recherche d'information serait peu acceptable, en raison des efforts à déployer et de l'importance des changements à réaliser pour implanter le nouveau système.

L'acceptabilité pratique exige donc que le système d'immersion à mettre en place sur notre terrain applicatif s'intègre aux procédés et outils déjà utilisés dans l'entreprise. Ainsi que nous l'avons évoqué plus haut, il n'est par exemple pas envisageable à l'heure actuelle de modifier la phase de récupération de l'information sur les bases de données en ligne. Par conséquent, le matériau premier et les méthodes par lesquelles il est acquis ne peuvent être changés. De même, le format d'import des données doit lui aussi être respecté. Le système d'immersion doit être adapté à ces contraintes, ce qui impose d'ores et déjà la méthode de constitution de l'ensemble documentaire ainsi que son format, en l'occurrence une base de données relationnelle.

La notion d'utilisabilité impose que le système d'immersion réponde au mieux aux cinq critères qui la définissent. Tous convergent vers une prise en compte de la facilité d'utilisation, que celle-ci concerne l'apprentissage de l'utilisateur ou l'activité de recherche elle-même. Par ailleurs, l'utilisation est d'autant plus facilitée que le système est fiable. Tous ces critères rentrent en jeu dans la satisfaction subjective de l'utilisateur, également fondée sur un certain confort de prise en main et sur le fait que le système soit globalement agréable à manipuler. En somme, la conception d'un tel système doit passer par la prise en compte de l'ensemble de ces dimensions pour qu'il soit effectivement utilisé.

Enfin, la notion d'utilité pourrait sembler triviale : un système conçu dans un certain objectif devrait inévitablement y répondre. Pourtant, il arrive souvent que des systèmes soient implantés sans que cet objectif ait été réellement étudié, dans sa spécificité comme dans ses différentes exigences. C'est par exemple le cas de certains systèmes de recherche d'information qui ont été conçus pour un contexte donné, puis qui ont été implantés dans d'autres contextes, et ce sans succès. Chaque situation est en effet différente, et en ce qui concerne la recherche d'information particulièrement, elle « n'a de sens que contextualisée » [Tricot, 2003].

Dans notre cadre applicatif, nous avons constaté que certaines caractéristiques de la tâche étaient extrêmement contraignantes pour l'activité de recherche d'information. Par ricochet, ces

9. Pour illustrer l'acceptabilité sociale d'un système, nous pouvons citer l'exemple d'outils tels que les générateurs de clés, qui permettent d'utiliser des logiciels d'ordinaire payants sans s'acquitter des droits d'utilisation. Ces outils génèrent une clé qui permet de déverrouiller les logiciels en question, et sont donc illégaux. Il est clair que tous les utilisateurs, en fonction de leurs valeurs, n'auront pas la même attitude vis-à-vis de ce type d'outils. Ces derniers peuvent donc être soumis à une faible acceptabilité sociale.

caractéristiques influent fortement sur la conception d'un système d'immersion efficace. Nous avons en particulier observé que les différentes missions ont toutes en commun (ou presque) le matériau premier de recherche, à savoir des documents scientifiques et techniques aux genres très contraints. La première phase de recherche se déroule par ailleurs toujours de la même manière, et il n'est pas possible aujourd'hui d'intervenir sur cette première étape. Enfin, il existe une tâche de rédaction associée à chaque mission, qui se fonde en partie sur les résultats de cette recherche.

Nous avons montré, dans la section 1.2 page 17 et au cours de cette section, que la diversité des besoins et des objectifs de recherche d'information était l'une des caractéristiques les plus saillantes de l'activité. Nous avons par ailleurs vu que dans un tel cas, mais également dans un contexte multi-documentaire, il était important de rendre à l'utilisateur son libre arbitre, et ainsi de lui laisser la possibilité de construire lui-même sa représentation des données et son information pertinente (voir 2.2.3.1 page 38). Cela exclut donc de le mettre face à des résultats de recherche statiques, et/ou sous forme de listes plates. Au contraire, des résultats de recherche dynamiques et évolutifs lui permettent d'ajuster son but de recherche au fil de son activité, et de passer de vues globales de l'ensemble documentaire en fonction de certains critères à des vues plus précises sur un document particulier par exemple.

Ainsi que nous l'avons vu en 2.2.3.1 page 38, plutôt que de laisser l'utilisateur en-dehors d'un système de recherche d'information, avec la possibilité d'interagir avec lui de manière limitée, nous jugeons pertinent de le placer à l'intérieur du système et de l'ensemble documentaire qu'il a lui-même constitué. Nous estimons en effet qu'en tant que concepteur de l'ensemble documentaire comme en tant qu'interprétant de cet ensemble, il influence sa signification et celle de chaque document, et il construit cette signification au cours de son parcours interprétatif. Ainsi, le diptyque d'influence entre local et global, c'est-à-dire entre document et ensemble documentaire (figure 2.3 page 36, partie gauche), passe à un triptyque avec l'ajout de l'utilisateur-concepteur-interprétant au sein même du système (figure 2.3 page 36, partie droite). Pour avoir un système utile, utilisable et acceptable du point de vue ergonomique, nous faisons donc le choix de placer l'utilisateur au cœur de l'ensemble documentaire qui fait l'objet de sa recherche, de l'y immerger, tout en lui donnant des clés d'interprétation lui permettant de rechercher efficacement, et d'établir des liens entre les documents de l'ensemble par le biais de certains critères qu'il aura lui-même choisis.

Ces clés d'interprétation prennent la forme de connaissances au sujet de l'ensemble documentaire étudié. Celles-ci ne sont pas données explicitement avec les documents, et il est donc nécessaire d'en faire une extraction, puis une représentation afin de pouvoir les exploiter. Cette extraction et cette représentation sont des processus et méthodes relevant de l'intelligence artificielle, et plus particulièrement de l'ingénierie des connaissances. Il s'agit donc à présent de voir en quoi cette discipline permet d'ajouter une *couche* de connaissances entre les documents et l'utilisateur, en quoi elle permet à ce dernier d'interpréter ces documents et l'information qu'ils

véhiculent en lui fournissant les clés nécessaires.

2.3 Construire le sens avec la machine

Nous venons d'examiner la recherche d'information en général, ainsi que sur notre terrain applicatif en particulier, sous l'angle de l'ergonomie cognitive. D'autres points de vue peuvent compléter cette vision de l'activité de recherche d'information, et notamment celui de l'ingénierie des connaissances. Cette discipline interroge le rôle de l'artefact fondé sur des connaissances dans l'activité humaine. Elle fait appel à certaines notions du constructivisme, qui permettent de considérer la connaissance comme un construit.

2.3.1 L'ingénierie des connaissances

Nous avons vu (section 2.2.1 page 30) que, dans un système de recherche d'information, l'utilisateur est celui qui a suffisamment de connaissances pour savoir qu'il a besoin de plus de connaissances pour résoudre un problème, répondre à des objectifs de mission. L'acquisition de ce savoir passe par la consultation d'un système de recherche d'information, lorsqu'il va chercher dans une mémoire externe artificielle les connaissances qui viendront compléter les siennes. Si l'ergonomie se place bien de ce point de vue, c'est également le cas de l'ingénierie des connaissances (IC). Cette discipline s'intéresse particulièrement aux systèmes informatiques utilisant les connaissances nécessaires à une application visée, ou plus exactement leur représentation. Ainsi que nous l'avons mentionné, utiliser de telles représentations permet d'ajouter une *couche* [10] entre les documents ou autres vecteurs d'information et l'utilisateur d'un système contenant ces informations. Cette couche doit être « lisible » par la machine comme par l'humain, et doit permettre à ce dernier d'interpréter les informations fournies par le système. C'est donc l'IC qui structurera les clés d'interprétation fournies à l'utilisateur, lui donnant par là les moyens de construire le sens. Il pourra de la sorte être placé à l'intérieur d'un système, et pourra construire du sens au sein même de l'ensemble documentaire.

L'ingénierie des connaissances renvoie donc dans une certaine mesure à l'ergonomie, bien que l'objet de la discipline soit quelque peu différent. L'IC se place du côté de l'artefact destiné à l'homme, alors que l'ergonomie s'intéresse plus précisément à l'homme en interaction avec l'artefact. L'artefact est, selon [Beguin & Rabardel, 2000], cité par [Pernin, 2007], un outil « initialement conçu et réalisé pour répondre à un objectif précis » [11].

Cela ne signifie pas que l'ingénierie des connaissances ne s'intéresse pas à l'homme placé devant l'outil, au contraire. Mais elle appuie plus sur les aspects propres à l'outil informatique, et

10. Nous empruntons le terme de couche à [Hernandez, 2005], qui parle de *couche sémantique* ajoutée entre les utilisateurs et l'information pour enrichir les systèmes de recherche d'information.

11. Les auteurs distinguent l'artefact de l'instrument, ce dernier résultant « de l'usage de l'artefact à travers un ensemble de schèmes d'utilisation » [Pernin, 2007].

sur la manière de le concevoir pour qu'il soit efficace. Pour cela, cette discipline reprend la notion de résolution de problème également utilisée en ergonomie, et l'applique à l'outil informatique pour la résolution automatique ou assistée de problème. Elle s'en sert pour établir un modèle du problème à intégrer dans l'outil visé, puis pour opérationaliser ce modèle. La résolution de problèmes passe généralement par l'acquisition des connaissances nécessaires à cette résolution, souvent grâce à des mémoires externes artificielles. C'est donc en toute logique qu'elles viennent se placer comme fondamentales dans ces systèmes informatiques.

Pour exploiter des connaissances dans la résolution de problème automatique ou assistée, il convient de les définir sur deux points : d'abord, il faut déterminer quelles sont les connaissances pertinentes. Par ailleurs, il est nécessaire de se poser la question de leur forme. Un modèle formel doit donc être établi, et doit permettre de projeter les raisonnements et connaissances humains dans l'artefact. Pour cela, selon [Charlet, 2004], l'IC met en place des méthodes permettant de « modéliser des raisonnements et le domaine d'activité dans lequel ils s'insèrent ». Cette modélisation se fait sur deux plans distincts. Le premier plan concerne le niveau conceptuel du futur système à base de connaissances : le comportement du système doit faire l'objet d'une description abstraite, en termes de domaine, de méthodes de raisonnement, et de la tâche dévolue au système. Le second plan se place au niveau opérationnel du système à base de connaissances, et est défini au niveau des programmes ; il a trait à la manière dont le modèle conceptuel va être concrètement implémenté.

Comme pour répondre aux principes de l'ergonomie (cf. 2.2 page 28), une large place est donc laissée à la notion de contexte : d'abord, la situation d'utilisation du système à concevoir est déterminante. L'ingénierie des connaissances se place donc en cela du point de vue de Tricot entre autres, qui appuie sur l'importance de ce contexte d'utilisation des outils de recherche d'information [Tricot, 2003].

D'autre part, pour que l'individu en situation d'utilisation puisse interpréter correctement le fonctionnement du système et se l'approprier, il a besoin du contexte d'énonciation des connaissances. Or, pour être exploitables informatiquement, ces mêmes connaissances doivent en même temps être fortement formalisées. Pour résoudre cette tension, il est donc nécessaire de conserver et de rendre accessible au sein même de l'outil conçu ce contexte d'énonciation, c'est-à-dire, dans notre cadre et plus généralement dans le cas de documents textuels, le texte duquel elles sont tirées. En effet, au moins dans les documents textuels, la langue est le matériau qui permet d'exprimer les connaissances.

Ce point de vue s'oppose quelque peu aux visions très normatives et universalistes qui ont présidé à la constitution de ressources dites *universelles*, telles que celle initiée par [Wüster, 1968] en terminologie par exemple. Ce dernier postule en effet l'immuabilité et l'invariabilité des concepts, ce qui implique qu'une représentation de concepts et/ou de connaissances vaut dans l'absolu, indépendamment de l'application à laquelle elle est destinée. Or, ce principe a été fortement remis

en cause par la suite, en particulier par le rapprochement de la terminologie et de l'informatique, donnant naissance à des applications toujours plus variées, pour lesquelles les besoins diffèrent de plus en plus. Le groupe de travail Terminologie et Intelligence Artificielle [12] en particulier, auquel Charlet appartient, pose au contraire la variabilité des ressources en fonction des applications, et donc un besoin de pouvoir retourner à la source de l'information formalisée dans la structure de représentation des connaissances.

Finalement, l'IC comme l'ergonomie s'intéressent aux deux versants du contexte dans une situation de communication où l'émetteur et le récepteur sont séparés, comme c'est le cas d'un document écrit par exemple. A travers la prise en compte de ce contexte, les deux disciplines reconnaissent, implicitement ou pas, l'objectif de cohérence de l'émetteur du message. L'ingénierie des connaissances cherche à conserver le contexte de production d'une information pouvant devenir connaissance. Certains courants de l'ergonomie, comme celui dans lequel s'inscrit [Schnotz, 2001] par exemple, voient dans la production d'un document la volonté de la part de l'émetteur de permettre la construction d'une représentation cohérente du message par le récepteur. Du côté du récepteur du message, le contexte de réception est lui aussi considéré comme crucial. L'ergonomie considère que la tâche et le contexte du récepteur du document, ou de l'ensemble documentaire, jouent inévitablement sur sa recherche d'information et sur son parcours de lecture (cf. 2.2.2 page 34). L'IC modélise le contexte d'utilisation d'un système à base de connaissances, à travers le domaine d'activité, les méthodes de raisonnement, etc., des utilisateurs.

Puisque ces deux champs de recherche font appel à la notion de connaissance, il est important de comprendre comment elle est abordée dans ces deux disciplines. Ces réflexions relèvent de problèmes épistémologiques, et ont notamment été abordées par [Charlet, 2004]. L'IC va assez loin dans le sens où même en tant que discipline d'ingénierie, elle interroge la notion même de connaissance, c'est-à-dire l'une des notions fondatrices de l'ensemble des disciplines scientifiques.

Pour l'auteur, en premier lieu, la connaissance est directement dépendante de la technique. La première prend forme et existe grâce à la seconde, qui la mémorise et permet d'y accéder. Si auparavant, l'écriture puis l'imprimerie constituaient l'environnement technique exclusif lui servant de support, c'est aujourd'hui l'informatique qui, majoritairement, permet de la stocker, de la véhiculer, de la faire évoluer. En somme, hors la technique, point de connaissance. A partir de là, « il y a connaissance et représentation des connaissances quand les manipulations symboliques effectuées par la machine *via* des programmes, prennent un sens et une justification pour les utilisateurs interagissant avec ces programmes » (*ibid.*). Ce sont alors les utilisateurs qui jouent le rôle d'instances d'interprétation, et qui confèrent à ces manipulations symboliques le statut de connaissances. D'autre part, « la pensée repose sur la médiation externe du signe » (*ibid.*). C'est donc, là encore, la technique qui permet de formuler la pensée puis d'en faire des

12. `http://tia.loria.fr/TIA/`

connaissances mémorisées. De ces éléments de définition, il se dégage que c'est l'interprétation humaine qui donne son existence à la connaissance. Par ailleurs, son support est, dans le cadre de l'IC, l'outil informatique.

Par ailleurs, Charlet pose la connaissance comme étant étroitement liée à l'humain et aux traitements que ce dernier effectue. Pour lui, « il y a présomption de connaissance si la faculté d'utiliser l'information à bon escient est attestée ». Une connaissance n'existe donc qu'à travers les traitements humains. De fait, elle n'a de valeur qu'en fonction du contexte dans lequel elle est élaborée et interprétée, et n'existe pas *a priori* : elle est construite à partir d'un projet propre au modélisateur, c'est-à-dire à celui qui l'élabore. Cette conception de cette notion est issue du constructivisme.

Etant donnée cette définition, l'objectif d'un système à base de connaissances devient clair. Il devra fournir des données susceptibles d'être interprétées par l'utilisateur, tout en conservant le contexte d'utilisation d'origine de ces données. Enfin, il devra donner à l'utilisateur « les moyens - informatiques - d'agir et de réécrire les résultats de son interprétation ». L'accent est donc clairement mis sur l'utilisateur, et le lien est ici évident avec l'ergonomie, en particulier avec l'ergonomie cognitive.

L'ergonomie, qui décrit l'activité des individus, donne des clés à l'IC pour la modélisation qui permet de construire l'outil. Elle permet en particulier de définir la forme que doivent prendre les connaissances dans l'outil, la façon dont elles seront présentées à l'utilisateur. [Charlet, 2004] décrit par exemple l'outil Hospitexte, dont il est l'un des concepteurs, et qui a pour point de départ la prise en compte de la manière dont un praticien consulte un dossier médical papier : elle a en effet une influence sur la compréhension de ce dernier. Le projet Hospitexte est né du constat que les outils informatiques de consultation des dossiers de patients dans le domaine médical étaient peu utilisés. Il s'avère que cette sous-utilisation vient du fait que dans ces outils, l'information médicale issue des dossiers est stockée dans des bases de données conservant mal ses caractéristiques contextuelles. Or, une information n'a pas de sens médical par elle-même, mais est fonction du contexte [Charlet et al., 1999]. De fait, ces informations ne peuvent être répertoriées efficacement indépendamment de leur formatage linguistique et documentaire [Bachimont, 2001]. C'est pourquoi Hospitexte conserve la structure documentaire du dossier patient, de manière à permettre les processus de lecture et d'appropriation que les praticiens ont mis en place à partir des dossiers papiers. Par exemple, il est courant qu'un médecin prenne tous les documents papiers d'un dossier et les étale sur une table pour avoir une vision générale du cas. Pour conserver cette structure documentaire, un dossier patient virtuel est créé à partir d'informations éparses. Par ailleurs, une « station de travail » est élaborée dans le but de résoudre les problèmes de surcharge informationnelle. Hospitexte fait intervenir trois types de documents :

1. les **documents originaux** ;

2. les **documents de navigation**, qui sont des hyperdocuments : ils permettent à l'utilisateur de se rendre à un point particulier du dossier. Il peut s'agir d'une table des matières, d'une liste chronologique des documents, etc., d'où il est possible d'accéder au document original ou à l'un de ses passages correspondant à l'objet cliqué ;

3. les **documents de lecture** contiennent les annotations de l'utilisateur pendant sa navigation.

L'utilisateur peut donc accéder aux documents originaux, grâce à une navigation dans l'hypertexte, et par là, construire son parcours de lecture grâce aux fonctionnalités d'annotation. En cela, les méthodes d'ingénierie des connaissances qui ont présidé à la conception d'Hospitexte ont bien pris en compte le modèle de l'activité de l'individu dans cette situation pour élaborer le modèle de l'outil. Par exemple, en rendant possible la navigation hypertextuelle tout en conservant l'accès au document intégral et dans sa forme initiale, l'outil permet de reproduire, virtuellement, le fait d'étaler les documents d'un dossier sur une table. L'utilisateur peut alors naviguer du document à l'ensemble documentaire, du local au global, puis en tirer une information. Hospitexte respecte donc bien les caractérisations de l'ingénierie des connaissances, tout d'abord en ce que sa conception a pris en compte le contexte d'utilisation : pour conserver le contexte d'élaboration des connaissances, les textes eux-mêmes sont conservés. D'autre part, des connaissances nouvelles sont créées, par la génération de tables des matières et autres documents de navigation, des nouveaux documents par agrégation de parties des documents originaux, etc. Par là, « l'informatique crée de nouvelles proximités entre informations médicales, [et donc] de nouvelles connaissances » [Charlet, 2004]. Enfin, du point de vue technique, c'est l'informatique qui est utilisée pour sa « capacité à produire du sens au regard des usages » (*ibid.*).

Dans le même ordre d'idée, Charlet cite le système OncoDoc, guide de bonnes pratiques cliniques de [Bouaud et al., 1998]. Ils ont conçu un système où le guide de bonnes pratiques est implémenté entre texte et formalisation. L'utilisateur lit le guide à l'aide d'un hypertexte, ce qui permet une plus grande souplesse qu'un système d'aide à la décision tout automatique. Cela permet à l'utilisateur d'opérationaliser les connaissances en fonction d'un patient donné pour lequel il doit prendre une décision, dans le cadre de traitements de cancers du sein [Bouaud et al., 1999]. L'objectif est ici que le praticien arrive à déterminer le meilleur traitement possible pour son patient. Le praticien construit lui-même un équivalent « formel » de son patient et de ses caractéristiques dans le système OncoDoc, de manière à pouvoir confronter ces caractéristiques aux connaissances contenues dans le guide. Par là, il peut prendre une décision plus éclairée, et surtout contextualisée, loin des patients « canoniques » décrits par les guides ou construits entièrement automatiquement, qui ne correspondent pas à la réalité de la pratique. Ainsi, là encore, l'accent est mis sur l'utilisateur et sa pratique, son contexte d'utilisation du système.

Ces deux systèmes offrent donc un mode de lecture délinéarisé par rapport au document ori-

ginal : les textes de départ n'ont pas à être lus un à un et du début à la fin, puisque Hospitexte comme OncoDoc fournissent des hyperdocuments permettant de créer de nouveaux parcours de lecture, affranchis des contraintes de linéarité du document papier. Ces deux outils permettent donc à l'utilisateur de réaliser ce que [Lebarbé, 2009] appelle des *lectures rhizomatiques* : l'utilisateur « accède à un ensemble d'éléments textuels linéarisés selon un thème alors qu'ils sont matériellement éparpillés et sans hiérarchisation ».

Au regard de ces exemples et de ce qui a motivé leur conception, il est clair que l'IC part du principe qu'une représentation des connaissances n'équivaut pas strictement pour un outil à véhiculer une connaissance « finie », dans l'absolu. L'outil propose des interprétations privilégiées par rapport à un système de normes déterminé. Par voie de conséquence, le modèle de l'outil n'est pas la représentation du sens, mais un instrument pouvant intervenir dans l'interprétation [Charlet, 2004]. Ainsi, l'IC propose des machines « qui donnent à penser et non des machines qui pensent » (*ibid.*). Dans ce cadre, c'est là encore l'utilisateur qui fait donc figure d'interprétant, et finalement de concepteur de la connaissance et du sens : c'est lui qui créera du sens à partir des inscriptions numériques manipulées par l'outil, qu'il utilisera alors comme des connaissances.

Dans cette perspective, l'ingénierie des connaissances donne des moyens pour la construction de tels systèmes, et en particulier relativement à la couche de « connaissances », ou d'inscriptions numériques, à intégrer dans le système. Ces connaissances peuvent prendre différentes formes, en fonction du degré de formalisation nécessaire à l'application visée. Ces modèles de représentation des connaissances vont du simple lexique à l'ontologie formelle, en passant par des thésaurus ou des taxinomies. Quoi qu'il en soit, toutes contiennent des éléments d'information qui pourront être interprétés par la suite, lors de l'utilisation de l'outil, comme des éléments de connaissances construites dans un contexte donné.

2.3.2 Un peu d'épistémologie : les apports du constructivisme

Nous nous situons donc clairement du côté de l'ingénierie des connaissances comme l'entend [Charlet, 2004], et plus largement les tenants du groupe TIA [13], à l'image par exemple de [Bourigault & Aussenac-Gilles, 2003]. Puisque nous avons pris le parti de laisser l'utilisateur-concepteur construire son parcours interprétatif, cela implique qu'il construira lui-même ses connaissances pertinentes. A ce titre, nos travaux se placent sans équivoque du côté des épistémologies constructivistes.

En effet, comme leur nom l'indique, ces épistémologies postulent que la connaissance relève d'une construction, et n'est pas une simple copie du réel. Savoir ce qu'est la connaissance, ou quelles sont les connaissances certaines, sont des questions propres aux épistémologies positivistes. Or, le constructivisme est né des questionnements beaucoup plus concrets de [Piaget, 1967],

13. `http://tia.loria.fr/TIA/`

qui cherchait avant tout à savoir comment un enfant peut arriver à l'acquisition de ce que nous appelons connaissance. Une connaissance serait donc une construction à partir du réel, une élaboration de structures cohérentes et non contradictoires, qui permettrait à l'homme de s'adapter à son milieu [von Glasersfeld, 1994].

Cette position est donc bien loin des épistémologies positivistes, et en particulier de leur axiome ontologique, pour lequel l'objet est un morceau de la réalité à décrire, indépendant des observateurs qui la décrivent [Le Moigne, 1995]. La connaissance serait alors le résultat de cette description, une « copie » de ce morceau du réel.

Pour le constructivisme, si une connaissance n'a pas a être vraie, elle doit en revanche être viable. Cette notion de viabilité, énoncée par [von Glasersfeld, 1994], renvoie à la relation existant entre connaissance et réalité. Selon l'auteur :

> « [...] on jugera « viable » une action, une opération, une structure conceptuelle ou même une théorie tant et aussi longtemps qu'elles servent à l'accomplissement d'une tâche ou encore à l'atteinte du but que l'on a choisi. »

A cela s'ajoutent les actions, opérations, etc., qui ont échoué : elles représentent une connaissance viable de ce qui ne fonctionne pas pour accomplir une tâche ou atteindre un but donné.

Cette connaissance viable se construit à l'aide de deux types de « réalités » : tout d'abord, le constructivisme ne rejette pas en bloc la notion de réalité ontologique, indépendante de la perception et de la pensée humaines. Cependant, le constructivisme radical la considère tout au plus comme une « fiction utile » [Kant, 1787], (cité par [von Glasersfeld, 1994]), qu'il est nécessaire d'utiliser pour appréhender le réel. D'autre part, ce courant épistémologique prend en considération la réalité vécue et tangible de notre expérience. C'est de cette expérience que nous tirerions l'ensemble de nos connaissances.

Selon von Glasersfeld, l'expérience est « constituée par les sensations et par les abstractions empiriques et réfléchissantes dont nous sommes conscients ». C'est donc l'interaction qui crée l'expérience. Il s'agit là de l'axiome phénoménologique du constructivisme, qui postule que c'est l'interaction entre le phénomène à connaître, c'est-à-dire l'objet, et le sujet connaissant qui forme la connaissance de l'objet et le mode d'élaboration de la connaissance, c'est-à-dire l'intelligence [Le Moigne, 1995]. D'autre part, l'axiome téléologique pose l'intentionnalité de l'acte cognitif[14] : une connaissance construite par un sujet est conditionnée par l'objectif qu'il poursuit. Puisque le sujet connaissant a un rôle décisif dans la constitution de la connaissance, à travers son interaction avec l'objet, alors il faut prendre en compte la finalité du sujet, et interpréter son comportement cognitif en termes de « causes finales ». De plus, la détermination de ces finalités semble être le plus souvent endogène, et donc attribuable au sujet lui-même. En d'autres termes, le sujet,

14. La téléologie est « l'étude des fins, de la finalité » (Trésor de la Langue Française Informatisé). Du grec ancien *télos* : « fin, but ».

lorsqu'il construit une connaissance, le fait dans un objectif déterminé, avec une intention relative à l'objet.

En conséquence de quoi, les principes méthodologiques rattachés au constructivisme sont fondés sur l'étude de la connaissance en tant qu'acte. La systémique privilégie la modélisation de cet acte, qui exprime l'interaction complexe du sujet et de l'objet, plutôt que celle de la chose. Le principe d'action intelligente [Simon, 1981], quant à lui, est à rapprocher de la résolution de problème, évoquée en 2.2 page 28. En effet, dans ce principe, les processus cognitifs alternent la mise en œuvre de moyens adaptés à des fins intermédiaires, lesquelles suggèrent de nouveaux moyens, qui évoquent alors d'autres fins possibles. Pour cela, des heuristiques tirées d'expériences antérieures sont utilisées. Elles n'ont pas à être vraies : elles doivent seulement permettre de construire des connaissances viables, faisables. La construction de ces connaissances par l'individu permet d'établir des modèles, qui servent à générer « un monde plus ou moins régulier et prévisible » [von Glasersfeld, 1994]. L'élaboration de ces modèles est motivée par la croyance humaine selon laquelle l'expérience future ressemblera inévitablement à l'expérience passée (concept d'induction chez [Hume, 1963], cité par [von Glasersfeld, 1994]), en tout cas en ce qui concerne les régularités considérées comme viables.

C'est ce principe qu'utilise [Tricot, 2003] lors de la modélisation de la représentation du but, puisque cette représentation est soumise à réajustements constants au fur et à mesure de la recherche et des cycles d'évaluation des informations traitées par le sujet humain (cf. 2.2.1 page 30). Les connaissances acquises aux différentes étapes de recherche sont viables tant que rien ne vient les contredire : lorsque c'est le cas, le but de recherche est modifié, réajusté, pour acquérir de nouvelles connaissances, et ainsi de suite jusqu'à l'atteinte de l'objectif global de recherche.

La connaissance est toujours subjective, puisqu'elle vient avant tout de perceptions et traitements cognitifs humains. Cependant, cela n'empêche pas l'interaction sociale, grâce au développement d'une « intersubjectivité » stable [von Glasersfeld, 1994], permettant l'échange, le partage et la construction de ces connaissances.

La conséquence directe des points de vue constructivistes sur la science est la prise en compte prioritaire du caractère viable d'une solution, d'un modèle ou d'une théorie, en somme, d'une connaissance, sur son éventuel caractère « vrai ». Les critères de sélection d'une solution sont alors l'économie qu'elle permet, sa simplicité ou son élégance, sa compatibilité avec d'autres solutions, et non plus sa vérité ontologique supposée.

Ces prises de position permettent de raisonner sur des disciplines scientifiques dont les épistémologies positivistes ne pouvaient rendre compte, comme la physique quantique, ou bien, plus proche de nos travaux, les sciences de l'artificiel, d'où vient directement l'ingénierie des connaissances. Selon [Simon, 1981], qui a travaillé sur les sciences de l'artificiel, la connaissance n'est pas objective, et n'existe pas *a priori*. Elle est le produit d'un projet propre à un sujet connaissant, à

un modélisateur. D'après [Charlet, 2004], dans le cadre plus restreint de l'IC, cela remet donc en cause la réutilisabilité des ontologies, particulièrement de celles qui ont été conçues avec l'objectif de créer des structures de représentation de données universelles, ou du moins générales. En effet, si l'acte cognitif est intentionnel, alors le sujet connaissant mènera une modélisation dépendant de son projet, et donc de la tâche visée par l'outil, qui sera difficilement réutilisable pour d'autres tâches. Allant dans le même sens, le fait qu'une connaissance ne soit accessible qu'à travers une représentation implique qu'une connaissance et sa représentation sont totalement indissociables.

D'autre part, la modélisation systémique privilégie le projet sur l'objet, l'action sur la chose. Ainsi l'ingénierie des connaissances modélise-t-elle, à l'instar de l'ergonomie, des interactions entre l'homme et l'utilisateur pour la conception de systèmes à base de connaissances, et non des connaissances dans l'absolu, indépendantes de leur usage. C'est d'ailleurs pourquoi à l'heure actuelle, l'ingénierie des connaissances devient de plus en plus interdisciplinaire, bénéficiant des apports de disciplines telles que l'ergonomie, nous en avons parlé, mais aussi de la terminologie, de la gestion, etc. A ces disciplines s'ajoutent également celles dont l'IC tente de modéliser les connaissances.

2.4 L'immersion documentaire pour les documents scientifiques et techniques (et un peu juridiques aussi, quand même)

Nous avons expliqué dans la section précédente que pour respecter le processus d'acquisition, et même de création de connaissances, il était important de laisser à l'utilisateur la possibilité d'agir réellement sur son parcours de lecture et d'interprétation. Nous partons en effet du postulat constructiviste selon lequel la connaissance se construit en interaction et en fonction d'un projet. L'utilisateur d'un système de recherche d'information doit donc avoir la possibilité de construire lui-même son chemin, par des retours en arrière, des changements de direction, etc. Pour cela, il doit en outre être guidé par des clés d'interprétation, qu'il sera libre de manipuler efficacement en fonction de son objectif. En tant qu'agent interprétant, ou en l'occurrence agent « connaissant », il conçoit lui-même sa connaissance par le biais qui correspond le mieux à la situation dans laquelle il se trouve au moment où il effectue sa recherche. Ainsi, l'utilisateur peut être comparé au marcheur de [Machado, 1917], auquel l'auteur rappelait que c'était à lui de construire son propre parcours : « Marcheur, il n'y a pas de chemin, le chemin se construit en marchant ». L'utilisateur est donc ici une sorte d'explorateur, qui doit se frayer un chemin à travers les documents pour arriver à son but.

2.4.1 Au-delà de la navigation : un système d'immersion pour experts scientifiques

Nous avons vu, en 2.2.3.2 page 44, que nous considérons que l'utilisateur doit être entièrement intégré à un système de recherche d'information, comme l'une des composantes de ce système, en ce sens que dans le cadre de notre terrain pratique, il est à la fois le concepteur et l'interprétant de l'ensemble documentaire. C'est en effet lui qui constitue (construit) l'ensemble documentaire dans lequel il cherchera l'information, et celui qui interprète (construit le sens) les documents et l'information qui lui est fournie par le système, pour en tirer du sens. Selon les principes constructivistes, il est aussi modélisateur (constructeur) de la connaissance qu'il tire du système, par son interaction avec ce dernier.

Notre enjeu est donc de faciliter la modélisation de cette connaissance par l'utilisateur, et ce quel que soit le besoin d'information. Nous avons constaté que certains besoins, et donc certains objectifs, étaient invariants. Par exemple, la recherche d'information est presque systématique, de même que la rédaction d'un livrable pour le client. Dans tous les cas, les analystes cherchent à acquérir des connaissances sur un domaine qu'ils ne connaissent pas bien, ou en tout cas pas suffisamment pour répondre à la question qui leur est posée. Cependant, une constante de l'activité de la société TKM est également une très grande diversité des besoins d'information des analystes, suivant en cela la grande diversité des questions posées par les clients. Pour couvrir la grande variété des besoins d'information en présence, il est donc nécessaire de tirer parti des divers points de vue qu'il est possible d'avoir sur un document ou un ensemble documentaire. Ainsi que nous l'avons mentionné en 2.2.3.1 page 38, l'analyste peut avoir besoin de différents types d'information, bien que tous soient véhiculés par un même document ou ensemble de documents. La conception s'appuie donc sur le fait que les modes d'accès à ces documents doivent être variés, respectant ainsi l'axiome téléologique des épistémologies constructivistes : un objet de connaissance est toujours finalisé, dans le sens où l'intentionnalité du sujet fait que l'objet est abordé avec un objectif particulier. Un même objet est donc composé de plusieurs facettes, correspondant à autant d'objectifs de l'interprétant. Par conséquent, les types d'information *utiles* fournis par le document constituent autant de facettes par lesquelles appréhender l'objet document. La régularité de la présence de ces types d'information permet d'en tirer des modèles utiles pour couvrir la diversité des besoins auxquels vont être confrontés les utilisateurs.

Ainsi, notre système d'immersion fondé sur une représentation des connaissances permet de construire un modèle viable pour une activité donnée. Nous ne formalisons que la connaissance qui a de l'intérêt pour l'activité, et nous la considérons comme connaissance « juste », à défaut d'être « vraie » et exhaustive dans l'absolu, justement parce qu'elle est viable pour l'activité. Conformément aux principes constructivistes, nous faisons donc primer la viabilité et la simplicité pour une situation donnée sur la notion de « vérité ». Nous construisons un modèle d'outil adapté

à plusieurs situations différentes, et non un modèle vrai dans l'absolu, qui dans les faits ne serait pas pertinent pour toutes les situations dans lesquelles se trouvent les utilisateurs. De plus, dans une activité de recherche d'information, un individu doit gérer plusieurs processus cognitifs : cette activité implique pour lui en effet à la fois une planification, un contrôle métacognitif et une régulation de son activité [Tricot, 2003]. Tout cela peut rapidement entraîner une surcharge cognitive, qu'il est donc important de plafonner au minimum. C'est pourquoi nous avons pris le parti de la simplicité, et de ne formaliser que la connaissance utile, limitant ainsi le nombre de types d'informations à manipuler.

D'autre part, des études ont montré, nous l'avons mentionné en 2.2.2 page 34, que face à un ensemble documentaire multisources, les lecteurs experts établissaient une représentation de cet ensemble en créant une sorte d'indexation mentale de certains contenus à certaines sources des documents. En effet, en mémoire à long terme, certaines connaissances sont représentées par la combinaison de leur contenu et de leur source, et non pas seulement par leur contenu [Rouet et al., 1996]. Des accès par la source sont donc nécessaires pour correspondre aux connaissances construites ou à construire par les utilisateurs. Nous cherchons donc à assister l'utilisateur dans son indexation partielle, par un traitement qui indexe les documents en fonction de certains critères. Par conséquent, il s'agit de rendre accessibles tous les liens potentiellement « utiles » entre contenus et sources, pour répondre à l'ensemble des besoins potentiels, tout en permettant une actualisation ciblée de ces liens durant l'utilisation du système, pour que l'utilisateur-explorateur soit aidé dans la construction de sa représentation de l'ensemble documentaire. Ce parti-pris répond également au principe de modélisation systémique des approches constructivistes, qui cherche à exprimer la complexité de l'interaction entre le sujet et l'objet de connaissance, et non la complexité de l'objet seul et pour lui-même. Ainsi, il ne s'agit pas de modéliser l'ensemble des liens possibles entre documents ou entre contenus et sources, mais de modéliser ceux qui sont viables et utiles dans l'activité des utilisateurs du système. Aussi, l'exhaustivité n'est pas nécessaire, et n'est pas non plus souhaitable. Dans notre contexte d'application, la systémique amène à se poser des questions du type : « Que se passe-t-il lorsqu'un sujet connaissant est en interaction avec un document ou un ensemble documentaire ? », et plus précisément :

1. Que fait le sujet avec le document ?,
2. Qu'utilise-t-il dans le document pour cela ? Quels éléments ?

Nous notons que cette dernière question tend à mêler modélisation systémique et modélisation analytique : se poser la question des éléments ou parties du document rentrant plus particulièrement en interaction avec l'utilisateur revient à décomposer un tout (le document) en parties, raisonnement propre au principe analytique des épistémologies positivistes. Les positions positivistes ne sont donc pas exclues du modèle de notre système, au moins en tant que « fictions

utiles ».

Par ailleurs, nous avons vu qu'en ergonomie tout comme en recherche d'information, des relations doivent être établies entre des vues d'ensemble, en l'occurrence des vues de l'ensemble documentaire dans sa globalité, et des vues plus focalisées, soit sur un sous-ensemble, soit sur un document seul. Puisqu'une activité de recherche d'information est itérative ou cyclique, ainsi que la décrit le modèle EST [Tricot, 2003] (cf. 2.2.1 page 30), un système de recherche d'information efficace est un système qui présente des vues dynamiques de l'ensemble documentaire et de ses unités, de manière à permettre le déploiement de l'ensemble du cycle sans interruption. C'est en effet le seul moyen de permettre des allers-retours entre les niveaux global et local, pour que l'utilisateur construise sa propre représentation d'un ensemble documentaire, à travers son propre parcours interprétatif.

D'un point de vue pratique, les résultats de recherche doivent donc présenter à l'utilisateur des visualisations de cet ensemble documentaire et des informations qu'il contient qui dépassent ce que peuvent offrir des listes plates de résultats ou des vues statiques. Il s'agit donc, nous l'avons expliqué, de mettre en place des visualisations dynamiques, permettant de se déplacer du global au local et inversement, en fonction de critères de recherche sur lesquels l'utilisateur, en tant que composante du système, peut agir à n'importe quel moment. Ainsi, l'utilisateur est placé au centre d'un système multidimensionnel, chaque dimension correspondant à un critère, lui-même correspondant à un type d'information véhiculé par les documents, comme les thématiques techniques ou bien les organisations publiant un document. Les liens entre ces dimensions, qui établissent des liens sémantiques entre les documents, constituent donc un réseau dynamique, adaptable, au sein duquel l'utilisateur peut se déplacer à l'envi, et selon les critères souhaités. De là, il peut accéder à un ou des documents particuliers, pour les consulter plus précisément et en tirer l'information dont il a besoin. De plus, puisque le document et son existence même sont d'un intérêt primordial pour les analystes (cf. 2.2.3.1 page 38), il est fondamental de mettre en valeur l'objet document en tant que combinaison du support et de son contenu, à la fois en tant que contexte langagier et formel de la connaissance produite et en tant que connaissance « à part entière ». L'ensemble de ce parcours, ainsi que l'information tirée soit d'un document particulier, soit de la vue globale de l'ensemble, lui permet de construire sa connaissance pertinente pour un objectif donné. Ainsi, la machine et ses inscriptions numériques, si elles ne raisonnent pas à la place de l'utilisateur et ne déterminent pas ce qu'est la connaissance pertinente, assistent en revanche ce dernier et l'aide à « voir du nouveau » comme à « penser autrement » [Charlet, 2004]. Par exemple, dans le cadre d'une activité de veille technologique, elle fournit les éléments, conceptuels et visuels, qui permettront à l'analyste de détecter les signaux faibles du domaine surveillé.

Bien entendu, dans son parcours de recherche et de lecture, il est possible que les documents obtenus ne correspondent pas à l'objectif de l'utilisateur, et que ce dernier n'atteigne pas tout de

suite l'information désirée. Auquel cas, l'échec constaté est pris en considération et l'activité est corrigée en fonction de cet échec. La recherche peut alors être relancée avec d'autres critères de sélection, de manière à naviguer différemment, dans les mêmes dimensions ou dans des dimensions différentes, mais en éliminant et/ou en rajoutant des documents correspondant aux critères. L'acquisition des connaissances dont il a besoin peut donc passer par un apprentissage par essais et erreurs de la part de l'utilisateur, dans une démarche heuristique, en accord avec le principe d'action intelligente (voir la sous-section 2.3.2 page 54).

Les critères de sélection sont donc les clés d'interprétation qui guident l'utilisateur lors de la construction de son parcours. Ces clés d'interprétation sont ensuite matérialisées par les instances des types d'informations fournies par les documents. Ainsi que nous l'avons mentionné, les organisations publiant les documents représentent un type d'information parmi d'autres, soit un critère de sélection. L'organisation *Université de Grenoble* publiant un document, par exemple, est donc une instance du type « organisation ».

Concrètement, les visualisations dynamiques peuvent par exemple prendre la forme de cartographies géographiques, répartissant ainsi les documents par pays, régions du monde ou villes, de réseaux de collaboration, répartissant les documents par auteurs ou organisations communs, etc. Les figures 2.5 page suivante et 2.6 page 62 montrent des exemples de ces vues potentielles.

FIGURE 2.5 – Exemple de projection cartographique : répartition des brevets par pays à l'aide de Google Earth

La figure 2.5 présente une visualisation des brevets dans le domaine très large des nanotechnologies, répartis par pays dont les organisations qui les déposent sont originaires. La figure 2.6 montre un réseau de collaboration, créé grâce à une répartition des documents d'un ensemble par

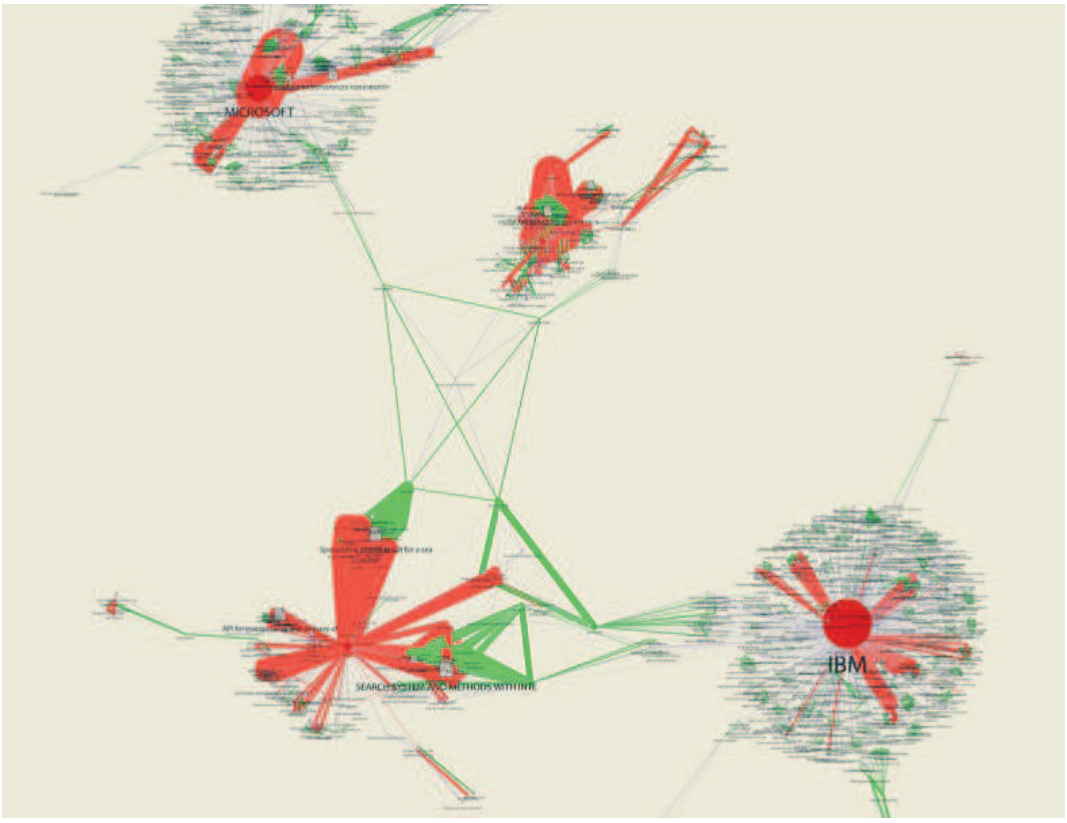

FIGURE 2.6 – Exemple de projection cartographique : réseau de collaboration à partir des organisations co-publiantes

organisations. De là, il est possible de voir quelles sont les organisations qui ont collaboré sur des dépôts de brevets ou sur des publications d'articles. Le type d'image ou d'illustration présentée influant sur la performance de compréhension d'un sujet en fonction du type de tâche qu'il doit réaliser [Tricot, 2003], les visualisations doivent être adaptées aux types de recherche d'information. En l'occurrence, une carte géographique est pertinente pour présenter une répartition des documents par pays ; un réseau est plus adapté pour mettre au jour les liens entre organisations à travers leurs collaborations.

Puisqu'il s'agit de visualisations dynamiques, c'est-à-dire d'hyperdocuments, chaque point ou élément sur la carte est cliquable, et mène à un document ou à un sous-ensemble de documents, avec d'autres visualisations dynamiques ou des vues directes sur la matérialité du document. Nous sommes consciente que le risque de ce type de représentations hyperdocumentaires est une surcharge cognitive induisant de plus faibles performances que des représentations linéaires [Tricot, 2003]. C'est pourquoi, comme nous l'avons expliqué, seuls les types d'informations utiles font l'objet de points d'accès pour la représentation de l'ensemble documentaire, de manière à limiter cet impact négatif et la possibilité de « perdre » l'utilisateur dans son immersion. Sur ces

deux figures, nous constatons que les clés d'interprétation sont bien matérialisées par les instances des types d'information : les pays d'origine des organisations sont les instances prégnantes pour la première carte, tandis que la seconde repose principalement sur les organisations elles-mêmes et les relations qu'elles entretiennent. L'une ou l'autre carte peut être utilisée en fonction de l'objectif d'information de l'utilisateur. Ce dernier peut ensuite accéder aux documents d'origine eux-mêmes, puisque c'est le moyen le plus pertinent de fournir à l'utilisateur l'information qu'il recherche. En effet, le document original induit des traitements cognitifs propres à sa forme elle-même [Charlet, 2004]. Dans notre cas, dans le cadre d'une activité de conseil en innovation, un brevet d'invention n'est pas toujours abordé et lu de la même manière qu'un article scientifique, ne serait-ce parce que les implications juridiques ne sont pas les mêmes.

2.4.2 Kartoo : la navigation cartographique

Ce type de représentation des résultats de requêtes en recherche d'information n'est pas sans rappeler, entre autres, le méta-moteur de recherche KartOO, qui présentait les résultats sous forme de cartographies. A partir d'une requête utilisateur, KartOO récupérait les résultats correspondants dans des moteurs de recherche comme Google ou Yahoo !, et les présentait à l'utilisateur sous forme de carte. Le parti-pris des concepteurs était de faire appel au « cerveau droit » de l'utilisateur, grâce à des présentations « visuo-graphiques » [Baleydier, 2004], voire « visuo-spatiales ». Cela permet à l'utilisateur d'embrasser et de comprendre un grand nombre d'informations en un temps très limité (*ibid.*). L'interface de résultats de KartOO se présentait comme dans la figure 2.7, ici à partir de la requête « moteur de recherche » [15].

Sur cette interface, les résultats sont donc présentés sous forme d'un réseau cartographié, où les sites ou pages de sites renvoyés comme résultats sont positionnés les uns par rapport aux autres, plus ou moins proches visuellement en fonction de leur proximité « sémantique ». De plus, sur la partie gauche de l'écran sont placées des « thèmes », comme *services* ou *actualité*, à partir desquelles il est possible de cibler sa recherche, et ainsi de se focaliser sur certains documents ou sous-ensembles de documents. Ces thèmes permettent en effet d'établir des liens entre documents, et donc de les regrouper.

Ce méta-moteur était novateur dans le sens où intégrer une interface de visualisation cartographique permettait de présenter autrement l'information, et d'induire chez les utilisateurs des traitements cognitifs différents de ceux mis en place pour la lecture de listes plates. Cependant, KartOO n'existe plus aujourd'hui. Les raisons de cet échec sont en partie financières, puisque la société a mal résisté à la crise économique.

Cependant, selon [Jdey, 2010], il existe également des raisons imputables à un positionnement

15. La société KartOO ayant fermé ses portes en janvier 2010, et le méta-moteur n'étant donc plus disponible, nous n'avons pu réaliser nous-même de copie d'écran. Nous empruntons donc cette image au site Abondance (`http://actu.abondance.com/2010/01/kartoo-cest-fini.html`)

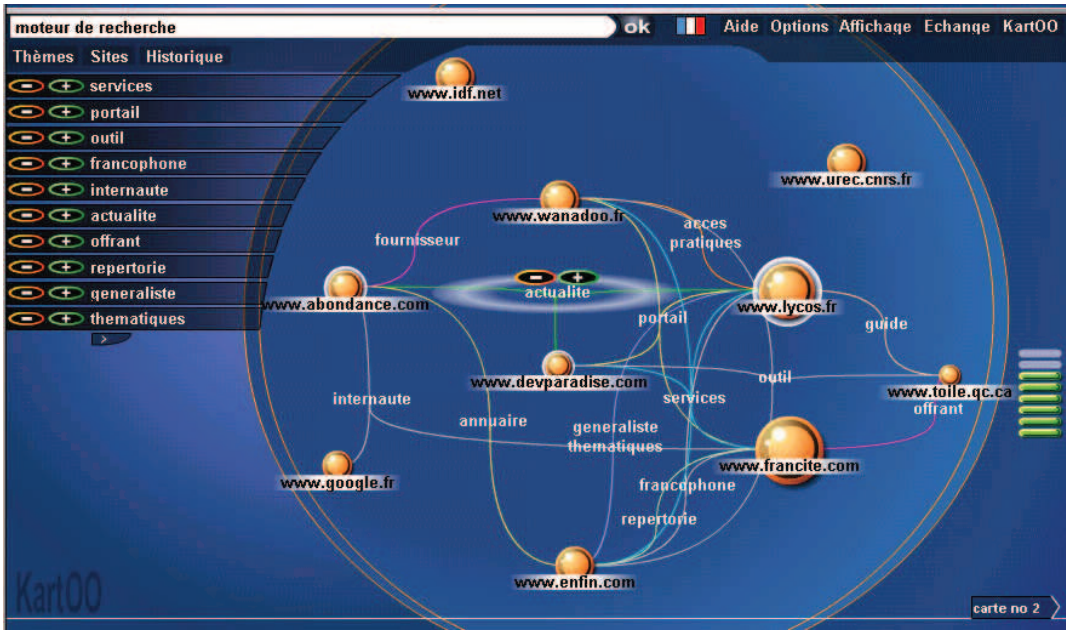

FIGURE 2.7 – Exemple de résultat de recherche du méta-moteur KartOO

commercial manquant de clarté. KartOO était en effet positionné sur deux segments : la société proposait une offre professionnelle, destinée aux entreprises spécialisées dans la veille, la recherche d'entreprise ou la cartographie. Or, d'après [Jdey, 2010], cette offre n'était pas assez spécialisée et manquait de pertinence pour une exploitation industrielle.

De l'autre côté, KartOO était aussi un méta-moteur grand public, censé toucher l'ensemble des utilisateurs d'Internet. L'offre s'adressait donc à un public très large et non spécialiste. Or, dans ce cas précis, l'interface nécessitait un apprentissage minimum : d'une part, un utilisateur désireux de se servir de KartOO devait d'abord apprendre à changer ses habitudes, généralement prises sur des moteurs comme Google, Yahoo ! ou Voilà. D'autre part, puisque les traitements cognitifs ne sont pas les mêmes entre une liste plate et une représentation cartographique, les premières utilisations peuvent manquer de confort, et il a pu être difficile de fidéliser une clientèle sans formation, même rapide, au préalable. En effet, [Jdey, 2010] appuie en particulier sur le manque d'ergonomie de l'interface, qui ne permettait pas une appropriation immédiate de l'outil.

Plutôt qu'un manque d'ergonomie de surface de l'interface, nous y voyons plutôt une cible mal définie : il semble en effet difficile de demander à un utilisateur lambda, formaté à l'utilisation de moteurs de recherche plus classiques, de passer du temps à se former lui-même à un outil alors qu'il a par ailleurs un accès immédiat aux informations dans un format qu'il connaît. Il est probable que, concernant les solutions d'entreprises, le problème ne se soit pas situé sur le même plan, c'est-à-dire sur les difficultés d'appropriation de l'interface et de la présentation des informations. Cependant, il est à l'heure actuelle très difficile d'obtenir de l'information sur les

anciennes offres professionnelles de KartOO, et c'est pourquoi nous nous garderons de conclure sur ce sujet.

Quoi qu'il en soit, il semble que le problème majeur de l'offre grand public relève avant toute chose d'un manque d'acceptabilité et/ou d'utilisabilité (cf. 2.2.3.2 page 44), pour les raisons que nous venons d'évoquer : ce système allait à l'encontre des habitudes des utilisateurs. Or, ceux-ci ne prenaient pas forcément le temps de s'adapter à ce nouveau moteur puisque d'autres, qu'ils connaissaient et savaient manipuler, se trouvaient déjà implantés de longue date dans leur environnement.

De plus, et concernant le manque d'utilisabilité de cet outil, il semble possible que la nature des liens servant à naviguer dans l'ensemble des résultats puisse elle aussi être mise en cause. Nous avons vu ci-dessus (voir page précédente) que ces liens sont en fait des thèmes communs à plusieurs sites ou pages de sites renvoyés en résultats. La liste en est reprise à gauche de l'écran, ainsi que nous l'avons déjà signalé. Or, à regarder cette liste de plus près, nous trouvons des thèmes tels que « offrant » ou « répertorie ». Bien que nous n'ayons pu accéder au détail de la technologie utilisée par KartOO, il nous semble probable que ces « thèmes » soient en réalité des mots récurrents et communs à plusieurs documents. Si certains d'entre eux peuvent à la rigueur être représentatifs des thèmes des documents, comme « actualité », « services » ou « portail », bien que ces mots soient trop larges pour être réellement porteurs de sens, il est difficile de considérer que « offrant » ou « répertorie » soient des thèmes.

De ce point de vue, la technologie KartOO ne semble donc pas reposer sur des traitements sémantiques, mais tout au plus sur des traitements statistiques, dont sont d'ailleurs exclus les mots complexes, c'est-à-dire les mots composés de plusieurs suites de caractères alphabétiques séparées par des caractères d'espacement. Ainsi, les clés d'interprétation fournies par KartOO restent limitées, et ne reposent pas sur des catégories sémantiques claires et définies. Le guide permettant de construire un parcours de recherche et de lecture n'est donc pas totalement fiable. Enfin, l'utilisateur n'est pas entièrement intégré au système, dans le sens où malgré une navigation par thèmes, par sites ou par dates, ses choix et son potentiel d'actions sont limités, et il n'est pas considéré comme composante à part entière du système d'information.

Si notre approche s'inspire de celle de KartOO, elle s'en distingue de trois manières :

- tout d'abord, nous mettons du sens dans les liens qui unissent les documents, qui va au-delà de mots-clés communs ;
- d'autre part, les objets que nous traitons, c'est-à-dire les documents électroniques aux genres figés et contraints que sont les brevets d'invention et les articles scientifiques, offrent des possibilités d'analyse systématique plus importantes que des pages web et documents dont les genres peuvent être extrêmement variés ;
- enfin, le public visé par notre système d'immersion est bien différent de celui visé par l'offre grand public de KartOO.

La mise en place de clés d'interprétation, et de liens entre ces dernières, est un moyen de guider l'utilisateur au sein d'un ensemble documentaire afin qu'il construise son propre parcours d'interprétation. Cependant, pour que ce « guidage » soit efficace, une simple mise en avant des mots communs à plusieurs documents n'est pas optimale, car elle n'est pas nécessairement représentative du sens de ces documents. Nous jugeons que pour faire émerger le sens des liens entre les documents, et par là le sens des documents eux-mêmes, il est nécessaire d'intégrer à notre système d'immersion une couche de connaissances allant au-delà d'une liste de mots. Ainsi, nous avons mis en place un système de facettes, où chaque facette représente un type d'information véhiculé par le document (nous avons expliqué ce que nous entendions par *type d'information* en 2.2.3.1 page 38).

Ce système en facettes a deux avantages : tout d'abord, il permet de circonscrire précisément le contenu de chaque facette, et donc de clairement définir le type des informations qui s'y trouvent. Par exemple, une facette contient les noms d'organisations qui publient un document, et seulement cela. De cette façon, il est plus simple pour l'utilisateur d'y accéder lorsque c'est ce type d'information qui l'intéresse. D'autre part, puisque le contenu de chaque facette est bien défini, typer les liens de manière systématique devient parfaitement réalisable. Par exemple, une instance de la facette des organisations publiantes peut être reliée, par le biais d'un ou plusieurs documents, à une instance de la facette des auteurs ou inventeurs en tant que personnes individuelles, par une relation « organisation *emploie* auteur ou inventeur » ou sa réciproque « auteur ou inventeur *travaille pour* organisation ». Ce n'est donc plus un mot-clé qui unit deux documents, mais une véritable relation sémantique. Ces instances de facettes et ces liens deviennent alors un réseau riche, comportant autant de dimensions que de types d'information.

L'ensemble des clés d'interprétation, auquel s'ajoute l'ensemble des liens unissant ces clés, est donc à considérer comme un ensemble de connaissances potentielles, des inscriptions numériques au sens de [Charlet, 2004]. Elles permettent à l'utilisateur d'interpréter les données qui lui sont fournies par l'ensemble documentaire et de construire son parcours de lecture, et par là ses connaissances pour son objectif.

Injecter du sens dans notre système d'immersion est rendu possible grâce à la nature des documents que nous traitons, et au fait que nous connaissons leur genre à l'avance : nous avons rapidement mentionné (dans la section 2.2.3.1 page 38) que les brevets d'invention et les articles scientifiques sont deux genres contraints et très figés, limitant ainsi la variation linguistique et structurelle des documents. De fait, des régularités existent qui nous permettent de dégager un certain nombre de types d'informations se retrouvant dans tous les éléments d'un ensemble documentaire, fournissant ainsi les éléments de sens exploitables. Grâce à la connaissance du contexte d'application, il est possible de dégager à partir de ces régularités les types d'information qui fourniront les meilleures clés d'interprétation pour guider l'utilisateur dans la construction de sa connaissance, sans le surcharger d'informations. Sur cette base, les facettes peuvent être

constituées.

Notre système d'immersion est conçu pour un public restreint, spécialisé, et surtout, formé à ce type d'outils. Nous avons montré en 2.2.3.1 page 38 que les individus amenés à travailler sur un tel système d'immersion utilisent à l'heure actuelle un outil de recherche dans lequel il existe déjà un certain nombre de visualisations cartographiques. Ce type de représentations visuelles de l'information est donc d'ores et déjà ancré dans leur pratique professionnelle, bien que les visualisations soient à l'heure actuelle peu dynamiques. Par ailleurs, il existe une telle nécessité d'utilisation d'outils de recherche d'information dans leur métier qu'ils deviennent rapidement experts du système utilisé. Enfin, c'est justement cette représentation visuelle de l'information qui leur permet de dégager l'information utile, ou pertinente, pour leur objectif. L'acceptabilité du principe de visualisation de ce type ne pose donc pas problème, dans la mesure où d'une part elles s'intègrent naturellement à leur pratique déjà en place, et d'autre part répondent à des besoins clairement conscients et exprimés.

Ces utilisateurs sont donc *a priori* compétents dans l'utilisation d'un tel système d'immersion, qui leur permettra d'effectuer une tâche récurrente : ils deviendront donc rapidement experts dans cet usage. De même, ces analystes sont déjà habitués à certaines manipulations qu'ils seront amenés à réaliser dans le système, et sont conscients de l'apport d'un tel outil par rapport à des visualisations par listes plates par exemple. En somme, si les documents fournissent la connaissance et le sens injectés dans le système d'immersion, les utilisateurs sont quant à eux à même d'interpréter ces connaissances et ce sens.

La question fondamentale est donc de faire émerger du sens et des connaissances à partir des liens entre les documents.

2.4.3 Injecter du sens dans l'immersion

Les connaissances potentielles, ou inscriptions numériques manipulables par la machine, sont intégrées à notre système d'immersion, afin d'en faire un système à base de connaissances [Charlet, 2004]. Afin d'être exploitables informatiquement, mais aussi compréhensibles par l'humain qui utilise l'outil, les clés d'interprétation et leurs liens doivent donc être intégrés à une structure de représentation formelle des connaissances, mais toujours en lien avec leur contexte d'utilisation. En somme, les informations tirées des documents doivent être reliées à leur document source, de manière à en laisser l'accès permanent aux utilisateurs. De cette façon, les utilisateurs, qui représentent un public spécialisé et sont formés à l'outil, sont capables de dégager le fonctionnement de ce dernier, grâce à ces clés d'interprétation. La représentation des connaissances est donc à la fois un outil de structuration de l'information, et un outil de lecture de ces informations.

La structure de représentation des connaissances choisie doit donc rassembler l'ensemble

des caractéristiques nécessaires à l'actualisation des connaissances pertinentes pour l'utilisateur ayant un objectif donné, et à l'accès au document en tant qu'élément unique d'un ensemble documentaire. Encore une fois, il s'agit donc de naviguer du global au local, et inversement. Nous avons expliqué en 2.4.1 page 57 qu'un système utile dans notre contexte doit être multidimensionnel, avec autant de dimensions que de types d'information, qui constituent les clés d'interprétation dans l'utilisation du système, afin de pouvoir croiser ces dimensions entre elles pour la recherche d'information. Il est important d'isoler ces différentes dimensions dans la représentation des connaissances, tout en établissant des liens entre leurs instances. En effet, en donnant à l'utilisateur des liens entre les connaissances, ce dernier peut alors construire le sens par une sélection délibérée de certains d'entre eux. La séparation entre les différentes dimensions permettra de rentrer des critères de recherche distincts, et les liens entre leurs instances permettront à l'utilisateur de visualiser à la demande les relations unissant ces instances. Ces liens et clés font l'objet d'un engagement sémantique, c'est-à-dire d'une spécification formelle contraignant le sens d'unités linguistiques [Bachimont, 2000], dont le niveau sera certes restreint, puisque limité à quelques dimensions, mais sûr.

L'injection de connaissances potentielles est donc fondamentale dans notre système d'immersion, pour garantir la pertinence des liens choisis par l'utilisateur. La question à résoudre est alors de savoir quelle structure sera la plus adaptée à notre cadre d'application, en fonction des données, de leur type, et de l'utilisation qui en sera faite.

Chapitre 3

Une représentation des connaissances pour l'immersion : exploiter les connaissances thématiques et contextuelles

Sommaire

Un concept est une invention à laquelle rien ne correspond exactement, mais à laquelle nombre de choses ressemblent. Friedrich Nietzsche

Notre objectif est de proposer des modèles pour un système d'immersion documentaire dans des domaines scientifiques et technologiques. Pour obtenir un système d'immersion efficace, il apparait fondamental de le fonder sur un modèle permettant de représenter un certain nombre d'informations ou connaissances contenues dans les documents auxquels l'utilisateur souhaite accéder. Cette ressource est ainsi un intermédiaire entre l'immersion et la navigation d'une part et les documents eux-mêmes d'autre part, en tant qu'unités de sens. Dans l'idéal, elle permet un accès aux documents qui est adapté à l'information qu'ils peuvent véhiculer, mais qui est également en adéquation avec l'utilisation qui est faite du système.

Les composants de l'ensemble documentaire, c'est-à-dire les brevets d'invention et les articles scientifiques, appartiennent à des genres extrêmement codifiés et contraints. Il est fondamental de prendre en compte ces deux genres pour la constitution d'une ressource, ainsi que l'objectif du message qu'ils véhiculent. Ainsi, la ressource doit cibler en priorité les caractéristiques essentielles de ces documents et de leur message, du point de vue de l'émetteur comme de celui des récepteurs [16]. Pour cela, nous considérons qu'il convient de rendre l'utilisateur partie prenante de la constitution de la ressource.

Les ressources utilisées pour la représentation des connaissances ou des informations relèvent classiquement de ressources comme des ontologies, des terminologies, des thésaurus, etc. Tous ces modèles de représentation ont en commun de mettre en jeu des réseaux de termes pour modéliser le contenu des documents et permettre ainsi un meilleur accès à la connaissance [Aussenac-Gilles & Condamines, 2004]. Toutes ces représentations ne sont cependant pas semblables, puisque leur degré de structuration et de formalisation varie d'un type de ressource à un autre (*ibid.*). Cependant, puisqu'elles impliquent toutes l'usage de termes et de relations entre ces termes, nous les rassemblons, à la suite de [Aussenac-Gilles & Condamines, 2004] ; [Bourigault & Aussenac-Gilles, 2003], sous le terme de ressources termino-ontologiques (RTO).

Ces auteurs, membres du groupe de travail TIA, mettent par ailleurs en avant le fait que les textes sont à la fois des sources pour la construction de RTO, et des sources directes de connaissance pour l'utilisateur. Cette exploitation de données textuelles pour le développement de RTO est un des points centraux des travaux du groupe TIA. La vision sous-tendue par cette approche est que les textes sont une source objective de données, ces dernières véhiculant des connaissances relativement stabilisées, partagées et consensuelles. De ce fait, les documents mêmes de l'ensemble documentaire que l'utilisateur doit explorer peuvent être à l'origine de la structure de représentation qui aidera à cette exploration. Cette méthode garantit l'adéquation de la ressource aux besoins induits par la nature des documents à explorer.

16. Nous empruntons les termes d'émetteur et de récepteur aux sciences de la communication.

3.1 Une représentation des connaissances orientée sur l'objectif de communication des documents

3.1.1 L'objectif général de communication des documents scientifiques et technologiques

Ainsi que nous l'avons établi dans le chapitre précédent, l'objectif du système d'immersion est de fournir un accès à des documents scientifiques et technologiques. Plus précisément, ces documents sont tous des brevets d'invention et des articles scientifiques. Ces deux genres sont relativement différents, puisque le brevet relève autant du document juridique que du document scientifique, alors que l'article scientifique s'intéresse exclusivement à des composants scientifiques. Cependant ces genres, très codifiés par ailleurs, ont des objectifs de communication sensiblement comparables. Nous pouvons énumérer ces objectifs comme suit :

- énoncer et prouver une relation de propriété entre le contenu scientifique et/ou technologique et son auteur ;
- fournir une information scientifique et/ou technologique aux lecteurs des documents issus de ces genres ;
- dans le cas des brevets d'invention en particulier, fournir également une information juridique au sujet du contenu scientifique et technologique du document.

Or, selon [Strawson, 1970], qui prônait alors le glissement de la linguistique « traditionnelle », de la langue en tant que système, à une linguistique du discours prenant comme objet d'étude des unités allant au-delà de la phrase, les objectifs de communication d'un discours sont fondamentaux. D'après lui, « le contexte d'un énoncé affecte ce qu'on dit », et il est par ailleurs impossible de « comprendre le discours si nous ne tenons pas compte du but de communication ».

Par conséquent, dans notre processus de conception, il est nécessaire de prendre en compte cet objectif de communication, ainsi que son contexte, pour être capable de fournir aux destinataires de documents les clés permettant leur interprétation.

Nous reprenons le schéma de la communication verbale de [Jakobson, 1963] dans la figure 3.1 page suivante.

Un message n'existe pas « dans le vide » : il prend place et prend vie entre un destinateur, qui « émet » ce message, et le destinataire, qui le « reçoit ».

Cette transmission du message prend place dans un contexte donné, dans une situation de contact (ou de non contact) particulière, et utilise un code de communication.

C'est l'ensemble de ces facteurs qui définit donc un acte de communication verbale. Ce schéma a été maintes fois critiqué, du point de vue de son contenu, de la pertinence de la notion de code, ou encore de son exhaustivité. La polémique a en particulier porté sur le fait que le code, en l'occurrence la langue, dans ce schéma, était envisagé comme parfaitement univoque, alors qu'il est

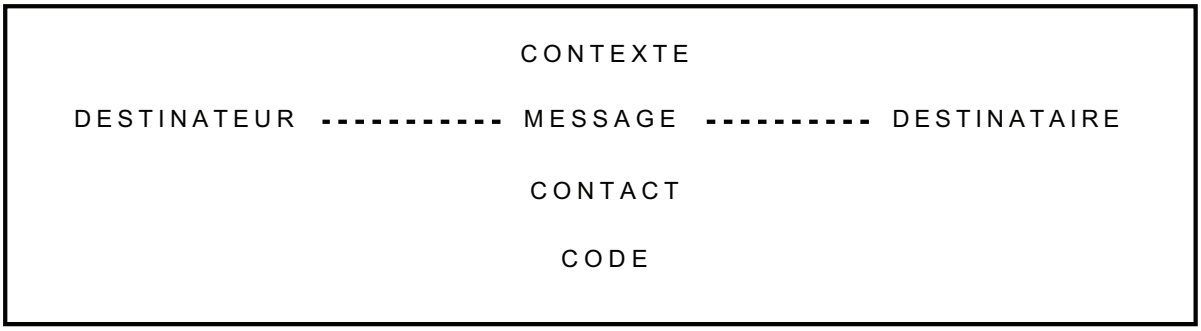

FIGURE 3.1 – Schéma de la communication de Jakobson [Jakobson, 1963]

par essence ambigu. D'autre part, certains détracteurs ([Ducrot, 1972] en particulier) rejettent la communication comme un simple échange d'informations, délaissant par là la dimension pragmatique, extra-linguistique, des actes de langage pour eux-mêmes. Nous ne rentrerons pas dans les détails de la polémique suscitée par ce schéma pendant de nombreuses années. Nous reprenons par contre le schéma qu'a proposé [Kerbrat-Orecchioni, 1980], qui vient compléter, plus qu'il ne remplace, celui de Jakobson. Nous le reproduisons dans la figure 3.2.

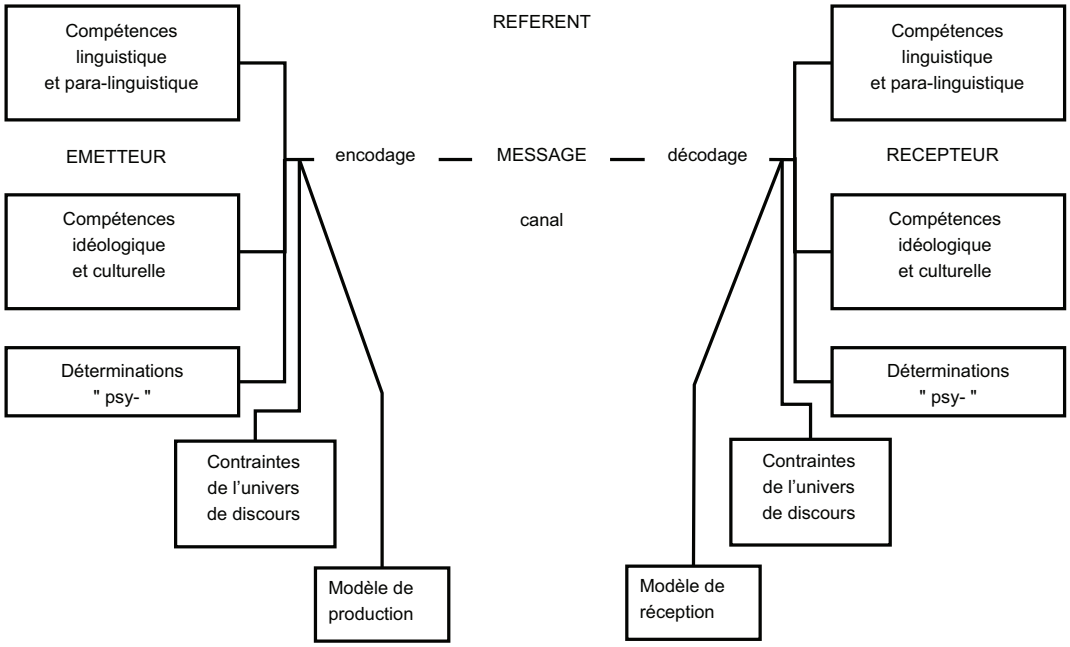

FIGURE 3.2 – Reformulation du schéma de Jakobson par [Kerbrat-Orecchioni, 1980]

Beaucoup d'éléments sont rajoutés ici ; ils permettent d'embrasser plus largement un acte de communication et les facteurs impliqués. Certains facteurs, extérieurs au message linguistique lui-même, nous paraissent particulièrement pertinents. Il est visible sur le schéma que les contraintes véhiculées par l'univers de discours, ainsi que les modèles de production et d'interprétation, sont

pour beaucoup dans l'encodage et le décodage du message lui-même. Nous développons donc ces deux points plus particulièrement.

L'univers de discours Pour [Kerbrat-Orecchioni, 1980], l'univers de discours est très complexe, et réunit deux types d'éléments.

- Les données situationnelles, tout d'abord, concernent la nature du locuteur et des allocutaires, et les conditions de production du point de vue matériel, social, etc. Dans ce cadre, ces données prennent la forme d'images, qui vont quelque peu contraindre la communication. L'émetteur comme le récepteur du message ont une image d'eux-mêmes et de leur partenaire discursif leur permettant de se positionner dans la communication l'un par rapport à l'autre. Les questions de légitimité de l'émetteur et du récepteur se situent donc ici.
- Les caractères thématiques et rhétoriques, d'autre part, imposent des contraintes, généralement matérialisées par le « genre » dans lequel un message s'inscrit.

Les modèles de production et d'interprétation Ces modèles sont constitués, d'après Kerbrat-Orecchioni, des connaissances que les sujets ont sur leur langue, et qu'ils mobilisent pour un acte énonciatif donné, d'émission ou de réception. A partir de ces connaissances, ils « font fonctionner des règles générales qui régissent les processus d'encodage et de décodage » (*ibid.*). Rentrent aussi en jeu, dans ces modèles, les facteurs relatifs aux compétences culturelle et idéologique, et les données situationnelles disponibles dans l'univers de discours.

Ainsi, communiquer met en jeu un ensemble de paramètres qui dépassent largement le cadre linguistique proprement dit. Les données de natures diverses gravitant autour du message sont à prendre en considération pour pouvoir l'interpréter correctement. Ainsi, « on ne peut décrire un message sans tenir compte du contexte dans lequel il s'enracine, et des effets qu'il prétend obtenir » (*ibid.*). Les « effets » vers lesquels tendent le message dépassent bien souvent le caractère purement informationnel de ce dernier. Ils s'inscrivent dans sa dimension pragmatique, et lui confèrent une valeur illocutoire.

La dimension pragmatique, en effet, concerne les relations entretenues entre les signes et leurs usagers [Morris, 1938]. Un message peut être considéré comme un signe, et en l'occurrence, les usagers en sont l'émetteur et le récepteur. La pragmatique, en tant que discipline, étudie donc l'énonciation, par exemple chez Morris, mais aussi les « actes de langage » (terme emprunté à [Searle, 1969]), en tant qu'ils permettent d'avoir un effet sur le monde extra-linguistique. Kerbrat-Orecchioni résume ainsi la pragmatique illocutoire :

> « [...] l'hypothèse fondatrice [en] est la suivante : parler, c'est sans doute échanger des informations ; mais c'est aussi effectuer un acte, régi par des règles précises [...], qui prétend transformer la situation du récepteur, et modifier son système de croyances

> et/ou son attitude comportementale ; corrélativement, comprendre un énoncé c'est identifier, outre son contenu informationnel, sa visée pragmatique, c'est-à-dire sa valeur et sa force illocutoires. » [Kerbrat-Orecchioni, 1980]

C'est donc bien le rapport du message au monde extra-linguistique qui est au cœur des deux approches de la pragmatique. Dans le premier cas, cette dernière s'attache à relier le message à sa situation d'énonciation ; dans le second, elle porte sur l'effet de ce message sur le monde. Dans tous les cas, il est question d'une interaction forte entre la communication verbale et le non verbal qui l'entoure. La pragmatique considère donc que la communication verbale ne vaut, et sa signification ne peut être comprise, que si les conditions extra-linguistiques sont prises en compte, soit pour l'interprétation du message même, soit pour en percevoir les effets.

Un énoncé, ou message, a donc un contenu sémantique, ou propositionnel, qui concerne ce qui est dit, et une ou des valeurs illocutoires, qui ont à voir avec l'objectif de cet énoncé ou message. C'est la combinaison de ces deux types de valeurs qui permet la compréhension totale d'un message.

Si nous revenons aux genres de l'article scientifique et du brevet d'invention, et si nous reprenons les objectifs de communication communs à ces deux genres que nous avions identifiés plus haut, nous pouvons désormais préciser et attribuer un statut à chacun de ces objectifs, et donc à chacune des valeurs illocutoires en jeu. Nous les avons formulés comme suit :

1. prouver une relation de propriété entre le contenu scientifique et/ou technologique et son auteur ;
2. fournir une information scientifique et/ou technologique aux lecteurs des documents issus de ces genres.

Ils correspondent clairement aux valeurs illocutoires suivantes :

1. la preuve de la propriété d'une avancée scientifique et/ou technologique relativement à son auteur ;
2. l'information sur cette avancée.

Ainsi, pour ces deux genres, la valeur illocutoire est (au moins) double : prouver et informer ; quant au contenu propositionnel, ou sémantique, il s'agit d'un contenu scientifique et/ou technologique.

Les documents issus de ces genres ont donc pour objectif de véhiculer de l'information considérée comme objective, relative à un contenu scientifique ou technologique souvent pointu. Mais au-delà de ce contenu sémantique riche, ils permettent aussi d'exprimer le fait que le contenu du document « appartient » à son auteur. Cette relation de propriété peut être résumée par l'expression *X a fait Y*, et par sa réciproque *Y a été fait par X*, où *X* est l'auteur du document, et *Y* est le contenu du document. Dans les deux cas, cette information relève de la dimension pragmatique inhérente à tout acte de communication.

Selon le point de vue choisi, l'expression ou sa réciproque pourra être privilégiée : si l'auteur est l'information primordiale, alors l'expression *X a fait Y* est la plus pertinente ; en revanche, si le point de focalisation porte sur le contenu, alors la réciproque *Y a été fait par X* est plus appropriée.

De manière générale, la situation d'énonciation entourant ces documents est toujours sensiblement la même. Tout d'abord, c'est le canal de l'écrit qui est utilisé - le produit de la communication est donc un document écrit. Par ailleurs, tout document publié est par essence publié par une personne au moins, physique ou morale, dans un lieu et à une date donnés. Ces paramètres sont des données de la situation d'énonciation. Puisque l'une des valeurs illocutoires des articles scientifiques et des brevets d'invention est la « preuve » de propriété, ces informations revêtent une importance particulière. L'objectif du côté de l'émetteur du message est de faire savoir aux destinataires qu'un contenu scientifique précisément déterminé a été produit par lui ; l'objectif du côté du récepteur, qu'il soit chercheur, juriste, veilleur, etc., est de savoir que ce contenu scientifique a été produit par l'émetteur, ou que l'émetteur en question a produit un contenu scientifique déterminé [17]. Le fait qu'un document porte sur un thème particulier est donc lui aussi crucial, et nous touchons ici à la seconde valeur illocutoire du message : la valeur d'information. Cette valeur est étroitement liée, de fait, avec le contenu propositionnel lui-même du document.

En l'occurrence, les informations fondamentales à retenir dans ce message sont donc l'identité de l'émetteur d'une part, soit le *X* de l'expression, et la nature du contenu produit, c'est-à-dire le *Y* de l'expression, d'autre part. A cela s'ajoutent les informations de lieu et de date, qui sont nécessaires à la situation du message dans l'espace-temps. Sans cette situation, et en particulier la situation temporelle, les documents produits auraient peu de valeur, en particulier en ce qui concerne la valeur illocutoire de preuve, puisqu'il serait alors impossible d'attribuer rigoureusement la propriété d'un travail scientifique à un émetteur en particulier. Pour les brevets, les revendications et le marquage d'antériorité sont là pour asseoir cette propriété juridiquement. En ce qui concerne les articles scientifiques, il est communément admis que le premier à publier un contenu scientifique donné en est le propriétaire, sinon du point de vue légal, au moins du point de vue moral.

Il est à noter que la seule distinction que nous faisons au sein des documents est celle qui concerne le genre des textes : les brevets sont distingués des articles scientifiques. Dans leur contexte d'utilisation, le même traitement est globalement réservé à ces deux genres : ils font l'objet d'une lecture, puis leur contenu est évalué par le lecteur. Du point de vue de l'utilisateur, la seule différence majeure réside, nous l'avons mentionné, dans la valeur juridique des brevets, qui n'est pas partagée par les articles scientifiques. D'autres divergences d'ordre stylis-

17. Les catégories de récepteurs que nous citons ici sont parmi les plus courantes et sont des cas prototypiques pour lesquels l'émetteur produit le document. Les analystes de TKM représentent un petit sous-ensemble de cette cible prototypique. Il peut cependant exister, bien entendu, des détournements de ces documents pour des utilisations autres. Par exemple, nous-même les utilisons dans le cadre de nos travaux pour les modéliser.

tique, argumentatif ou terminologique existent entre ces deux genres ; cependant, elles ne sont pas prises en compte comme telles par les utilisateurs, qui de plus travaillent très majoritairement sur les résumés. Malgré leurs différences, les documents sont donc considérés comme des objets similaires.

Nous aurions pu affiner la distinction binaire entre brevets et articles par la prise en compte de mesures comparatives. Par exemple, le facteur d'impact de Thomson Reuters [18] pour les articles, ou des mesures de *patent rating* pour les brevets, auraient pu être exploités pour organiser les documents en fonction de leur score. Cependant, ces mesures sont intrinsèques à chacun des genres de documents, et il n'existe pas, à notre connaissance, de mesure commune permettant d'établir des distinctions et des recoupements inter-genres de ce point de vue.

Nous prenons donc le parti de nous en tenir aux paramètres de la situation de production des documents que nous avons déjà énoncés, et qui sont communs aux documents des deux genres : l'auteur ou l'inventeur, l'organisation propriétaire d'un document, la localisation géographique de l'organisation, la date de publication, et enfin les thèmes des documents.

3.1.2 Le document comme medium et le document comme signe : prise en compte du contexte de production des documents pour la représentation des connaissances

La double facette du document est prise en compte dans le message porté par les articles scientifiques et les brevets d'invention : dans les deux cas, le document est à la fois considéré comme signe et comme medium. D'une part, le document est un signe, et à ce titre, il est un contenu sémantique précis. Cette « facette » du document est donc à relier au contenu propositionnel du message qu'il communique. D'autre part, il véhicule un certain nombre d'informations qui viennent s'ancrer dans un contexte et une pratique sociale déterminés. De ce point de vue, le document a une force illocutoire, pragmatique.

Le collectif R.T. Pédauque [Pédauque, 2003] définit le document comme signe, comme forme et comme relation. Un document est un signe en tant qu'il est porteur de sens ; il est une forme en tant qu'objet matériel ou immatériel possédant une structure déterminée ; enfin, il est une relation. Cette dernière caractéristique du document, qui correspond pour nous à la notion de medium, revêt toute son importance dans le cadre d'articles scientifiques et de brevets d'invention. Elle interroge en effet la nature du document en tant que composante des relations sociales. D'après les auteurs, le document « est un élément de systèmes identitaires et un vecteur de pouvoir » (*ibid.*). Ceux-ci caractérisent le document comme relation de la manière suivante :

> « Un document donne un statut à une information, à un signe. C'est une preuve

18. Voir le site de Thomson Reuters : `http://thomsonreuters.com/products_services/science/free/essays/impact_factor/`.

> qui fait foi d'un état de choses. C'est une annonce qui prévient d'un événement. C'est un discours dont la signature le rattache à un auteur, etc. Ce statut ou cette légitimité s'acquiert sous deux conditions : le sens doit dépasser la communication intime (entre quelques personnes privées), et il doit s'affranchir de l'éphémère (dépasser le moment de son énonciation). Ces conditions impliquent que si tout signe peut être un document, un signe particulier [...] ne l'est pas nécessairement. par exemple, un journal intime n'est pas un document, sauf si quelqu'un prend l'initiative de le rendre public [...]. »

Pour être un document au sens de Pédauque, un signe doit donc être ancré dans une pratique sociale, et être public.

D'après cette caractérisation, un brevet d'invention ou un article scientifique sont des cas typiques de documents en tant que relations. Ce sont à la fois des « discours dont la signature les rattache à un auteur », et des « preuve[s] qui [font] foi d'un état de choses ». En cela, cette dimension relationnelle du document est primordiale pour ces deux genres. Il est donc fondamental de la prendre en compte pour permettre d'y accéder. En l'occurrence, il s'agit de récupérer le contexte social et pragmatique de la production du document, exprimé linguistiquement par les paramètres de la situation d'énonciation du document. Il est à noter que si un signe, pour être un document, « doit s'affranchir de l'éphémère », cela ne signifie pas qu'il faille tout bonnement évacuer les dimensions temporelles et locales de la production d'un document. Au contraire, il est fondamental de conserver les paramètres qui se rapportent à l'espace-temps de la production du document, de manière à pouvoir le situer par rapport à d'autres documents, ainsi que nous l'avons signalé un peu plus haut. Ce n'est qu'ainsi qu'il conservera l'intégralité de ses fonctions pragmatiques (prouver et informer).

La dimension du document en tant que signe est aussi à prendre en compte. La raison d'être d'un document est aussi de véhiculer un sens, un contenu sémantique. A ce titre, il est impossible de se passer de cette caractérisation si nous voulons restituer la richesse d'un document.

Enfin, le document en tant que forme, c'est-à-dire en tant qu'objet, matériel ou immatériel, fournit les moyens de sa propre exploitation. Il ne pourrait exister en tant que signe ni en tant que relation s'il n'était pas une forme placée sur un support technique (voir la sous-section 2.3.1 page 49, sur la dépendance de la connaissance, véhiculée par les documents, vis-à-vis de la technique chez [Charlet, 2004]). D'autre part, la structure même de cet objet donne des indications formelles sur les informations à sélectionner, de manière à mieux les exploiter, informatiquement comme cognitivement.

En l'occurrence, les objets que nous traitons sont des objets numériques, et plus précisément des entrées de bases de données. A ce titre, nous travaillons sur des données structurées, où les paramètres de la situation d'énonciation sont isolés dans différents champs de chaque entrée, avec,

plus ou moins systématiquement, un type de paramètre par champ. Le corps du document est lui aussi dans un champ distinct ; en revanche, le corps lui-même n'est pas structuré en différents champs. De fait, nous travaillons à la fois sur des données structurées, pour les paramètres de la situation d'énonciation et donc pour la dimension pragmatique, et sur du texte plein en ce qui concerne le contenu sémantique. Le document en tant que forme distingue donc dans notre cas les éléments se rapportant au rôle de medium de ceux relatifs au rôle de signe.

La ressource termino-ontologique (RTO) fondée sur des documents et construite pour y accéder doit donc nécessairement englober les trois dimensions qui les définissent pour être performante. La RTO dans sa globalité intègre donc des informations concernant à la fois la situation d'énonciation d'un document, et son contenu sémantique. Cette intégration est rendue possible par sa matérialité : elle permet de prendre en compte la structure formelle du document pour repérer les informations contextuelles et sémantiques qu'elle renferme.

De manière globale, le message véhiculé par les articles scientifiques et par les brevets d'invention porte notamment sur quatre types d'informations à rattacher au document comme relation, pour sa fonction illocutoire, et au document comme signe, pour son contenu propositionnel : qui a produit le contenu, quand, où, et quel est ce contenu. Les trois premiers types décrivent le contexte de production du document, et nous les nommons donc, en toute logique, les informations contextuelles de chaque document. Formellement, elles prennent la forme d'entités nommées [19]. Il est à noter que le *qui*, qui correspond à l'émetteur dans la situation d'énonciation, peut être un émetteur complexe. En ce qui concerne brevets et articles scientifiques, il s'agit en effet d'un émetteur à deux niveaux, chaque niveau pouvant comprendre plusieurs émetteurs distincts, comme ce peut être le cas pour un article écrit en collaboration entre plusieurs organisations par exemple. L'émetteur peut donc être divisé en deux catégories : d'une part l'auteur ou les auteurs, en tant que personnes, et d'autre part l'organisation ou les organisations au nom de laquelle/desquelles est publié le document, et pour laquelle/lesquelles travaille(nt), ou a/ont travaillé, l'auteur/les auteurs. Le quatrième type d'information décrit le contenu sémantique du document, et relève donc du document comme signe.

Pour être pertinent vis-à-vis des documents et de leur double objectif de communication, l'accès à ces documents doit donc se faire prioritairement à l'aide de ces informations contextuelles et sémantiques : elles sont des composantes inhérentes à l'objectif de communication de l'émetteur.

Ici, c'est le versant *émetteur* de la situation d'énonciation qui a été majoritairement pris en compte, hors de tout contexte d'exploitation. Le versant *récepteur* quant à lui n'est pourtant pas négligé : les objectifs inhérents à ces genres, et donc aux émetteurs des documents, sont tous deux pris en compte par les récepteurs de notre contexte applicatif - les analystes. C'est pourquoi la ressource sur laquelle s'appuie le système de navigation tient compte des objectifs des deux

19. Nous revenons sur cette caractérisation formelle dans le chapitre 4.

partenaires de la communication : elle doit contenir et organiser ces informations, de manière à en faire des points d'accès privilégiés aux documents.

La RTO choisie et définie restitue des informations sur le contenu du document d'une part, considérant ainsi sa dimension sémantique, et d'autre part sur l'ensemble des paramètres de la situation d'énonciation, ce qui revient à prendre en compte sa dimension pragmatique. Il est en effet fondamental de considérer les documents comme la combinaison de ces deux dimensions : laisser l'une ou l'autre de côté reviendrait à perdre des moyens d'accès incontournables à ces documents, limitant ainsi les possibilités de compréhension des messages qu'ils portent.

De plus, c'est le versant *récepteur* qui, en orientant la représentation des connaissances, détermine la structure formelle des informations de la RTO finale.

3.2 Une représentation des connaissances orientée sur l'objectif d'exploitation des documents scientifiques et technologiques

3.2.1 L'application finale comme critère de définition de la RTO

Rappelons que l'application finale, pour laquelle nous avons construit la structure de représentation des connaissances, est un système d'immersion documentaire. L'objectif est de permettre à l'utilisateur d'accéder à ces documents afin qu'il obtienne l'information pertinente pour son propre objectif. C'est alors à lui d'interpréter l'information qui lui est présentée par le biais des documents, en fonction de leur contenu et/ou en fonction de son contexte de production. Nous cherchons donc à donner accès à l'information, non pas à la dissocier de son contexte ni à l'interpréter.

Cet accès doit répondre aux contraintes imposée par le milieu professionnel dans lequel le système d'immersion prend place : il sera utilisé par des analystes exerçant dans une société de conseil en innovation. Nous avons constaté dans le chapitre 1 page 9 que cette activité professionnelle « globale » recouvre en réalité une pluralité d'activités de conseil non entièrement prévisibles à l'avance : des états de l'art, des études de marché, des recherches de partenaires scientifiques, etc. De plus, les domaines couverts sont nombreux. Ainsi, pour une seule application du système existent une variété de besoins.

La ressource doit couvrir l'ensemble des besoins, ce qui revient à dire que la ressource doit fournir autant de points d'entrée que nécessaire pour donner à l'utilisateur une vue pertinente des documents sur lesquels il travaille pour son objectif, et ce dans un grand nombre de domaines scientifiques différents.

En considérant comme prioritaire la prise en compte de l'objectif et du contexte de l'application, nous adoptons le point de vue défendu par un certain nombre de chercheurs français dans le domaine de la représentation des connaissances. Plus précisément, nous suivons en cela

la position du groupe de travail Terminologie et Intelligence Artificielle (TIA) [20], qui œuvre pour une « terminologie textuelle » [Bourigault & Slodzian, 2000].

Les membres du TIA se démarquent fortement des approches normatives, qui posent le principe de représentations « universelles », au moins dans des domaines stables. Ils s'écartent de la définition première de l'ontologie, dont l'objectif était de représenter le monde dans sa globalité. Ils s'opposent également à la Théorie Générale de le Terminologie fondée par Wüster [Wüster, 1968], qui a posé, depuis les années 30, une vision unificatrice de la terminologie. Ce dernier postule en effet, d'après la critique de [Bourigault & Slodzian, 2000], « une signification conçue comme discrète ou discrétisable, objectivante et permanente qui caractériserait le terme a priori ». Ce courant a influencé les travaux pratiques en terminologie pendant longtemps, ayant pour conséquence des recherches orientées vers un objectif d'universalité des terminologies ou ontologies pour un domaine donné : puisque les concepts sont supposés stables, alors les termes, qui en sont le reflet linguistique, doivent l'être aussi, et des ressources telles que des ontologies ou des terminologies sont donc censées être transposables quelle que soit l'application visée.

Or, ce postulat est remis en question par le développement de l'informatique et son rapprochement avec la terminologie : ces disciplines donnent lieu à des applications variées, qui imposent un constat de variabilité des terminologies [Rastier, 1995], [Bourigault & Slodzian, 2000]. En ingénierie des connaissances, un constat similaire est fait au sujet des ontologies : une ontologie de domaine ne peut être définie une fois pour toutes et être réutilisée à l'envi et quelle que soit l'application [Charlet, 2002]. Pour un domaine donné, il existe donc en réalité autant de RTO que d'applications qui les utilisent. Elles diffèrent du point de vue des connaissances auxquelles elles renvoient, c'est-à-dire sur les unités retenues et leur description, mais aussi de celui de leur degré de structuration ou de formalisation [Aussenac-Gilles & Condamines, 2004].

Un premier clivage existe entre les types de ressources par rapport à leur contenu. Selon la classification qu'en fait [Hernandez, 2005], quatre types de ressources pour quatre types de contenu.

– **les ontologies génériques** tendent à définir des concepts génériques pour plusieurs domaines. Typiquement, WordNet relève de cette catégorie. Cette ressource est considérée comme une ontologie [Hernandez, 2005], bien que les relations existant entre les termes soient plus souvent d'ordre linguistique que conceptuel. Les groupes de synonymes, ou synsets, de WordNet expriment des concepts. Cependant, les relations sont des relations lexicales, ou sémantiques, telles que la synonymie, l'antonymie, les relations taxinomiques, etc. Ainsi, au niveau global, et de l'avis des concepteurs eux-mêmes [21], il s'agit avant tout d'une base lexicale. Ce type de ressource, cherchant à établir tous les sens possibles pour un terme, est difficile à exploiter automatiquement.

20. `http://tia.loria.fr/TIA/`
21. `http://wordnet.princeton.edu/`

- **les ontologies de domaine** s'intéressent à la conceptualisation d'un domaine particulier. A ce titre, l'objectif est de normaliser, et donc de limiter, le sens de chaque terme dans le domaine concerné. Cela en fait des ressources beaucoup plus exploitables du point de vue informatique, puisque l'ambiguïté est largement réduite.
- **les ontologies d'application** « contiennent toutes les définitions qui sont nécessaires pour modéliser la connaissance propre à l'élaboration d'une tâche particulière » [Hernandez, 2005]. En somme, une méthode spécifique de constitution est appliquée en fonction de l'application visée par la ressource.
- enfin, **les ontologies de représentation de la connaissance** sont généralement des ontologies de haut niveau, représentant des concepts abstraits. Ces ontologies peuvent par la suite être utilisées pour définir des concepts spécifiques.

Sur le plan du contenu, nous pouvons d'ores et déjà situer notre ressource par rapport à notre objectif : d'une part, les documents relèvent de domaines très spécialisés ; il est donc peu pertinent de considérer la possibilité d'utiliser des ontologies génériques. Inversement, le nombre de domaines spécialisés à traiter est très important et non fermé ; définir une ontologie par domaine est donc de même difficile à concevoir. Cependant, cela n'exclut pas la nécessité d'intégrer des connaissances spécialisées dans la ressource finale. Les ontologies d'application sont par contre à prendre en compte, puisqu'elles permettent d'envisager une ressource comme une réponse à un problème posé précisément. Cependant, rappelons que la diversité des domaines oblige, non pas à un caractère universel, mais du moins à une couverture large pour une seule et même application et à une véritable cohérence de structure en ce qui concerne les unités impliquées. Enfin, les ontologies de haut niveau sont peu pertinentes dans notre cas, puisque nous sommes relativement éloignée de connaissances très abstraites.

Au final, nous situons donc notre ressource, du point de vue du contenu, entre les ontologies de domaine, pour leur caractère spécialisé, et les ontologies d'application, pour leur adéquation à un besoin précis.

Du point de vue de leur degré de formalisation, il existe un spectre important de possibilités de représentations. Elles se distinguent les unes des autres, nous l'avons expliqué, par leur degré de formalisation, qui fait suite à l'engagement sémantique pris au début de la définition de la ressource. L'engagement sémantique correspond au niveau de spécification formelle permettant de restreindre l'interprétation de chaque concept et ainsi d'en donner la sémantique [Bachimont, 2000]. En somme, plus l'engagement sémantique est important, plus l'ambiguïté potentielle entre concepts de la ressource est réduite. Les représentations des connaissances peuvent alors se ranger sur une échelle en fonction de cet engagement, et vont du vocabulaire contrôlé, avec très peu d'engagement sémantique, à la hiérarchie formelle couplée à des relations non taxonomiques et à des axiomes. Ainsi, les représentations se plaçant à la gauche de l'échelle ont une sémantique qui est uniquement définie dans l'esprit des personnes qui les utilisent : elle n'est

donc pas formalisée explicitement. A l'autre extrémité de l'échelle se placent des structures où la sémantique des composants est explicitée le plus possible, de manière à pouvoir tirer parti au maximum de cette information, par la machine comme par l'humain. Nous reprenons et adaptons cette échelle dans la figure 3.3, formalisée par [Hernandez, 2005].

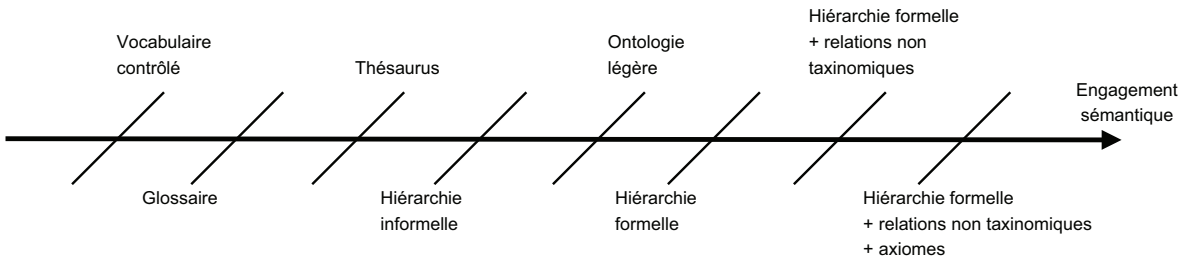

FIGURE 3.3 – Types de représentations de connaissances en fonction de leur degré d'engagement sémantique (adaptation d'après [Hernandez, 2005])

Ainsi, pour une application donnée, il est important de savoir à quel degré placer l'engagement sémantique. Il faut donc savoir à quel point l'interprétation des éléments de la ressource doit être contrainte, grâce à la normalisation sémantique [Bachimont, 2000]. L'application peut nécessiter un degré très profond d'engagement sémantique, ou au contraire, peut avoir besoin de peu de détails à cet égard.

L'engagement sémantique semble porter chez B. Bachimont sur les connaissances en tant que concepts, c'est-à-dire en tant que classes d'individus potentiellement réalisés, la plupart du temps. Dans ce cas, l'objectif est de contraindre l'interprétation des termes les représentant à une seule signification, pour un domaine et un contexte donnés.

Le cas des entités nommées, qui expriment linguistiquement les informations contextuelles qui nous intéressent, n'est pas abordé, et il ne nous paraît pas entièrement assimilable au cas général exposé par [Bachimont, 2000]. En effet, du point de vue référentiel, une entité nommée désigne dans tous les cas un élément bien précis du réel, qu'il s'agisse d'une instance actualisée d'un concept donné, comme dans des syntagmes nominaux définis comme *la voiture grise immatriculée 954 EGD 84* par exemple [22], ou d'un référent unique désigné par un nom propre, comme *Thom Yorke* [23]. Cependant, les entités nommées sont soumises aux mêmes variations que les unités de langue plus classiques [Ehrmann & Jacquet, 2006], et sont donc sujettes à des phénomènes d'ambiguïté.

L'un des cas prototypiques d'ambiguïté des entités nommées cités par [Ehrmann, 2008], non dans le cadre de l'engagement sémantique mais dans celui, somme toute proche, de la résolution des cas de polysémie et d'homonymie dans les entités nommées, est celui du mot *Vienne*, qui

22. La voiture grise immatriculée 954 EGD 84 appartient à la mère de l'auteur de ces lignes.
23. Thom Yorke est le chanteur et leader du groupe britannique de rock alternatif Radiohead.

peut désigner la capitale autrichienne ou bien la ville iséroise en France. Ce cas d'homonymie peut être géré de différentes façons en fonction de l'application visée, et en fonction du modèle de représentation adopté. Dans l'hypothèse d'un modèle contenant uniquement une classe *lieu*, alors *Vienne*, lorsqu'il est rencontré par le système utilisant ce modèle, pourra être identifié comme lieu, mais le traitement s'arrêtera là. Il n'existerait donc qu'un seul *Vienne*, ambigu de fait. En revanche, dans un modèle possédant une classe *ville* et une classe *pays*, alors *Vienne* pourrait être rattaché soit à *France*, soit à *Autriche*. Dans ce dernier cas, l'engagement sémantique de départ serait donc plus profond que dans le premier, puisqu'en l'occurrence, le fait même de l'existence d'une classe pays permet de restreindre l'interprétation de l'entité nommée *Vienne* et empêche l'ambiguïté. Dans le cas des entités nommées, il ne s'agit donc pas de « définir » leur sens, mais bien de lever leur potentielle ambiguïté référentielle.

Reste à savoir si le système employant la ressource sera en mesure, sur les données qu'il doit traiter, de désambiguïser cette entité nommée. Pour cela, notre engagement sémantique se place aussi au niveau de la fiabilité que nous attribuons aux données saisies par l'auteur d'un document. Cette fiabilité est d'autant plus probable dans notre cas que les documents traités sont des articles scientifiques et des brevets d'invention, et qu'une partie de leur but de communication est d'avoir valeur de preuve d'une propriété entre un auteur et un contenu. Par conséquent, les données sont *a priori* saisies avec soin. Pour reprendre notre exemple des villes de Vienne, il paraît peu probable qu'un auteur situe dans le mauvais pays la ville dans laquelle il se trouve ou se trouve l'organisation pour laquelle il travaille. Ainsi, bien que cet engagement sémantique de fait puisse induire quelques erreurs, elles restent marginales, et ne parasitent pas réellement les traitements de désambiguïsation.

En dernier lieu, il convient d'adopter un engagement sémantique suffisamment profond pour résoudre les cas de synonymie entre entités nommées. C'est par exemple le cas des deux entités nommées d'organisations *Université de Toulouse 2* et *Univ Toulouse II*, qui en tant qu'hétérographes synonymes, désignent toutes les deux la même université toulousaine. En effet, les auteurs de documents peuvent saisir l'un ou l'autre indifféremment, mais nos traitements nécessitent de regrouper les variantes d'un même nom d'organisation. Pour atteindre l'engagement sémantique nécessaire à l'identification de ces variantes, nous faisons appel à un processus de normalisation.

L'engagement sémantique de notre ressource doit donc permettre de résoudre les problèmes de synonymie, ainsi que de désambiguïser les éléments pertinents pour nos traitements dans le cadre de notre application. Puisque nos travaux portent majoritairement sur des entités nommées, ce sont elles que l'engagement sémantique concerne en priorité. La désambiguïsation et la résolution des cas de synonymie que cet engagement doit permettre sont donc obtenues par des moyens différant de ceux employés pour les unités lexicales « classiques », mais n'en est pas moins réel, et sûr.

De fait, une RTO ne peut être optimale que dans la mesure où sa construction est en permanence guidée à la fois par la nature des éléments à représenter, le corpus ou la collection de documents à traiter, et l'application finale dans laquelle elle sera utilisée. Nous avons traité en 3.1 le paramètre relatif au corpus des documents à traiter et à leur nature. Nous faisons état à présent de la manière dont l'application finale, un système d'immersion documentaire, son contexte d'utilisation, et la nature des éléments à représenter, ont influencé la constitution de la ressource, de manière à la rendre efficace.

3.2.2 Représentation des informations pour l'immersion documentaire dans un environnement industriel

L'application et le contexte de son utilisation tiennent un grand rôle dans la définition d'une ressource. En effet, la contrainte professionnelle pose d'emblée une série de conditions à remplir pour obtenir une ressource efficace :

- une adéquation de la ressource à l'objectif de communication des documents ; du point de vue de la production du message, nous avons vu plus haut que les informations contextuelles étaient fondamentales ;
- un accès à des documents issus de deux genres proches quant à leurs objectifs de communication, mais de domaines extrêmement divers d'un objectif de mission à l'autre ;
- un coût de conception et de développement limité ;
- une acceptabilité par l'intégration aux outils et modes de stockage de l'information déjà présents dans l'entreprise ;
- un coût cognitif limité de l'activité de recherche d'information ;
- un accès adapté à des objectifs de mission très différents ; ici, c'est le versant « récepteurs » de la situation d'énonciation qui doit être traité.

Nous examinons dans les sous-sections suivantes ces conditions.

3.2.2.1 Adéquation de la ressource aux documents

L'adéquation de la ressource à l'objectif de communication des documents a déjà été traitée en 3.1 page 71. Nous nous contenterons donc de rappeler que les informations contextuelles, relatives à la situation d'énonciation, sont primordiales pour respecter l'objectif de communication des documents et leur fonction pragmatique.

3.2.2.2 Diversité des domaines, limite des coûts de conception et acceptabilité

La diversité des domaines traités et la nécessité de limiter le coût de conception de la RTO ne permet pas de pencher pour la constitution d'ontologies formelles. En effet, la présence de domaines extrêmement spécialisés aurait pu inciter, dans un premier élan, à considérer cette

possibilité. Cependant, puisque le nombre de domaines abordés est très élevé, et va de plus en augmentant, il serait contre-productif de chercher à constituer pour chacun d'entre eux une ontologie de domaine. La création d'une ontologie scientifique « générale » n'a quant à elle même pas été envisagée, étant donnée l'impossibilité de cette tâche. [Charlet, 2002] donne une définition « primaire » de ce qu'est une ontologie en ces termes :

> « Ensemble des objets reconnus comme existant dans le domaine. Construire une ontologie c'est aussi décider de la manière d'être et d'exister des objets. »

Puisque nous traitons un nombre potentiellement infini de domaines et sous-domaines, et étant entendu qu'il est dans ce contexte inenvisageable de formaliser « l'ensemble des objets reconnus comme existant » dans chacun d'eux, l'éventuelle constitution d'ontologies formelles dans leur sens classique aurait été contreproductive, puisque coûteuse, et aurait eu une couverture insuffisante, entraînant une perte d'information importante.

Pour les mêmes raisons de coût, il est important de prendre en considération l'ensemble des outils et ressources déjà présents dans un milieu industriel donné, et plus précisément dans une entreprise : des habitudes sont ancrées qu'il est impossible de bouleverser totalement, des choix ont été faits quant aux outils et modes de stockage déjà en place. A ce titre, une RTO devant être intégrée aux outils pré-existants devra avoir une forme adaptée à la fois aux informations qu'elle porte et aux objectifs qu'elle doit servir, mais aussi aux moyens techniques et concrets utilisés avant son implantation. En somme, elle doit être acceptable du point de vue pratique en termes ergonomiques (cf. 2.2.3.2 page 44). Pour cette raison, notre RTO prend la forme d'une base de données relationnelle, forme nécessaire et suffisante pour que son utilisation soit possible d'une part, et efficace d'autre part.

3.2.2.3 Un coût cognitif limité de l'activité de recherche d'information

Comme nous l'avons montré en 2.2 page 28, toute activité de recherche d'information implique un nombre important de processus cognitifs, tels que la planification, la régulation, etc., qui peuvent mener rapidement à une surcharge cognitive. Cette surcharge est d'autant plus probable que la recherche d'information est une activité secondaire, et que l'individu doit la gérer de front avec l'activité principale qui la nécessite. Nous cherchons à limiter cette surcharge, c'est pourquoi la ressource doit inclure l'ensemble des types d'information utiles à l'activité, appelés à permettre la construction de connaissances viables, et seulement ceux-là.

3.2.2.4 Diversité des objectifs de l'immersion documentaire pour la recherche d'information

L'utilisateur doit pouvoir sélectionner le type des informations relatives à la situation d'énonciation qui lui permettront d'accéder aux documents les plus pertinents pour lui, et de la manière

la plus « parlante » pour un objectif donné. Ainsi, la ressource doit à la fois prendre en compte les circonstances entourant la production des documents, et la situation de réception de ces documents pour l'utilisateur. Cette dernière correspond au contexte dans lequel se trouve l'utilisateur [Hernandez, 2005] : quel est son objectif de recherche ? Quel est son thème ? L'utilisateur s'intéresse-t-il plus particulièrement au contenu sémantique des documents, ou bien au contexte, géographique, professionnel, etc., dans lequel ils ont été produits ? Nous pouvons d'ores et déjà poser les jalons de la réponse à cette dernière interrogation : dans la majorité des cas, l'utilisateur aura besoin de prendre en compte les deux dimensions du document, mais à des degrés variables en fonction de l'objectif.

Les conditions de réception des documents se placent, nous l'avons vu, dans le cadre de la recherche documentaire pour de la veille ou du conseil en stratégie de l'innovation. Nous avons également constaté que cette unité « globale » recouvre en réalité une pluralité, non prévisible à l'avance, d'activités de conseil : des états de l'art, des études de marché, des recherches de partenaires scientifiques, etc.

De fait, la ressource doit couvrir l'ensemble de ces besoins, ce qui revient à dire que la ressource doit fournir autant de points d'entrée que nécessaire pour donner à l'utilisateur une vue pertinente des documents sur lesquels il travaille pour son objectif.

Avec cette priorité en tête, nous avons donc pris le parti de construire une ressource présentant autant de plans, ou facettes, que de types d'informations en présence, faisant référence à la situation d'énonciation ou au contenu sémantique des documents, à leur énoncé. Les cinq plans de la ressource sont donc les suivants :

1. plan des **dates** de publication des documents ;
2. plan des **auteurs** des documents ;
3. plan des **lieux** de publication des documents ;
4. plan des **thèmes** des documents ;
5. plan des **organisations** publiant les documents.

Une facette contient donc un type d'information spécifique, distinct de tous les autres. Du point de vue de l'application finale dans laquelle la ressource s'insère, cette structure en facettes autorise l'utilisateur à interroger le système d'immersion sur la base des types d'informations qu'il aura choisis. Cela implique en réalité quatre choses :

- l'utilisateur a la possibilité de choisir autant de points d'entrée que nécessaire : il peut par exemple accéder aux documents par la combinaison des plans des auteurs et des organisations publiantes, afin de révéler des réseaux de collaboration, à travers des auteurs travaillant dans plusieurs organisations ;
- les facettes permettent à l'utilisateur de construire ses visualisations : par exemple, la facette des lieux fournit des informations exploitables pour projeter d'autres types de données

sur une carte géographique ;

- les facettes sont utilisables en tant que filtres : l'utilisateur peut ne s'intéresser aux réseaux de collaboration entre organisations que sur une période donnée ou sur une certaine zone géographique par exemple. Auquel cas, le plan temporel et/ou le plan géographique serviront de filtres, ou de cadre à la recherche ;
- il n'est pas nécessaire de sélectionner tous les points d'entrée : tous les plans peuvent s'avérer nécessaires si l'ensemble des utilisations potentielles est pris en compte. En revanche, pour une utilisation donnée en fonction d'un objectif précis de l'utilisateur, certaines facettes sont parfois superflues. Par exemple, la facette géographique se révèle inutile si les documents sélectionnés au départ ont fait l'objet d'une recherche sur un seul pays, et si des informations géographiques plus précises telles que les villes de publication sont non pertinentes pour l'objectif visé.

Pour résumer, grâce à cette structure en facettes, l'utilisateur choisit dans le système d'immersion ses points d'accès aux documents, et seulement ceux-là, et peut visualiser et filtrer les documents à l'aide d'une ou de plusieurs facettes.

La navigation par facettes (*faceted navigation*) est à l'heure actuelle utilisée dans plusieurs systèmes. Selon [Hearst, 2008b] :

> « Faceted navigation is a proven technique for supporting exploration and discovery within an information collection. The underlying data model is simple enough to make navigation understandable while at the same time rich enough to make navigation flexible in a wide range of domains. »

L'auteur travaille sur des systèmes de navigation par facettes, et dirige entre autres le projet de recherche Flamenco, consacré à la conception d'une interface de recherche fondée sur un tel système. L'élément clé de cette interface est l'utilisation de métadonnées classées en catégories, servant à la fois à structurer, indexer et interroger la collection d'informations traitée. Les catégories de métadonnées sont en effet présentées sur l'interface comme autant de facettes. Pour chacune d'elles, les différentes instances peuvent être sélectionnées. Les informations correspondantes sont alors affichées, et il est possible de raffiner ou d'étendre la recherche en sélectionnant de nouveaux éléments de facettes, ou en exécutant une recherche par mots-clés. Nous présentons dans la figure 3.4 page suivante un exemple d'interface dynamique, portant sur les titulaires de prix Nobel dans le monde.

Dans le même ordre d'idée, le projet Scriptorium [Folch & Habert, 2000] utilise un système de facettes fondé sur les métadonnées des documents. L'objectif du projet Scriptorium, mené à partir des années 1990, était d'identifier des sujet saillants ou émergents grâce à l'analyse automatique de discours des différents acteurs sociaux de l'entreprise EDF [Lahlou, 1996]. Pour cela, [Folch & Habert, 2000] utilisent des Topic Maps, tirées de la norme [ISO, 2000]. Topic Maps est

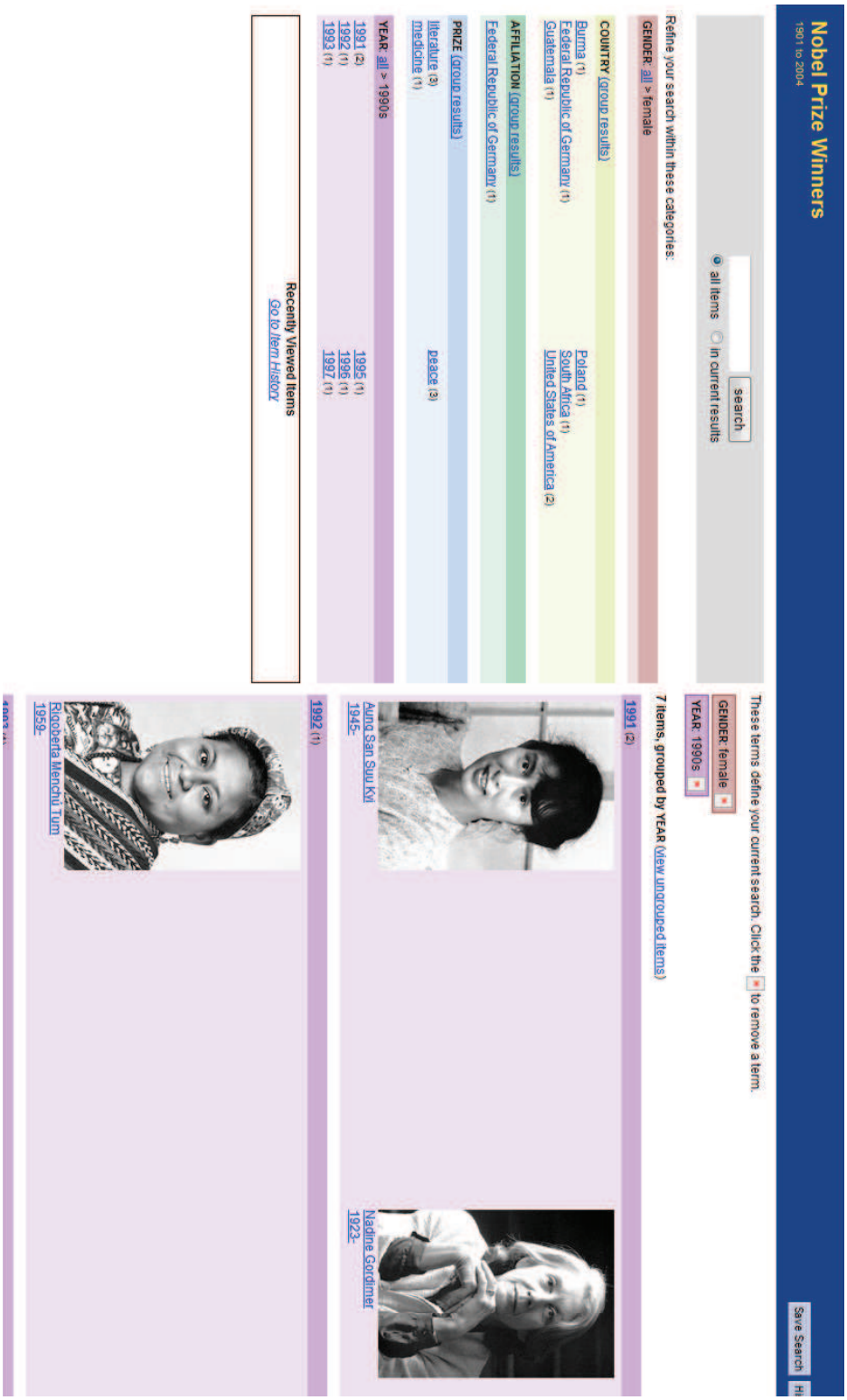

FIGURE 3.4 – Exemple de recherche *via* l'interface Flamenco [Hearst, 2008a] sur les femmes titulaires d'un prix Nobel obtenu dans les années 1990

un « *international standard providing a language [...] to construct a layer of topics and relations aimed at classifying and semantically tagging a collection of documents* » [Folch & Habert, 2000].

Ces Topic Maps permettent des « multiple, concurrent views of sets of information objects. The structural nature of these views is unconstrained ; they may reflect an object oriented approach, or they may be relational, hierarchical, ordered, unordered, or any combination of the foregoing » [ISO, 2000]. Cette norme, qui permet donc une certaine souplesse dans la constitution de Topic Maps et dans les modes de navigation, s'appuie entre autres sur des facettes. En effet, en plus des informations contenues dans l'ensemble à traiter, il est possible de constituer des paires d'attributs-valeurs, appelées facettes, et de les rattacher aux objets textuels.

Dans le cas du projet Scriptorium, ces facettes contiennent des métadonnées associées aux objets textuels extraits. Ces métadonnées concernent le contexte de production du texte ou bout de texte, soit le nom de l'auteur, son affiliation, la date de publication, etc. [Folch & Habert, 2000] opposent les facettes, contenant les métadonnées externes, aux « topics », informations inductives dérivées du texte à l'aide d'Alceste [24], également intégrés à la Topic Map. Les facettes permettent de créer des « vues » (*views*) de la collection de textes, c'est-à-dire des sous-corpus construits à partir de la collection complète et contenant des textes ou éléments de textes correspondant aux critères d'une ou de plusieurs facette(s) sélectionnée(s). Dans ce cas encore, l'utilisation de facettes fondées sur les métadonnées des éléments d'information permet une certaine souplesse et un dynamisme de l'utilisation de l'outil s'appuyant sur une telle ressource.

De manière moins spécialisée, ces systèmes à facettes sont de plus en plus présents dans l'offre commerciale grand public, bien qu'ils ne portent généralement pas sur des données textuelles. La navigation dans les données contenues dans un iPod [25], par exemple, est réalisée à l'aide d'une telle structure. En effet, un même titre musical est accessible par divers moyens en fonction de la facette qui est considérée par l'utilisateur : celui-ci peut y accéder par la ou les listes de lectures qui le contiennent, par son interprète, par l'album auquel il appartient, etc. Le chemin choisi par l'utilisateur dépend de son besoin immédiat et des critères de sélection qui découlent de sa représentation. Dans la figure 3.5 page suivante, nous présentons deux chemins permettant d'accéder au morceau *Sweepstakes* du groupe Gorillaz : l'un par la sélection de la facette *Artistes*, l'autre par la facette *Genres*.

De même, des sites de vente par correspondance, comme Sarenza.com, sont construits sur ce type de modèle. Un même produit, comme une paire de chaussures, peut être renvoyé en résultat de plusieurs requêtes, en fonction des facettes interrogées et des instances de ces facettes qui le caractérisent. Une paire de sandales compensées par exemple (voir la figure 3.6 page 91) peut être caractérisée en fonction de facettes telles que « Type de chaussures » (*compensées*),

24. Alceste est un logiciel de statistiques textuelles, conçu au CNRS et commercialisé par la société Image. Page web : `http://www.image-zafar.com/index_alceste.htm`.

25. Baladeur lecteur de mp3 de la marque Apple.

FIGURE 3.5 – Accès au titre *Sweepstakes* de Gorillaz, par sa facette *Artiste* en haut, et par sa facette *Genre* en bas.

« Style » (*talon bois*), « Hauteur du talon » (*plus de 8cm*), « Matière » (*cuir*), « Prix » (*entre 50 et 100€*) et « Saison » (*printemps/été 2011*). La pointure, la marque ainsi que la couleur peuvent également être prises en compte. Les critères se contraignent les uns les autres, puisque certaines instances des facettes ne sont disponibles que pour des sélections données : par exemple, il n'existe pas de chaussures compensées de plus de 8 cm de talon et de couleur rose. Quoi qu'il en soit, l'utilisateur du site croise donc les différents critères à travers les facettes pour obtenir uniquement les chaussures correspondant à son « besoin ».

Quoi qu'il en soit, dans le cadre plus précis, et beaucoup plus spécialisé, de notre contexte applicatif, cette souplesse de la représentation à facettes, qui permet une flexibilité dans l'interrogation, autorise également à agencer de manière assez rigide le détail de certains plans.

La facette induit la notion de prisme, puisqu'elle représente un ensemble de données ou d'informations selon un point de vue déterminé. Elle permet, ainsi que nous venons de le voir, de filtrer cet ensemble à partir de critères qu'elle définit.

Nous dérivons de ce concept de facette la notion de plan. Celui-ci permet de structurer l'espace de représentation d'une facette particulière. Chaque plan contenant un type précis de données, il est nécessaire de le constituer en fonction des informations qu'elles véhiculent et du degré de détail nécessaire. Ici, la combinaison de la nature des données et de l'utilisation qui en sera faite guide la structuration de chaque plan.

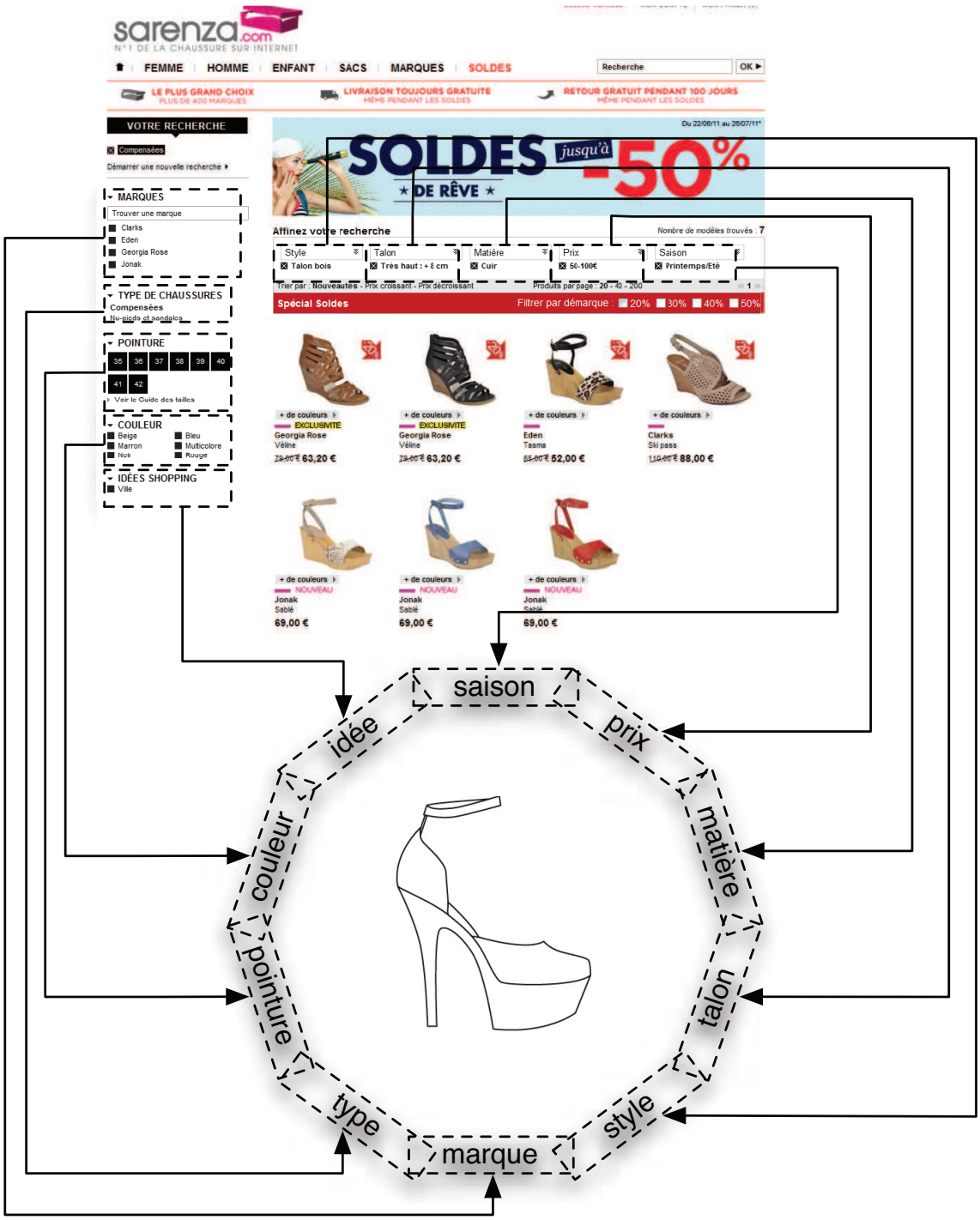

FIGURE 3.6 – Résultat d'une navigation par facettes sur le site Sarenza.com

3.2.2.5 Le plan des dates de publication

Un document publié l'est toujours à une date donnée : dans nos ensembles documentaires, l'information est systématiquement présente. Le plan des dates contient les dates clés et leur sémantique. Les articles scientifiques sont exclusivement associés à leur date de publication. En revanche, plusieurs dates sont liées à un même brevet : la date de dépôt de la demande, de priorité, ainsi que la période de validité. Des liens sont donc établis entre différents types de dates, en plus des liens fixés entre une (des) date(s) et un (des) document(s). Par conséquent, plusieurs dates sont associées à un même brevet.

L'année de publication est l'information minimale, éventuellement complétée par le jour et le mois pour les brevets. Dans ces cas-là, jour, mois et année de publication sont distingués dans la ressource, et une relation d'inclusion unit jour, mois et année de publication. En effet, une précision dans la structure implique la possibilité de précision dans l'immersion à partir du système, ce qui s'avère nécessaire si l'utilisateur a besoin de sélectionner des documents sur des dates spécifiques.

Le plan des dates est avant tout considéré comme une ligne chronologique, et prend donc la forme d'une énumération ordonnée représentant la linéarité du temps. Cependant, il peut être perçu comme hiérarchisé puisqu'il comporte des structures arborescentes permettant de moduler le niveau de détail des points temporels considérés. Nous présentons sur la figure 3.7 cette vision hiérarchique du temps.

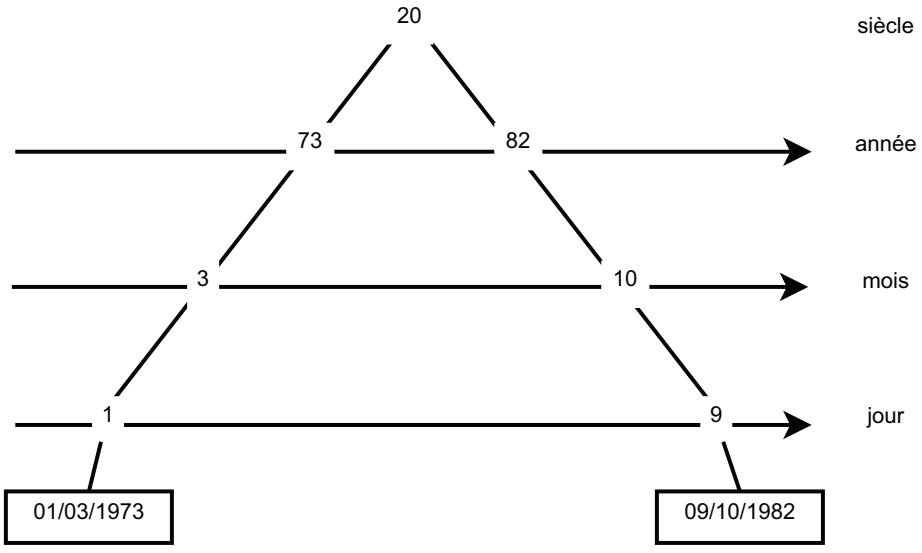

Figure 3.7 – Représentation hiérarchique de la ligne temporelle

Pour les deux exemples de dates complètes, elles s'organisent sur l'échelle temporelle en fonction du degré de granularité de leurs unités : un jour est rattaché à un mois, lui-même rattaché

à une année. Cependant, dans l'absolu, cette hiérarchie reste une extrapolation à partir d'une ligne temporelle plane : les éléments englobants comme les années peuvent être considérés, en fonction du point de vue adopté, comme des points, mais aussi comme des intervalles permettant de regrouper d'autres intervalles plus petits, les mois, et/ou des points, comme les jours.

Dans nos données, seuls les jours, mois et années sont disponibles. Cependant, à partir de là, il est également possible de regrouper les informations par siècles par exemple, ainsi que nous l'avons représenté sur la figure.

3.2.2.6 Le plan des auteurs

L'auteur d'un document est un humain, et est désigné linguistiquement par un nom de personne. De fait, ce plan est constitué de l'ensemble des noms d'auteurs d'un ou plusieurs documents composant l'ensemble documentaire. Les utilisateurs faisant une recherche sur les auteurs rentrent un nom complet ou seulement une partie de celui-ci, dans la plupart des cas un patronyme. Une structure plate est donc suffisante. De fait, cette facette prend la forme d'une liste, ou d'un lexique de noms de personnes.

3.2.2.7 Le plan des lieux de publication

Le lieu de publication d'un document est toujours présent dans les informations, au moins au niveau du pays. Cependant, dans un grand nombre de cas, la ville est elle aussi communiquée. Enfin, l'adresse exacte peut également apparaître. Or, pays, villes et adresses précises sont liés par des relations d'inclusion. Du côté de l'utilisateur, pouvoir situer précisément un lieu de publication pour un document est nécessaire pour atteindre certains objectifs.

C'est donc naturellement une structure hiérarchique, fondée sur des relations d'inclusion, qui a été définie pour la facette des lieux de publication. Le modèle hiérarchique adopté pour les lieux de publication est le suivant :

$$\text{Monde} > \text{Pays} > \text{Ville} > \text{Adresse précise}$$

où *Monde* est le nœud racine de la hiérarchie.

3.2.2.8 Le plan des thèmes

Un document porte toujours sur un thème (*a minima*). Le thème est ce dont il « parle », et dans le cas des documents textuels, « les "sujets" abordés dans les textes ou les segments de textes d'un corpus »[Pichon & Sébillot, 1999]. Le thème donne un aperçu du contenu du document, et peut être considéré comme le caractérisant : il est représentatif d'un ou de plusieurs éléments du monde extra-linguistique, c'est-à-dire du monde extérieur que le langage permet de décrire. En tant que propriété « générale » d'un texte, nous considérons qu'il est pertinent d'intégrer ce thème aux éléments de caractérisation du texte. Enfin, le thème répond à la question du *quoi*,

qui viendrait compléter les questions du *où, quand, qui* qui posent les paramètres d'une situation d'énonciation. Or, nous avons spécifié plus haut qu'un document ne peut être considéré que comme un tout possédant une dimension « signe » et une dimension « medium ».

Les thèmes caractérisant des documents isolés peuvent constituer, puisque ces documents appartiennent à un ensemble, des thèmes globaux relatifs à cet ensemble documentaire s'ils sont très fréquents. Ils peuvent également représenter des thèmes intermédiaires, ou sous-thèmes, s'ils sont spécifiques à un sous-ensemble donné.

Les ensembles documentaires sont caractérisés dès le départ d'un point de vue thématique général. Nous avons en effet indiqué (voir la sous-section 1.2.1 page 17) qu'ils étaient constitués par les analystes grâce à des équations de recherche contenant des mots-clés. Les thèmes globaux sont donc déjà connus, au moins en partie, lors de la phase de recherche d'information.

Matériellement, les thèmes d'un document ou d'un ensemble de documents s'expriment, d'une manière ou d'une autre, à travers des suites de mots. Dans le chapitre 4 page 101, nous explorons les moyens les plus adéquats pour extraire ces suites de mots thématiques, afin de les représenter pour les utilisateurs.

3.2.2.9 Le plan des organisations

L'organisation publiant un document est une entreprise ou un organisme académique et / ou public. Il peut aussi s'agir, plus rarement, d'un particulier, c'est-à-dire d'une personne individuelle publiant un document pour son propre compte. La distinction faite par les utilisateurs dans leur pratique est la suivante, et ce quel que soit l'objectif de la recherche :

- entreprise : toute organisation privée, à l'exception des hôpitaux et des universités ou écoles privées
- organisme public : toute organisation publique, et par extension, les hôpitaux, universités et écoles privés. Ces derniers sont assimilés aux organismes publics puisque la majorité des organismes académiques et de santé sont publics.
- particulier : personnes individuelles publiant un document en leur propre nom. Dans l'immense majorité des cas, le particulier publie un brevet, et très rarement un article scientifique.

Puisque cette distinction est courante pour l'utilisateur, il est important de la restituer dans la facette des organisations. Par conséquent, cette dernière inclut au premier niveau trois concepts « de base » parmi lesquels les organisations publiantes viendront se répartir : entreprise, organisme public et particulier. D'autre part, il peut exister plusieurs niveaux de hiérarchie entre organisations. Par exemple, un document peut être publié par une université au niveau global, mais être plus précisément attribué à l'un des départements de cette université. Les utilisateurs prennent parfois en compte ces différents niveaux de hiérarchie ; plus précisément, dans des cas

donnés d'utilisation, ils peuvent s'intéresser seulement au niveau des universités, ou à l'inverse, cibler une recherche sur les départements d'une ou plusieurs universités. Ainsi, cette hiérarchie doit nécessairement être prise en compte dans la structure de la ressource. Nous ne sommes donc plus face à une structure plate comme celle des auteurs, mais au contraire devant une structure hiérarchisée. Deux types de liens hiérarchiques doivent être distingués :

- des relations taxinomiques, également appelées relations de subsomption [Hernandez, 2005], ou encore relations « est un ». Du côté de la linguistique, nous parlerions de relation d'hyperonymie. C'est le cas de la relation qui existe entre les concepts supérieurs et les instances de plus haut niveau de ces concepts. Par exemple, l'université de Toulouse-Le Mirail est un organisme public.
- des relations partie-tout, ou relations d'inclusion. Si nous considérons les termes, et non les concepts ou instances de concepts, nous parlerons de relation de métonymie. Ces relations existent entre l'instance de plus haut niveau d'un concept et la ou les instances de niveau inférieur. Par exemple, une relation d'inclusion existe entre l'université de Toulouse-Le Mirail et le département de Sciences du Langage, où l'université inclut le département.

Ces différents types de relations hiérarchiques sont révélateurs de la distinction qui est faite, au moment de l'utilisation de l'application, entre concepts et instances : les concepts de haut niveau permettent d'attribuer un type à l'organisation (entreprise, organisme public, particulier), tandis que les instances sont les unités existant dans le monde réel, extra-linguistique. Nous aurions pu, dans la facette des organisations, rajouter des paliers purement conceptuels et expliciter un plus grand nombre de concepts. Nous pourrions différencier par exemple, au sein des organismes publics, les universités des hôpitaux. Or, cette distinction ne correspondrait à aucun besoin réel de l'utilisateur : cela rajouterait des étapes de traitement informatique, et surtout, une étape de traitement cognitif supplémentaire et manquant de pertinence, ou d'à-propos.

En parallèle de ces liens hiérarchiques, un type de relation non taxonomique existe également entre une organisation et les variantes de son expression linguistique. Ainsi, chaque organisation est reliée à toutes ses variantes rencontrées dans les documents traités. Cette relation de synonymie permet au système d'identifier une même organisation derrière les variantes rencontrées. La variation en contexte est la résultante de plusieurs facteurs, dont la différence de normes de notation en fonction des sources des documents, les erreurs typographiques, ou encore le caractère multilingue des données. Or, il est impératif de gérer correctement cette variation pour obtenir des résultats d'immersion et de recherche efficaces.

Le plan des organisations comporte donc deux types de relations hiérarchiques différents. Il ne rentre pas strictement dans le cadre des hiérarchies formelles tel qu'il est énoncé par [Hernandez, 2005]. En effet, l'auteur pose que toute ressource établissant d'autres relations hiérarchiques qu'une relation de subsomption ne peut accéder au statut de hiérarchie formelle. Elle prend pour exemple la hiérarchie créée par le moteur de recherche Yahoo !, dont elle fournit une

sous-partie :

Accueil

> *Mode & Accessoires*

> *Pour la Femme*

> *Tous les Accessoires Femme*

> *Pierres/Perles*

> *Perle*

Or, au sens strict de la subsomption, les individus de la catégorie *Perle* devraient hériter des propriétés sémantiques de la catégorie *Pierres / Perles*, en vertu de la condition d'héritage des propriétés, ce qui selon [Hernandez, 2005] n'est pas le cas ici. Par exemple, une perle n'a pas la propriété *ayant fait l'objet d'une taille*, contrairement aux pierres [26]. En l'occurrence, d'après l'auteur, c'est la définition très floue des relations pouvant exister entre deux nœuds liés hiérarchiquement qui rend la hiérarchie informelle.

Dans le cadre de nos travaux, cette restriction de Hernandez ne nous paraît pas adaptée. Tout d'abord, ces arguments sont de peu d'importance pour l'utilisateur, qui a besoin que l'information soit structurée en fonction de sa propre représentation. Reprenant l'exemple de la hiérarchie fournie par Yahoo !, celle-ci répond au point de vue et à la « réalité » d'un utilisateur. En l'occurrence, l'utilisateur peut être un consommateur cherchant à acheter des bijoux ou des accessoires de mode. Dans ce cas, pierres et perles partagent les mêmes propriétés, et en particulier des propriétés « décoratives ». Dans ce cas, la relation existant entre *Pierres/Perles* et *Perle* est bien une relation de subsomption : les perles partagent les mêmes propriétés que les pierres en tant qu'accessoires. Il est évident que l'assimilation entre ces éléments ne correspondrait pas à la réalité d'un utilisateur dans un contexte spécialisé en gemmologie par exemple. Ainsi, même une hiérarchie de type aristotélicienne, ayant pour objectif de représenter *le* monde, ne peut rendre compte que d'une réalité donnée, pour un contexte et un point de vue donnés.

D'autre part, il existe certes plus d'un type de relations hiérarchiques dans le plan des organisations. Cependant :

- le nombre de types de relations est connu : il y en a deux, et il ne peut en exister d'autres ;
- le type de chacune des deux relations est déterminé : subsomption stricte entre concepts et instances, et inclusion entre instances de niveaux différents ;
- l'existence de ces relations se justifie par la nature des éléments qu'elles unissent : par exemple, une instance de département ne peut être une instance d'université.

Dans ce cas, et malgré la présence d'une autre relation que la subsomption, il apparaît donc que le caractère formel de cette hiérarchie n'est pas remis en cause.

26. Cet exemple n'est pas issu de [Hernandez, 2005], qui n'en donne pas spécifiquement, mais construit par nous pour les besoins de la démonstration.

En conséquence, nous étendons la définition fournie par [Hernandez, 2005] aux structures hiérarchiques cohérentes du point de vue sémantique, dont les types de relations sont identifiés, et dans lesquelles le nombre de ces types est limité. Nous considérons donc que le plan des organisations correspond bien à une hiérarchie formelle. A cette structure s'ajoute un type de relation non hiérarchique, puisqu'un terme et ses variantes linguistiques sont liés par des relations de synonymie.

Nous incluons également dans la catégorie des hiérarchies formelles la facette des lieux et la facette des dates.

Au final, pour les facettes relatives à la situation extra-linguistique d'un document, donc à sa situation d'énonciation, c'est la structure des éléments du réel auxquels réfèrent les expressions linguistiques des données qui dicte l'organisation de chaque plan de la RTO.

En tant que ressource pour l'immersion documentaire, la RTO trouve la cohésion de ses cinq facettes à travers les liens établis entre elles par les documents. En effet, chaque document est en quelque sorte l'épicentre de la ressource. Chaque facette contient au moins un élément qui pointe vers lui. De fait, des relations non taxonomiques sont établies entre les différentes facettes, et entre les individus de chacune d'entre elles. Ces relations transversales ne sont pas exprimées formellement, pour la simple raison qu'elles n'ont pas besoin d'être formalisées. Elles existent de fait, et chaque utilisateur est capable d'interpréter ces liens sémantiques. Du point de vue du document, ils peuvent être exprimés en une phrase : un document porte sur un ou plusieurs thèmes particuliers, et est produit à un moment et dans un lieu donnés, par un ou des auteurs qui travaillent chacun pour une ou des organisations précises.

Par conséquent, malgré un engagement sémantique inégal en fonction des facettes [Bachimont, 2000], la RTO globale contient assez d'informations pour permettre l'immersion documentaire. La RTO ainsi construite permet d'accéder aux documents qui ont participé à sa constitution. A ce titre, elle est une forme d'indexation de ces documents, qui offre plus de possibilités d'exploitation que les formes d'indexation par mots-clés.

Nous présentons dans la figure 3.8 page suivante une schématisation de notre RTO et des relations entre ses plans et les documents.

Chaque plan étant relié aux autres par le biais des documents, la ressource ainsi constituée permet d'établir des relations binaires entre les instances de deux plans. Par exemple, une instance du plan *Auteurs* est liée à une (ou des) instance(s) du plan *Organisations* par une relation formalisable par l'expression *travaille pour*, à travers les documents ou les deux informations sont en co-présence. A ces relations binaires s'ajoutent des relations plus complexes, impliquant les instances de plus de deux facettes. Par exemple, un auteur travaille pour une organisation à une date donnée. Dans ce cas, les plans des auteurs, des organisations et des dates de publication des documents sont impliqués. Ou encore, un thème a été abordé par une organisation à une date

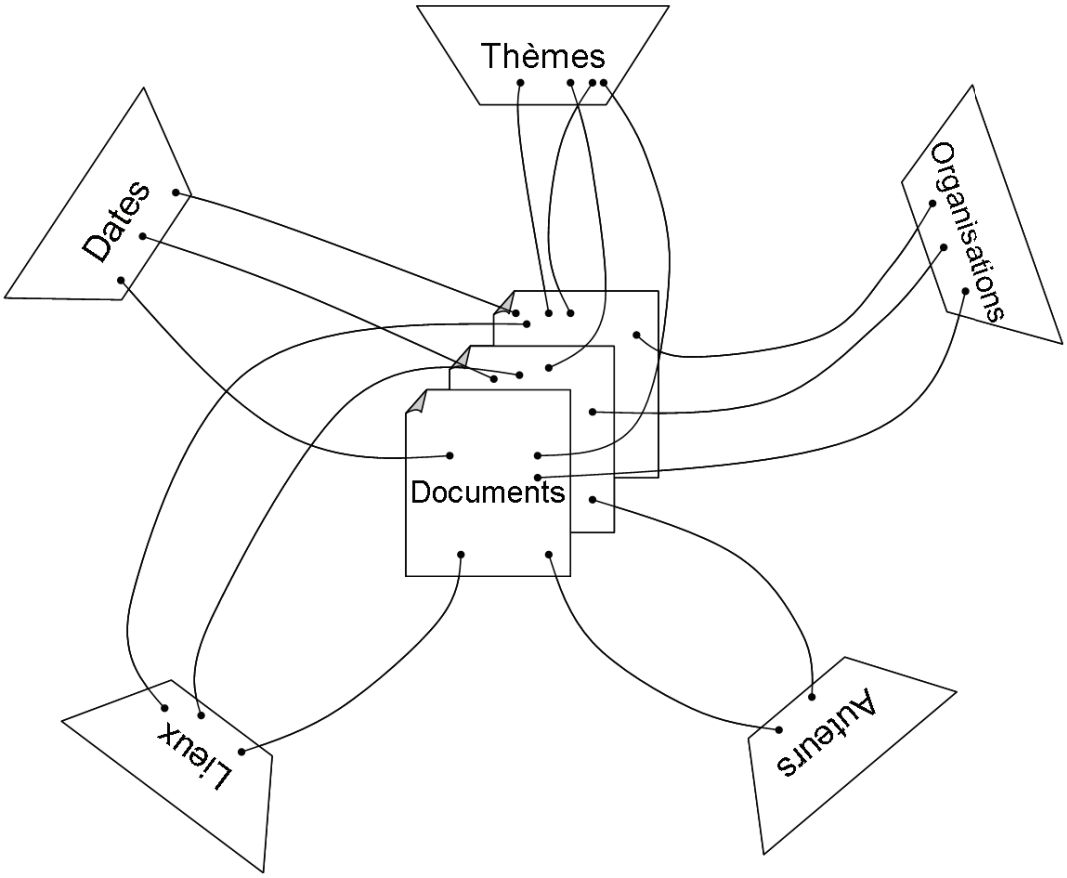

FIGURE 3.8 – Schématisation de la ressource termino-ontologique multi-plans

déterminée, qui correspond au moment où l'auteur X travaillait pour cette organisation. Ici, les facettes des thèmes, des organisations, des dates et des auteurs sont utilisées. En somme, chaque croisement de facettes, pouvant diverger quant aux types de facettes ou quant à leur nombre, donne lieu à des relations différentes ou, pour le moins, plus ou moins riches et complexes. Là encore, les types de relations étant interprétables par l'utilisateur grâce à la répartition en facettes des informations, il n'est pas utile d'en formaliser précisément la nature.

3.2.3 Définition de la ressource conçue

Au vu des différents types de connaissances intégrées dans nos facettes, et de la différence de structure qui existe d'une facette à l'autre, il peut paraître difficile d'attribuer un type précis à notre ressource globale. En effet, nous avons fait le constat que d'un plan à l'autre, la structuration adéquate aux besoins de l'application variait, allant de listes plates à des hiérarchies formelles. Par ailleurs, les relations non taxonomiques entre les plans sont bel et bien présentes, puisque le document est placé au centre de ces relations, et parfaitement interprétables par les utilisateurs. Pourtant, leur type n'est pas formalisé explicitement, justement parce que la structure globale couplée à la capacité interprétative de l'utilisateur suffisent à cela.

Bien que le terme de *ressource termino-ontologique* soit un terme générique, désignant un ensemble de ressources impliquant des termes structurés plus ou moins formellement pour rendre compte de connaissances, nous considérons qu'en l'occurrence, il peut désigner une ressource particulière, mettant en jeu un ensemble de « sous-ressources » dont la structure varie en fonction des informations ou connaissances qu'elle contient. A ce titre, nous opérons un glissement sémantique du terme de ressource termino-ontologique, qui n'est plus, dans le cadre précis de notre application, un terme générique, mais un terme spécifique à vocation englobante, désignant une ressource particulière. Ainsi, et dans notre contexte, nous définissons une ressource termino-ontologique comme suit :

Définition 1. *Une ressource termino-ontologique est une structure de représentation des connaissances pouvant inclure des sous-structures, autonomes mais non indépendantes, dont les formalisations et le degré de structuration varient, et dont la justification se trouve dans la divergence formelle des types de connaissances en présence.*

Chapitre 4

Connaissances contextuelles et thématiques : quelles unités linguistiques ?

Sommaire

La connaissance des mots conduit à la connaissance des choses. Platon

La ressource termino-ontologique que nous avons présentée dans le chapitre 3 page 69 est constituée d'éléments issus des documents. Certains d'entre eux sont d'ordre contextuel, comme l'année de parution d'un document, ou son auteur par exemple, et d'autres relèvent du ou des thème(s) des documents, du contenu abordé. Ces éléments sont véhiculés, dans notre cas du moins, par du texte. Les informations thématiques sont présentes dans le texte des résumés des documents, tandis que les informations contextuelles sont à rechercher du côté des métadonnées associées à chacun des documents.

Dans ce chapitre, nous présentons tout d'abord les séquences de mots graphiques qui peuvent être associées aux thèmes d'un document ou d'un ensemble documentaire. Nous pensons en effet que ces séquences sont potentiellement porteuses de ces thèmes, et nous précisons quel type de séquences de mots correspond le mieux à nos travaux. Dans une deuxième section, nous décrivons la manière dont les entités nommées expriment les connaissances contextuelles. Nous définissons d'abord précisément ces unités, puis présentons notre approche de normalisation pour la formalisation de ces entités.

4.1 Les séquences de mots graphiques : des unités révélatrices du thème d'un document

La séquence de mots graphiques est le terme générique que nous employons pour désigner les suites de plusieurs mots graphiques considérées comme cohérentes en fonction de critères variés. Selon les courants de pensée, les théories ou même les auteurs, elles peuvent prendre la forme de termes, de collocations, de segments répétés, de mots composés, etc. Dans tous les cas, il est question d'unités lexicales ou lexico-syntaxiques, révélatrices d'un emploi dans un contexte donné. Partant de cela, il est possible de considérer intuitivement que, dans certains cas, ces séquences de mots graphiques sont révélatrices du thème des documents dont ils sont extraits.

4.1.1 Le thème et les séquences de mots graphiques

4.1.1.1 Le thème : une notion ambiguë

Dans le langage courant, le terme *thème* renvoie la plupart du temps au principal sujet abordé dans un texte ou un discours. Le Trésor de la Langue Française Informatisé (TLFi) en donne la définition suivante [27] :

> « Idée, sujet développé dans un discours, un écrit, un ouvrage. *Thème d'un discours, d'une conférence, d'un roman, d'un sermon ; thème de propagande.* »

27. Site du TLFi : `http://atilf.atilf.fr/`

Au sens courant, un document quel qu'il soit a un (ou plusieurs) thème(s), qui est (sont) l'idée (les idées) ou le sujet (les sujets) qui y est (sont) développé(s). Réciproquement, le thème est donc véhiculé par un document, un texte, etc., dans sa globalité. En somme, l'unité vecteur du thème est, au sens courant, le document, texte, ouvrage, etc., dans son entier. Par exemple, le thème du présent document est l'immersion documentaire.

La notion de thème se précise dans des usages distincts, en fonction des domaines scientifiques qui la manipulent.

En recherche d'information textuelle, le thème concerne, comme dans le langage courant, le ou les sujets développé(s) dans un « bloc d'information » [Maisonnasse, 2008]. Ce bloc peut être le texte d'un document dans sa totalité, ou bien une partie de ce texte. Bilhaut, pour ses travaux en navigation documentaire, évoque le « thème en tant que "sujet" d'un segment textuel, "ce sur quoi il porte" ». Le thème ainsi considéré « présente un intérêt évident en RI, puisque la tâche d'indexation vise précisément à produire une représentation du contenu informationnel d'un texte, et s'applique assez naturellement à des segments plus importants que la phrase, par exemple sous la forme d'un thème textuel[...] » [Bilhaut, 2004]. En fonction des travaux se penchant sur la question, un thème peut s'exprimer à travers des structures arborescentes, comme le propose [Bilhaut, 2004] (voir l'illustration dans la figure 4.1), ou bien par des mots-clés à la structure plus ou moins complexe [Maisonnasse, 2008], [Haddad, 2002].

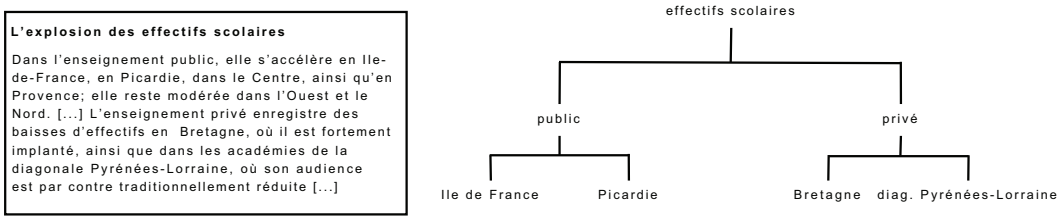

FIGURE 4.1 – Exemple d'arborescence thématique, emprunté à [Bilhaut, 2004]

Pour la linguistique, en revanche, le thème ne recouvre pas du tout la même réalité. De manière générale, dans cette discipline, le thème concerne avant tout la proposition, la phrase ou encore « l'unité syntaxique » pour [Péry-Woodley, 2000]. Dans ce cas, le thème n'est plus l'idée abordée et développée dans un discours, mais ce sur quoi porte une proposition. Ce thème linguistique s'entend en tant que l'une des deux composantes de la paire thème-rhème, proposée par Halliday dans [Halliday, 1967b], [Halliday, 1967a] et [Halliday, 1968], dans le cadre de sa grammaire systémique fonctionnelle. Ainsi, pour résumer de manière grossière, au sein d'une proposition, le thème est ce sur quoi porte cette proposition, et le rhème est ce qui est dit de ce thème. Par exemple, dans la phrase issue de nos données :

The initially released fraction proved complete retention of bioactivity.

le thème est *The initially released fraction*, et le rhème est porté par *proved complete retention of bioactivity*.

En fonction de la discipline qui l'utilise, la notion de thème peut donc prendre des sens différents, même si, nous le verrons, les différentes acceptions peuvent parfois se croiser. Dans ce qui suit, nous présentons un peu plus en détails les travaux rattachés à ces différentes acceptions, et en dégageons les avantages et inconvénients qu'ils impliqueraient si nous les appliquions à nos propres travaux.

4.1.1.2 Le thème textuel : utilisation et unités en recherche d'information

Dans la plupart des travaux en recherche d'information textuelle, cherchant à mettre en place des méthodes de navigation documentaire, le thème considéré est le thème qui se dégage du texte, ou au moins du bloc d'information, regroupant typiquement plusieurs phrases. Nous avons vu (dans la sous-section 3.2.2.8 page 93) que [Pichon & Sébillot, 1999], par exemple, entendent par *thèmes* « les "sujets" abordés dans les textes ou les segments de textes d'un corpus ». Selon les travaux, les objets porteurs de ce thème peuvent être de plusieurs sortes.

Dans ses travaux, [Bilhaut, 2004] cherche à dégager pour un texte sa « structure thématique arborescente », permettant de concevoir des systèmes de navigation documentaires fondés sur ces arborescences. Le thème est donc représenté par une structure logique, manipulable par la machine.

Pour d'autres chercheurs, le thème d'un texte peut être véhiculé par les unités lexicales qu'il contient. Au-delà de la méthode souvent employée par les moteurs de recherche web, qui procèdent à une indexation par mots-clés puis à un appariement entre cet index et la requête de l'utilisateur, un certains nombre de travaux prennent en compte non pas les mots graphiques simples, parfois appelés lexies (par exemple [Maisonnasse, 2008]) [28], mais des mots composés, également appelés *termes composés* (*ibid.*). En effet, l'utilisation de termes ou mots composés et non plus de mots graphiques simples « améliore la recherche d'information car [ces termes] reflètent le thème de la phrase » (*ibid.*). Ainsi, les thèmes locaux, contenus dans les phrases, permettent de mettre au jour le thème global d'un texte, ou d'un bloc d'information. Cette conception du thème peut être rapprochée de ce que [Fairthorne, 1969] a nommé l'*aboutness* extensionnel, exprimé par les unités sémantiques effectivement présentes dans le texte. L'auteur distingue cet *aboutness* de l'*aboutness* intentionnel, qui est le sens, la signification du texte dans son entier, ceux que son auteur a eu l'intention de transmettre.

Dans le même ordre d'idée, [Haddad, 2002] travaille pour la recherche d'information sur « les

28. Emploi erroné d'après la définition de *lexie* du dictionnaire du Centre National de Ressources Textuelles et Lexicales (CNRTL) : *Unité lexicale de langue constituée soit par un mot (lexie simple) soit par des mots associés (lexies composée et complexe).* (`http://cnrtl.fr/definition/lexie`). Cependant, nous conservons cette désignation ici, puisque c'est le terme employé par l'auteur.

informations exprimées dans un syntagme nominal défini sommairement comme un ensemble de termes respectant des lois de la morphologie et de la syntaxe, et possédant une signification propre. Plus précisément, [l'auteur s'intéresse] aux syntagmes nominaux en tant que thèmes ». Là encore, les syntagmes nominaux sont exploités pour restituer le thème global d'un bloc d'information.

Dans la plupart des travaux impliquant le thème et s'inscrivant dans la recherche d'information, les auteurs cherchent avant tout à dégager le thème global d'un document textuel. Pour cela, ainsi que nous venons de le voir, ils cherchent à extraire des mots, lexies et/ou termes, en somme des séquences graphiques qui soient représentatives de ce thème, pour indexer les documents, de manière à faciliter la recherche de l'utilisateur. Or, dans notre cas, nous ne cherchons pas seulement à dégager le thème global d'un document, puisque celui-ci est déjà partiellement déterminé (voir 1.2.1 page 17 et 3.2.2.8 page 93). Nous voulons également donner la possibilité de dégager ce que nous pourrions appeler les sous-thèmes d'un document, afin que l'utilisateur de notre système puisse rechercher de manière plus fine ce qui l'intéresse dans l'ensemble documentaire, ou dans une partie de celui-ci.

Ainsi, si les objectifs rapidement recensés ici diffèrent des nôtres, le type d'unités sélectionnées nous paraît pertinent pour la représentation des sous-thèmes présents dans les documents.

4.1.1.3 Le thème en sciences du langage : une notion fonctionnelle locale

Ainsi que nous l'avons signalé en début de section, le courant majeur en sciences du langage considère que la notion de thème « s'entend principalement à l'intérieur de la phrase ou de la proposition » [Bilhaut, 2004], voire dans « l'unité syntaxique » [Péry-Woodley, 2000].

Cette dernière résume la caractérisation de cette notion de thème en tant qu'il relève de la proposition ou de la phrase sur trois aspects, recensés dans la littérature :

1. « sa fonction : le thème est ce sur quoi porte la proposition (aboutness ou à propos [...]), le point de départ de l'énoncé [...] ;
2. la nature de son référent : le thème a un référent donné, connu, disponible [...] ;
3. sa position dans la proposition ou la phrase : la position initiale, et a fortiori la dislocation à gauche, sont intimement liées à la mise en place du thème [...]. »

Le thème concerne donc, dans la littérature, une proposition, et formellement, en est la partie la plus à gauche. Il est la partie de cette proposition qui est déjà connue, et ce sur quoi cette dernière porte. Sur bien des aspects, le thème se rapproche donc du terme utilisé en recherche d'information, en tout cas du syntagme nominal tel qu'il est présenté par [Haddad, 2002] (voir partie 4.1.1.2 page ci-contre). La différence majeure, cependant, est que [Haddad, 2002] prend en compte tous les syntagmes nominaux, dont ceux qui ne sont pas en position thématique, c'est-à-dire placés en initiale de la phrase ou de la proposition.

Ici encore, le thème a à voir avec l'*aboutness*, traduit par *à propos* chez [Berthoud, 1996]. Mais cette fois, cet à propos s'applique non plus à un texte, ou à un bloc d'information composé de plusieurs phrases, mais à une proposition.

Toutefois, dans son acception linguistique, le thème est une notion relationnelle, et est indissociable du rhème, c'est-à-dire de ce qui est dit de nouveau sur le thème. Typiquement, le rhème d'une phrase simple est souvent le syntagme verbal.

Cette approche du thème, loin de celle utilisée généralement en recherche d'information, est très fine. Elle est en particulier utilisée en analyse du discours, comme chez [Péry-Woodley, 2000]. Mais dans le cadre de nos travaux, l'extraction des thèmes au sens de Halliday [Halliday, 1967b], [Halliday, 1967a] et [Halliday, 1968] présenterait une série d'inconvénients qu'il serait difficile de dépasser. En effet, cette extraction serait extrêmement coûteuse à mettre en place, puisque pour chacune des propositions de chacune des phrases d'un document, il faudrait identifier le thème de manière fiable. Or, la notion de thème est loin d'être triviale et parfaitement systématique, et est donc difficilement formalisable. De plus, identifier le thème entraînerait l'élimination d'un certain nombre d'informations contenues dans ce qui est considéré comme le rhème d'une proposition. Dans ce cas, nous serions privée d'un ensemble important d'informations pouvant s'avérer pertinentes par rapport à un sujet, une tâche et un contexte donnés (voir section 1.2 page 17).

Pour ces raisons, nous avons fait le choix de ne pas utiliser cette conception du thème telle quelle pour nos travaux. Néanmoins, le thème en tant qu'objet de l'*aboutness* de la proposition, s'il est étendu au bloc d'information de manière générale, est à prendre en considération.

4.1.1.4 Le thème comme objet de l'*aboutness* : une notion relative

Dans le chapitre 1 page 9, nous avons expliqué que la pertinence d'un énoncé, et donc d'un document, était relative. Elle dépend en effet du sujet concerné, de la tâche à réaliser, et du contexte dans lequel la tâche prend place [Mizzaro, 1998]. Or, la notion d'*aboutness* appliquée à un texte, et plus particulièrement celle d'*aboutness* intentionnel, peut elle aussi être considérée comme relative. En effet, selon [Beghtol, 1986], [Fairthorne, 1969] ou [Hutchins, 1978] par exemple, cet *aboutness* est déterminé par la raison ou l'objectif pour lequel le document est utilisé [Fidel, 1994].

Ces deux notions, dans notre contexte de travail, c'est-à-dire dans le cadre d'un système d'immersion documentaire, sont donc à rapprocher étroitement. Si nous considérons le thème comme l'objet de l'*aboutness*, alors le thème lui-même d'un document devient relatif. Par voie de conséquence, les unités porteuses du thème peuvent être pertinentes ou non en fonction de la tâche à effectuer, du sujet, et du contexte.

4.1.2 La séquence de mots graphiques : des formes variées

Les termes utilisés en recherche d'information pour désigner des séquences de plusieurs mots graphiques sont variés. Les auteurs parlent de terme composé [Maisonnasse, 2008], d'unité sémantique [Fairthorne, 1969], de syntagme nominal [Haddad, 2002], etc. Il semble donc qu'en fonction des auteurs, des objectifs, des méthodes et des théories sous-jacentes, les séquences courtes considérées comme porteuses du thème d'un document sont de natures variées. Ces formulations ne désignent certes pas forcément les mêmes objets, mais ces derniers se recouvrent au moins partiellement. Dans tous les cas, il s'agit de séquences courtes de mots graphiques, permettant idéalement d'exprimer un thème, ou au moins une partie du thème d'un document. Par conséquent, nous pouvons parler, de manière générale, de collocations, ou a minima de co-occurrences, décrivant potentiellement un thème. Intuitivement, nous pouvons penser qu'une séquence de mots graphiques relevant de l'un des types que nous venons de citer permet de représenter une notion. A ce titre, une telle séquence pourrait être considérée comme un terme. Nous nous proposons donc de clarifier, brièvement, ce qui est considéré comme terme en terminologie, et ce qu'est une collocation en linguistique de corpus.

4.1.2.1 Le terme en terminologie

Dans le cadre de nos travaux, nous cherchons à rendre possible l'identification par l'utilisateur de thèmes et sous-thèmes au sein d'un document, d'un sous-ensemble ou d'un ensemble de documents. Nous avons en effet signalé que le thème global de cet ensemble était déjà identifié, et que nous nous intéressions également à des thèmes inclus dans le thème général, les *sous-thèmes* (voir la sous-section 4.1.1.2 page 104). Pour cela, dégager des séquences de mots graphiques semble plus pertinent dans bien des cas qu'une extraction de mots « simples » : des représentations visuelles comme des nuages de mots « basiques », constitués de ces mots simples, n'apporteraient que peu d'information constructive. Pouvoir extraire *high resolution imaging* d'un document par exemple est plus efficace que de se contenter d'extraire *high*, *resolution* et *imaging* séparément. En effet, un ensemble documentaire relevant du domaine du traitement de l'image ou de l'optique verra le mot simple *imaging* revenir de manière extrêmement fréquente : dans ce cas, ce dernier représente un thème global de l'ensemble documentaire, et est trop large pour être discriminant à l'intérieur de cet ensemble. En revanche, *high resolution imaging* peut être caractéristique d'un sous-ensemble de documents, et donc représenter un sous-thème.

Or, nous pouvons considérer *high resolution imaging* comme un terme du domaine de l'optique. Il peut donc sembler pertinent de s'intéresser aux termes contenus dans les documents, afin de dégager les sous-thèmes dont nous avons besoin. La notion de terme relève avant tout du champ de recherche de la terminologie.

La terminologie, d'après le site web Terminologie.net, est une branche de la linguistique, et

plus précisément « une discipline ayant pour objet l'étude des systèmes de désignations spécialisées et de leurs mises en œuvre dans le langage et les activités professionnelles » [29]. Une terminologie particulière est donc un « ensemble cohérent de désignations linguistiques de valeurs conceptuelles spécialisées » [30].

Le Centre National de Ressources Textuelles et Lexicales (CNRTL), quant à lui, définit cette discipline comme l'« art de repérer, d'analyser et, au besoin, de créer le vocabulaire pour une technique donnée, dans une situation concrète de fonctionnement de façon à répondre aux besoins d'expression de l'usager » [31]. L'objet qui en naît est alors l'« ensemble des termes relatifs à un système notionnel élaboré par des constructions théoriques, par des classements ou des structurations de matériaux observés, de pratiques sociales ou d'ensembles culturels » [32].

Dans les deux cas, la terminologie naît de l'aspect « concret » d'« activités professionnelles », qui nécessitent une structuration et un classement des mots employés dans une situation particulière.

Les objets manipulés en terminologie sont donc des « désignations linguistiques », autrement dit des « termes ». La terminologie est, au départ, une discipline normative : elle cherche à définir une norme pour un domaine donné, et ce dans l'objectif de permettre une meilleure communication entre spécialistes de ce domaine. C'est pourquoi il existe, autour de cette discipline, un grand nombre de normes, telles que la norme [AFNOR, 1987], ou [ISO, 1999], visant à définir précisément l'activité du terminologue. Dans ce cadre, la notion de terme a souvent été définie.

[AFNOR, 1987] définit ainsi le terme :

> « Mot ou groupe de mots employé pour représenter une notion. »

Cette définition est très proche de celle fournie par [Hudon, 1994], qui parle lui de « mot ou groupe de mots représentant un concept ». La définition fournie par le Bureau de la Traduction du Canada est plus détaillée quant à la forme du terme :

> « Mot, syntagme, symbole ou formule désignant un concept propre à un domaine d'emploi. Aussi appelé unité terminologique. »

Un terme peut donc prendre, selon cet organisme rattaché au Ministère des Travaux Publics et Services Gouvernementaux du Canada, des formes très diverses, du mot (que nous supposons graphique) au symbole. D'autre part, cette dernière définition est plus spécifique que les précédentes quant au concept désigné par le terme : le concept est propre à un domaine d'emploi, et la portée du terme comme forme d'expression dudit concept peut donc être limitée à ce domaine.

Enfin, nous citerons la définition fournie par la norme [ISO, 1999], plus particulièrement dédiée aux applications informatiques en terminologie :

29. `http://www.terminologie.net/reperes/rep_clarifi.htm`
30. *ibid.*
31. `http://cnrtl.fr/definition/terminologie`
32. *ibid.*

> « Dénomination, par le biais d'une expression linguistique, d'un concept particulier dans une langue définie. Un terme peut être constitué d'un mot unique ou d'une chaîne de plusieurs mots. La particularité du terme est qu'il correspond à un seul et unique concept, alors qu'une unité phraséologique peut combiner plusieurs concepts d'une manière lexicalisée afin d'exprimer des situations complexes. »

Cette dernière définition est à la fois plus large et plus spécifique que les précédentes : d'une part, elle attribue au terme la forme d'*expression linguistique*, sans plus de détails. Ainsi, toute expression linguistique peut prétendre, du point de vue de la forme, au statut de terme. En revanche, cette définition distingue clairement le terme de l'unité phraséologique : le premier renvoie à un et un seul concept, tandis qu'une unité phraséologique peut porter plusieurs concepts, pour l'expression de « situations complexes ».

En terminologie, le terme est donc avant tout considéré, par les différentes normes et approches normatives, comme l'expression non ambiguë d'un concept unique. Réciproquement, cela suppose que pour un concept donné dans un domaine précis, il existe un et un seul terme permettant de l'exprimer. La forme de cette expression semble par contre plus libre : une expression linguistique pour certains, un mot ou groupe de mots, syntagme, symbole, etc. pour d'autres. La focalisation ne semble pas se faire sur cet aspect.

Pourtant, traditionnellement, un terme est préférentiellement une forme nominale. Le lexique « *Sémiotique* » de [Rey-Debove, 1979] définit le terme comme un « nom définissable à l'intérieur d'un système cohérent ». Par ailleurs, [Otman, 1995], cité par [Rastier, 1995], affirme que « la base des terminologies est constituée d'unités nominales ». Cela s'explique par la tradition très normative de la terminologie depuis l'école wüsterienne. En effet, l'objectif de Wüster était de stabiliser les termes d'un domaine de manière à faciliter l'échange entre spécialistes de ce domaine, locuteurs de différentes langues. Or, « on associe à la forme nominale un haut niveau de stabilité et de pouvoir de désignation » [Condamines, 2005]. C'est pourquoi les terminologies sont traditionnellement fondées sur des formes nominales.

Cependant, certains auteurs se positionnent pour une ouverture du statut de terme à d'autres parties du discours. François Rastier, notamment, prône une prise en compte d'autres catégories qu'il juge tout autant informationnelles, telles que les verbes, mais aussi les adjectifs et les adverbes :

> « Le postulat nominal reste influent, et les terminologies sont de fait massivement constituées de noms. Mais ne gagneraient-elles pas à laisser place, de manière réfléchie et systématique, aux autres parties du discours [...] ? »

Pour justifier cette suggestion, il donne l'exemple des mots *contester*, *contestable* et *incontestablement*, qui ne sont pas moins riches en information sémantique que le nom *contestation*.

Par ailleurs, depuis quelques années est née une terminologie textuelle, c'est-à-dire une dis-

cipline fondée sur l'exploitation de textes pour la constitutions de terminologies. Elle s'éloigne de l'école wüsterienne en cela qu'elle se sert des usages attestés dans un domaine pour constituer une terminologie, alors qu'à l'inverse, Wüster rejetait l'usage de textes, considérés comme recelant une information non fiable, car contenant des éléments non entièrement stabilisés et ne pouvant donc être qualifiés de termes. Cette terminologie textuelle fait donc évoluer la tradition wüsterienne, et oriente la discipline vers une sorte de relativité d'une terminologie donnée, en fonction de l'application visée.

Quoi qu'il en soit, si les termes sont essentiellement des formes nominales à l'heure actuelle, cet état de fait peut être amené à évoluer rapidement, en fonction des objectifs et du contexte dans lesquels ils sont détectés. Cependant, un terme étant avant tout l'expression d'une notion ou d'un concept unique dans le domaine concerné (et, de plus en plus, pour l'application concernée), il se peut qu'il ne soit pas entièrement adapté à notre propre objectif. En effet, rappelons que pour qu'une séquence soit considérée comme un terme, elle doit véhiculer une et une seule notion [ISO, 1999]. De notre côté, nous cherchons avant tout à dégager non des notions unitaires, mais des thèmes ou sous-thèmes de documents. Or, il paraît fort probable qu'un même thème puisse couvrir plusieurs notions ou concepts différents. L'objet terme est donc trop restrictif pour nos travaux.

C'est pourquoi nous nous penchons à présent sur la notion de collocation, souvent présente en linguistique de corpus.

4.1.2.2 La collocation en linguistique de corpus

Nous venons de voir que la notion de terme n'est pas entièrement satisfaisante pour notre objectif d'extraction de thèmes et sous-thèmes de documents ou d'ensembles documentaires. Nous avons besoin d'une unité de traitement plus large, pouvant englober plusieurs notions ou concepts à la fois, pour peu qu'ils soient vecteurs d'un thème ou sous-thème. C'est pourquoi nous nous intéressons à présent aux collocations, définissables de manière très générale comme des co-occurrences régulières. Reprenant notre exemple de *on demand transportation*, si la co-occurrence entre ces trois mots graphiques est régulière dans un document ou un corpus, alors elle peut être considérée comme une collocation.

Partant de ce principe, et de celui qu'une co-occurrence régulière peut véhiculer un thème ou sous-thème d'un document, nous pouvons chercher à extraire, de la manière la plus systématique possible, les collocations présentes dans les textes. [Sinclair et al., 2004], dans un rapport rédigé en 1970 mais rendu public seulement en 2004, définissent la collocation comme « the co-occurrence of two items in a text within a specified environment ». En l'occurrence, un item est un mot graphique. Selon les auteurs, une collocation est donc l'apparition conjointe de plusieurs mots graphiques dans un empan de texte donnée, matérialisé par une fenêtre plus ou moins importante

en nombre de mots entourant le mot ciblé. Leur définition se précise lorsqu'ils apportent des spécifications d'ordre statistique : « Significant collocation is regular collocation between two items, such that they co-occur more often than their respective frequencies, and the length of text in which they appear, would predict. [...] Casual collocation means "non-significant" collocation ». C'est donc la notion de régularité, de répétition au sein d'un corpus, qui rend une collocation significative.

C'est l'idée que reprend [Millon, 2011] lorsqu'elle définit, de manière générale, la collocation comme une « co-occurrence lexicale régulière ». Selon l'auteur, « [d]e par leur nature, c'est-à-dire en tant que combinaisons lexicales conventionnelles, et, surtout, en raison de leur omniprésence [...] dans la langue en usage (écrite et orale) [...] on trouve dans les corpus de textes, de langue générale ou de langue de spécialité, des co-occurrences lexicales régulières récurrentes ». Ainsi, au sein d'une langue de spécialité, il est très probable de trouver des combinaisons lexicales récurrentes particulières à sa norme en vigueur. En effet, dans un tel cadre, utiliser certaines collocations permet à l'auteur de se faire comprendre mieux, et plus vite, par son interlocuteur. C'est notamment le cas des textes académiques [Howarth, 1998], cité par [Millon, 2011].

En mettant en avant l'aspect récurrent d'une collocation, [Sinclair et al., 2004] adoptent un point de vue à la fois textuel et statistique, et se positionnent du côté du contextualisme. Ils fondent en effet la portée significative (*significance*) d'une collocation sur son attestation récurrente en corpus. En somme, une collocation attestée fréquemment dans un corpus est une collocation typique de ce corpus. Si cette collocation est typique, il se peut qu'elle porte un sens particulier, propre au corpus examiné, éventuellement l'un de ses thèmes ou de ses sous-thèmes. Il est à noter que si cela s'applique à un corpus dans son entier, cela peut également s'appliquer à une sous-partie de ce corpus ou à l'un des documents qu'il contient.

Cette approche contextualiste, quantitative, rattachée à la linguistique de corpus, s'oppose à l'approche phraséologique des collocations. En effet, du point de vue phraséologique, le lien lexical unissant les mots graphiques d'une collocation est arbitraire : les normes de la langue en usage empêchent un certain nombre de combinaisons entre mots graphiques, qui seraient pourtant syntaxiquement et sémantiquement possibles. Ainsi, ces combinaisons n'existent pas parce qu'elles ne sont pas « naturelles » pour les locuteurs natifs de la langue. Pour les tenants de l'approche phraséologiques, les collocations sont donc des « co-occurrences lexicalement restreintes » [Millon, 2011].

Cette approche considère que les collocations ne sont pas de simples co-occurrences, mais obéissent à des structures syntaxiques significatives, formalisées par des patrons syntactico-catégoriels, de type *Nom + Préposition + Nom* par exemple [Millon, 2011]. Cette approche diffère donc de l'approche contextualiste en ce qu'elle adopte un point de vue beaucoup plus formel de la collocation. De plus, elle ne la considère pas forcément en contexte d'usage dans des corpus de textes attestés, mais implique plus souvent des méthodes d'introspection. Dans

ce cas, le linguiste lui-même décide de l'acceptabilité ou de la non acceptabilité d'une séquence. L'inconvénient de cette approche est donc le faible accord entre linguistes, voire les contradictions au sein de travaux d'un même auteur. Des tests effectués par [Mel'cuk & Wanner, 1996] ont d'ailleurs montré que les jugements d'acceptabilité d'un ensemble de collocations soumis à un panel d'individus étaient extrêmement disparates, et qu'il était difficile d'arriver à un consensus pour bon nombre d'entre elles. La définition de la collocation comme co-occurrence lexicalement restreinte en est affaiblie, puisque la restriction lexicale ne va pas de soi pour les locuteurs, même natifs.

Puisque nous travaillons essentiellement sur corpus, et que nous nous situons de fait plus du côté de la linguistique de corpus que du côté de la linguistique introspective, nous mettons de côté l'approche phraséologique. Comme le préconise [Sinclair, 2004], nous prenons le parti de « faire confiance au texte » pour en dégager des collocations significatives. Dans notre cas, cela revient à considérer ces collocations comme potentiellement porteuses de thèmes ou de sous-thèmes des documents d'un ensemble documentaire.

4.1.2.3 Extraire les collocations pour détecter les thèmes

Nous venons de voir que les collocations, en tant que co-occurrences lexicales régulières, permettent de dégager des éléments significatifs au sein d'un corpus. Par *significatif*, Sinclair entend que ces collocations, en tant que combinaisons lexicales, sont caractéristiques d'un usage déterminé d'une langue. Cependant, par extension, nous pouvons postuler que ces collocations significatives, ou du moins une partie d'entre elles, véhiculent des thèmes ou sous-thèmes, au sein d'un seul et même document, d'un ensemble documentaire, ou d'une partie de cet ensemble documentaire.

Nous avons exposé, en 4.1.2.2 page 110, le caractère statistique de la collocation en linguistique de corpus. La notion de régularité fait en effet appel à un aspect statistique, et renvoie à une fréquence relativement élevée d'une co-occurrence donnée au sein d'un corpus.

Pratiquement, [Sinclair, 1991] définit ainsi la collocation :

> « Collocation is the occurrence of 2 or more words within a short space of each other in a text. The usual measure of proximity is a maximum of four words intervening. Collocations can be dramatic and interesting because unexpected, or they can be important in the lexical structure of the language because of being frequently repeated. »

Une collocation est donc composée de deux ou plusieurs mots graphiques. D'après cette définition, les mots en jeu dans la collocation peuvent être séparés par au plus quatre autres mots. Là encore, Sinclair, bien qu'il ne nie pas l'intérêt de collocations rares, appuie sur l'aspect récurrent des collocations révélant la structure du langage :

> « This second class of collocation, often related to measures of statistical significance, is the one that is usually meant in linguistic discussions. [...] Collocation in its purest sense, as used in this book, recognizes only the lexical co-occurrence of words. [...] In this book, the attention is concentrated on lexical co-occurrence, more or less independently of grammatical pattern or positional relationship. In most of the examples, collocation patterns are restricted to pairs of words, but there is no theoretical restriction to the number of words involved. »

Ainsi, pour l'auteur, la collocation est considérée en tant que co-occurrence lexicale, indépendamment de sa structure grammaticale. De cette façon, la collocation n'est pas restreinte à une partie du discours particulière. Cela permet, dès lors, d'obtenir des collocations verbales, adjectivales, adverbiales, etc.

La récurrence est également mise en avant chez [Stubbs, 2001] :

> « Repeated events are significant. The first task of corpus linguistics is to describe what is usual and typical. Unique events certainly occur, but can be described only against the background of what is normal and expected. The frequent occurrence of lexical or grammatical patterns is good evidence of what is typical and routine in language use. »

L'auteur appuie sur l'importance de cette récurrence des phénomènes observés. Son objectif n'est certes pas de dégager des thèmes à partir des collocations. Cependant, il recherche ce qui est typique dans la langue en usage. Or, nous pouvons considérer que ce qui est typique du point de vue lexical l'est du point de vue sémantique, et par là, peut amener au(x) thème(s) des textes du corpus étudié.

Dans notre contexte, ces collocations peuvent permettre, en fonction de leur portée et de leur fréquence, de différencier un thème global de l'ensemble documentaire d'un sous-thème : une collocation présente massivement dans un ensemble peut être représentative d'un thème général. En revanche, une collocation qui se manifeste sur un sous-ensemble de documents peut être considérée comme un sous-thème, qui peut servir de base à l'identification de groupes de documents, puisqu'à l'intersection lexicale de ceux-ci.

Ainsi, adopter la notion de collocation en linguistique de corpus nous permet d'avoir une unité paraissant pertinente pour notre objectif, c'est-à-dire l'extraction des thèmes et sous-thèmes des document d'un ensemble documentaire.

Cette extraction des collocations n'est pas effectuée pour elle-même : elle a une visée fonctionnelle, puisqu'elle doit permettre par la suite l'identification des thèmes et sous-thèmes par l'utilisateur. Ces unités sont donc destinées à être utilisées dans notre système d'immersion documentaire, et peuvent à cet égard endosser trois rôles : tout d'abord, elles permettront d'indexer les documents, en fonction des collocations qu'ils contiennent et/ou partagent. Elles peuvent

donc être un point d'entrée dans le système. D'autre part, elles sont exploitables en tant qu'axes de représentation, dans le cadre d'une navigation heuristique, puisqu'en liant plusieurs documents, elles créent des arcs entre eux, utilisables par l'utilisateur. Enfin, elles sont un facteur de sérendipité : l'exhaustivité de l'extraction permet en effet de laisser la place à des découvertes de l'utilisateur, qui peut alors emprunter des chemins qu'il n'avait pas envisagés avant de se trouver confronté à certaines collocations.

4.2 Les entités nommées : les unités privilégiées pour exprimer les connaissances contextuelles

Les connaissances que nous utilisons pour la constitution de notre ressource termino-ontologique, et par là pour celle du système d'immersion documentaire, sont d'une part des connaissances thématiques, et d'autre part des connaissances contextuelles. Nous venons de voir, dans la section précédente, que les connaissances thématiques peuvent s'exprimer, dans notre contexte applicatif, à travers les données textuelles contenues dans les documents, et en particulier à travers les collocations relevées. Les connaissances contextuelles, c'est-à-dire celles qui ont trait à la situation de production des documents, sont en revanche véhiculées par les métadonnées associées à chaque document. Ces métadonnées, en tout cas dans notre cadre de travail, prennent la forme d'entités nommées.

4.2.1 Origine et définition(s) du concept d'entité nommée

4.2.1.1 Les connaissances contextuelles pour la caractérisation des documents d'un ensemble documentaire

Nous avons jugé pertinent, dans le cadre d'un système d'accès à l'information, de nous appuyer largement sur les données « externes » caractérisant le document. Ces données qualifient les documents du point de vue de leurs conditions de production, et servent d'indicateurs et de clés d'interprétation pour l'individu en situation de réception du message que ces documents véhiculent (voir 2.4.1 page 57). En d'autres termes, la situation d'énonciation permet de typer chacun des éléments de la base documentaire, sur la triade bien connue du *qui, quand, où* ?

- l'auteur du document ;
- l'organisation dont il émane ;
- son lieu de publication ;
- sa date de publication.

Le dernier type d'information que nous avons exploité est le thème des documents. Nous avons considéré que ce thème était véhiculé par les suites de mots graphiques caractéristiques contenus dans les résumés de chaque document. Le thème est en somme la réponse donnée à la

question *quoi* pouvant s'ajouter à la triade. Et nous obtenons alors la répartition suivante :

- l'énonciation : l'énonciateur, en l'occurrence l'auteur et l'organisation dont il dépend, la date d'énonciation, le lieu d'énonciation, soit *qui, quand, où* ;
- l'énoncé : le thème, soit le *quoi*.

4.2.1.2 Quelles entités nommées pour quel objectif ?

Le concept d'entité nommée existe depuis le milieu des années 1990, et plus particulièrement depuis la conférence MUC-6 (Sixième édition de la Message Understanding Conference) de 1995 [Grishman & Sundheim, 1996]. Depuis sa première édition en 1987, cette conférence est dédiée à la promotion et à l'évaluation de la recherche en extraction d'information. Dans ce cadre, les entités nommées sont donc avant tout un objet pratique et ont été définies d'un point de vue applicatif. De fait, l'assise théorique de ces premières définitions est peu développée, avant tout parce que leur objectif premier était de répondre à des besoins pratiques. Finalement, en quinze ans de manipulation courante et de travail sur ces objets, très peu de chercheurs ont tenté une définition réellement appuyée sur la théorie linguistique. Pourtant, les entités nommées, en tant qu'unités de la langue, relèvent bien entre autres de cette discipline. Ehrmann propose une définition et une caractérisation des entités nommées et leur donne un statut d'objet TAL (Traitement Automatique des Langues), mais aussi d'objet linguistiquement défini et circonscrit. Dans la suite de ce chapitre, nous nous appuyons notamment sur ses travaux [Ehrmann, 2008] pour présenter les entités nommées, en axant cette présentation sur les aspects qui seront les plus utiles à notre argumentation.

L'entité nommée est née du TAL La dénomination du domaine n'a pas toujours été stable. L'appellation de TAL date en réalité du début des années 1990 selon [Cori & Léon, 2002]. Cette instabilité est due en majeure partie aux « tensions » inhérentes à la nature de ce champ : il est issu de domaines scientifiques très différents, et a des objectifs en apparence contradictoires.

Le TAL gravite en effet autour de plusieurs champs scientifiques, dont, de manière prégnante, la linguistique et l'informatique [Cori & Léon, 2002]. D'autre part, le TAL est partagé entre ses objectifs de recherche théorique et fondamentale, hérités des disciplines dont il est issu, et ses objectifs de recherche applicative et industrielle, qui furent saillants dès les débuts de la discipline. Cependant, il « semble avoir acquis un rôle fédérateur entre recherches théoriques [...] et applications » [Cori & Léon, 2002].

C'est également ce qui ressort de la définition du TAL donnée par Lebarbé :

> « Le traitement automatique des langues a pour objet la modélisation linguistique et informatique, fondée sur une analyse de corpus en contexte, afin d'exploiter les langues, résultant en des applications logicielles et/ou des enrichissements de la connaissance des langues. » [Lebarbé, 2010]

Ce qui caractérise le domaine scientifique du TAL est donc son caractère pluridisciplinaire et hautement applicatif : il est né de besoins en traduction automatique, et s'est inscrit dans un contexte de demande sociale et industrielle forte. Toutefois, le TAL est bien sous-tendu par des fondements théoriques, tantôt venant des disciplines dont il est issu au départ, tantôt appartenant au TAL lui-même. De plus, il apporte lui-même des réponses par l'enrichissement des connaissances sur les langues, en construisant des outils permettant de faire émerger des caractéristiques du langage et des langues. Ainsi, le « dialogue entre dimension théorique et exigences opératoires » est « caractéristique du TAL » [Ehrmann, 2008].

Alors que la plupart des objets utilisés par la discipline TAL sont importés d'autres domaines, les entités nommées ont été créées par le TAL. Pourtant, les éléments théoriques rattachés nativement au TAL sont plus de l'ordre des processus que des objets. Ehrmann distingue ces deux types d'unités, affirmant que le TAL est « plus concerné par des processus que par des objets proprement dits. Ces derniers ne sont toutefois pas absents, mais il s'agit pour une grande majorité de termes "importés" d'autres disciplines, n'étant pas à proprement parler des "objets TAL" » [Ehrmann, 2008]. Or, elle démontre que l'entité nommée, qu'elle caractérise comme un objet, fait figure d'exception relativement à cette configuration des éléments théoriques du TAL.

Le concept d'entité nommée est en effet né de besoins propres au TAL, et plus précisément à l'extraction d'information. Il a donc été créé par et pour la discipline TAL, dans un objectif concret.

En conséquence, la définition qui leur a été donnée a, avant tout, une portée applicative et donc pratique. La notion d'entité nommée est apparue pour la première fois, nous l'avons dit, dans le cadre de la conférence MUC. Les tâches MUC consistent à remplir des formulaires pour décrire un événement, en renseignant son type, l'agent de l'événement, le temps et le lieu, etc. Les corpus de texte à partir desquels est pratiquée cette extraction d'information ont une thématique variant au fil des conférences, pouvant aller de messages militaires à des rapports d'événements terroristes, en passant par la fabrication de circuits électroniques.

La tâche *entités nommées* est créée en 1995 dans le but d'identifier plus précisément et de manière « isolée » les acteurs d'un événement, ainsi que ses circonstances. De plus, les organisateurs veulent prouver par là que certaines « sous-tâches » participant de l'extraction d'information sont suffisamment autonomes pour être transposables rapidement et sans coût excessif d'un domaine à un autre. Dans ce cadre, l'objectif est d'extraire toutes les entités nommées apparaissant dans un texte. Il s'agit alors d'« identifier les noms de toutes les personnes, organisations et localisations géographiques dans un texte » [Grishman & Sundheim, 1996]. A cela s'ajoutent, dans les spécifications finales, « les expressions de temps, de montants monétaires et de pourcentages ».

Pour la conférence MUC-7, la définition reprend les mêmes éléments, de manière plus formelle, et moins *ad hoc* :

> « On the level of entity extraction, Named Entities (NE) were defined as proper names and quantities of interest. Person, organization, and location names were marked as well as dates, times, percentages, and monetary amounts. » [Chinchor, 1998]

La notion linguistique de nom propre (*proper name*) est avancée pour la première fois, puisqu'elle n'avait pas été citée strictement pour MUC-6, qui ne faisait référence qu'aux *organization names*, *person names*, etc.

[Poibeau, 2003] se positionne sensiblement de la même manière :

> « On appelle traditionnellement "entités nommées" (de l'anglais *named entities*) l'ensemble des noms de personnes, d'entreprises et de lieux présents dans un texte donné. On associe souvent à ces éléments d'autres syntagmes comme les dates, les unités monétaires ou les pourcentages repérables par les mêmes techniques à base de grammaires locales. »

Il ajoute cependant quelques éléments formels et méthodologiques complémentaires, puisqu'il parle de syntagmes d'une part, et de repérage par des grammaires locales d'autre part.

Toutes ces définitions restent énumératives, et orientées vers l'aspect informationnel du concept d'entité nommée beaucoup plus que sur la forme ou le contenu linguistique des éléments relevant de ce concept : même si les termes de nom propre et de syntagme apparaissent, la focalisation se fait sur la dimension extra-linguistique de ces éléments, sur leur intérêt relativement à un contexte. Pourtant, de par sa nature même, l'entité nommée a aussi à voir avec la linguistique [Ehrmann, 2008].

L'auteur relève, dans les systèmes d'extraction d'information existants, un certain nombre d'entités nommées. Nous reproduisons une partie de ces exemples en figure 4.1 page suivante. Il se dégage de cet ensemble une certaine hétérogénéité, et il peut sembler difficile, de prime abord, de trouver le fil conducteur permettant de justifier cette liste en tant qu'ensemble cohérent.

C'est pourquoi Ehrmann s'est attachée à donner une assise linguistique à la notion d'entité nommée :

> « Etant donnés un modèle applicatif et un corpus, on appelle entité nommée toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus. » [Ehrmann, 2008]

Cette définition recouvre un ensemble de notions qu'il nous paraît important d'expliciter quelque peu.

Tout d'abord, en employant le terme d'« expression linguistique », Ehrmann souligne le caractère linguistique de l'entité nommée, en tant qu'unité de la langue. Il est donc légitime de s'intéresser aux entités nommées de ce point de vue.

Renault
Coca Cola
l'Ile de France
la gare Montparnasse
Jacques Chirac
le PRésident du conseil
Paris
lyzozyme
le 21 avril
http://www.lemonde.fr
...

TABLE 4.1 – Exemples d'entités nommées dans différents systèmes d'extraction empruntés à [Ehrmann, 2008]

Concernant leur forme, les entités nommées sont liées de manière privilégiée aux noms propres. En effet, les personnes, organisations et lieux, comme *Boris Vian*, *Coca Cola Company* ou *Toulouse*, sont préférentiellement désignés par des noms propres.

L'autre manière de faire référence à des objets uniques du monde est d'utiliser des descriptions définies. Selon [Kleiber, 1981], une description définie a la forme minimale d'un syntagme nominal de type *le + substantif*, où le substantif doit être *individuant*, c'est-à-dire qu'il doit permettre d'identifier un individu au sein d'une classe donnée. C'est le cas par exemple des noms *voiture*, *livre* ou *ministre*, qui sont des substantifs permettant de désigner, en discours et donc dans un contexte donné, un élément unique du réel. Ainsi, *la voiture rouge*, mais aussi *la voiture*, peuvent désigner, en contexte, une voiture bien particulière, un référent unique et identifiable. Dans certains cas, le substantif peut également être *globalisant*, mais uniquement s'il est intégré à un syntagme réduisant à l'état d'objet individuel une notion abstraite et/ou continue. C'est par exemple le cas de *fer* dans le syntagme *le fer du portillon de l'entrée*, où la notion « continue » de *fer* est individualisée par le syntagme prépositionnel *du portillon de l'entrée*.

D'autres expressions linguistiques peuvent acquérir, selon le contexte, le statut d'entité nommée. Dans le tableau apparaît par exemple une adresse de site internet, *http ://www.lemonde.fr*. La seule restriction semble être l'obligation de l'unité de la langue en question de relever de la

catégorie nominale ou en tout cas de pouvoir y être assimilée.

D'autre part, la notion de référence est mise en avant. La référence est le lien qui existe entre une expression linguistique et l'élément du réel auquel elle renvoie. Mais la manière de référer à une entité peut varier en fonction de la forme de l'expression linguistique en question.

Pendant des années, les noms propres ont été considérés comme « vides » [Mill, 1843]. Ainsi de [Kripke, 1972], qui considère qu'un nom propre n'a pas de contenu lexical, de traits descriptifs, mais que ce qui lui permet de référer à un objet particulier du monde est uniquement une convention. Plus tard en revanche, [Kleiber, 2004] propose une autre vision des choses : le nom propre aurait bien un sens, mais un sens instructionnel plus que descriptif. Dans ce cas, le nom propre donne « l'instruction de chercher et de trouver dans la mémoire stable le référent qui porte le nom en question ».

Les descriptions définies, quant à elles, ont un sens descriptif : elles conduisent à la référence par des traits descriptifs constants, et non plus par un sens instructionnel ou procédural. L'élément du réel faisant l'objet de la référence doit donc correspondre aux traits, ou conditions d'application, véhiculés par la description définie employée. Les expressions temporelles (*le 21 avril 2002*, *2008*, etc.) et les expressions numériques (*20€*, *150 kg*, etc.) fonctionnent comme des descriptions définies.

Le référent lié à l'expression linguistique doit quant à lui être une entité unique du monde pour que cette expression rentre dans le champ des entités nommées. Les entités nommées désignent des particuliers. En cela, elles font référence à un élément unique et unitaire. Une entité nommée sous forme de nom propre désigne directement le référent ; en revanche, une description définie acquiert le statut d'entité nommée en désignant une instance particulière d'un concept - et non le concept lui-même ou une infinité de ses instances - grâce à ses propriétés descriptives.

Il est à noter que le particulier désigné n'a pas à être « unaire » (nous empruntons le terme à [Ehrmann, 2008]) : l'individuel ne s'oppose pas au collectif, mais au général. De fait, en fonction du niveau de granularité adopté, *Clio* en tant que modèle de voiture différent de tous les autres modèles de voiture pourra avoir le statut d'entité nommée, bien que dans un autre contexte, cette expression puisse désigner un ensemble d'objets.

La référence doit se faire de manière autonome. Le nom propre permet classiquement d'identifier directement le référent visé. Par exemple, *Robert Plant* permet d'accéder directement à l'élément du réel correspondant, soit le chanteur du groupe de rock Led Zeppelin. Nous notons simplement pour l'instant que cet état de fait n'est pas absolu, puisqu'un prénom non précisé par un nom de famille, tel que *Mick*, peut nécessiter un apport du contexte pour l'identification du référent.

Pour les référents désignés par des descriptions définies, les choses sont moins catégoriques. A côté de descriptions définies complètes qui permettent d'identifier le référent grâce des connaissances du monde mais sans avoir besoin de recourir au contexte, il existe des descriptions définies

incomplètes pour lesquelles la référence est moins immédiate. Ainsi, l'expression *le chanteur du groupe de rock Led Zeppelin* est autonome par rapport au contexte [33]. En revanche, *le chanteur* ne peut mener au référent que grâce aux informations qui seront fournies par le contexte entourant la production langagière. Dans ce cas, c'est le cadre applicatif, et en particulier le corpus de l'application, qui déterminera le degré d'autonomie de la description incomplète.

Ce cadre applicatif, se manifestant à travers les notions de modèle applicatif et de corpus, relève du domaine spécifique du TAL. C'est lui qui déterminera les unités linguistiques qui endosseront le statut d'entités nommées. En effet, ces dernières, contrairement aux noms propres dont elles peuvent pourtant partager certaines caractéristiques, n'existent pas dans l'absolu. Elles existent par et pour une application donnée, et donc en fonction d'objectifs et de besoins. Ces objectifs et ces besoins impliquent l'utilisation d'un corpus donné et la conception d'un modèle applicatif particulier. C'est donc la combinaison de ces paramètres fondamentalement applicatifs qui déterminera l'appartenance ou non d'une unité de la langue à l'ensemble « entités nommées ».

Un modèle applicatif détermine et structure un ensemble de données pertinentes pour une application donnée. Il spécifie les objets à traiter, ainsi que les relations entre ces objets. De fait, les unités qui devront être repérées en tant qu'entités nommées sont définies dans le modèle applicatif. Ce modèle peut intégrer divers niveaux de précision, qui influeront directement sur les résultats des traitements effectués.

Nous comprenons que le corpus, ici, est vu comme un ensemble de données langagières auxquelles est appliqué le traitement. Un corpus peut être de différentes natures et être plus ou moins spécialisé, de la langue « générale » à des sous-langages très restreints. C'est en grande partie son degré de spécialisation qui conditionnera l'autonomie des unités linguistiques en présence, et donc qui influencera la définition des entités nommées pour l'application visée. C'est ainsi que *le Président du conseil*, peu autonome dans un corpus très général, peut le devenir si tous les textes sont issus de comptes-rendus de réunions d'un syndic de copropriété par exemple : l'expression gagne en autonomie car le contexte fourni par le corpus permet d'identifier le référent visé de manière certaine. Dans ce cas, et si cette unité présente un intérêt du point de vue de l'application, alors elle pourra devenir une entité nommée pertinente.

Au vu de cette définition, et à l'image de *le Président du conseil*, toutes les expressions présentées en exemple ci-dessus peuvent relever de l'ensemble des entités nommées : en fonction de l'objectif de l'application et du modèle applicatif défini en conséquence, chacune d'entre elles peut acquérir ce statut. Il est certes difficile d'envisager que *lyzozyme* et *Renault* seront considérées comme entités nommées dans un même système. Cependant, une application visant

33. Mais son interprétation, comme celle des noms propres d'ailleurs, dépend grandement des connaissances du monde des locuteurs impliqués dans l'acte de communication. Il y a par exemple fort à parier que, durant une réunion du fan club officiel de Justin Bieber, la description définie *le chanteur du groupe de rock Led Zeppelin*, même très détaillée, ne se suffise pas à elle-même. Ce sont bien ici les connaissances du monde partagées - ou non - qui rentrent en jeu dans l'interprétation.

à extraire des informations dans un corpus de biologie pourra avoir besoin de modéliser *lyzozyme* comme une entité nommée ; l'expression *Renault* pourra être une entité pertinente pour une extraction d'information dans des textes financiers ou juridiques.

Confrontation de nos unités informationnelles à cette définition... et inversement Les unités d'information dont nous disposons sont, ainsi que nous l'avons mentionné en 3.2.2 page 84, des personnes, des organisations et des lieux, ainsi que des dates. A cela s'ajoutent les séquences de mots graphiques issus des résumés des documents. Nous présentons en figure 4.2 un échantillon de ces formes issues des métadonnées de divers documents. Elles sont présentes telles quelles dans les métadonnées de chaque document. A l'exception des suites de mots graphiques, extraites des résumés, elles sont tirées de leur en-tête, et leur type global est au moins partiellement identifié. Par exemple, *22-01-2007* est répertorié comme date, *Anna M. Blom* comme auteur, et *Nippon Paper Industries Ltd, Japan* comme organisation. *Homocysteine* et *geothermal circulation* viennent quant à eux des résumés.

CNRS - Universite Nancy I
Université Paris 7
Lund University, Department of Medicine
Laboratory of Immunology, *Georges-Pompidou European Hospital*
Laboratoire de Photonique et de Nanostructures, *Route de nozay 91460 Marcoussis*
Istanbul University. gdincsimsek@yahoo.com
Toshiba Corp. *University of Messina, Italy. www.unime.it*
Alcatel SA *Anna M. Blom*
Bruno O. Villoutreix *Danone SA, Paris, France*
22-01-2007 *09-10-1999*
Nippon Paper Industries Ltd, Japan
Northwestern University, Chicago
homocysteine *geothermal circulation*
encodes 18 exons

TABLE 4.2 – Exemples de formes issues des métadonnées des documents de la base de la société TKM

Du point de vue formel, certaines de ces expressions tombent de manière évidente sous la portée de la définition des entités nommées d'Ehrmann. Les noms d'organisation comme *Toshiba Corp.* et *Alcatel SA*, et les noms de personne comme *Anna M. Blom* ou *Bruno O. Villoutreix*, sont fortement autonomes quels que soient l'objectif et le corpus, réfèrent à une entité unique et ne sont donc pas sujets à discussion.

Il en va de même pour les dates qui, si elles n'ont pas une forme développée de description définie comme *le 22 janvier 2007*, n'en sont pas moins des expressions temporelles autonomes, qui sont une sorte de version abrégée de leur équivalent sous forme de description définie.

Une expression telle que *Université Paris 7* peut elle aussi rentrer dans les critères établis par la définition de départ. Il ne s'agit certes pas d'une description définie au sens strict : l'absence de déterminant défini déroge à la forme canonique *le + substantif*. Cependant, il peut là encore s'agir d'une forme courte, abrégée de description définie, qui s'explique par le fait que ces expressions ne s'insèrent pas dans un discours, dans un texte, mais font partie de métadonnées externes au texte auquel elles sont rattachées.

En revanche, l'expression complète tirée des métadonnées brutes comporte parfois plusieurs éléments, repérables visuellement de manière immédiate et, en quelque sorte, intuitive. Un cas de figure particulier peut d'ores et déjà être distingué des autres : une expression du type *Nippon Paper Industries Ltd, Japan* semble contenir deux types d'unités différentes du point de vue référentiel. Si *Nippon Paper Industries Ltd* désigne un objet de type organisation, *Japan* fait quant à lui référence à un objet de type lieu. Il est donc fondamental de parvenir à séparer ces deux types d'unités, qui ont de grandes chances d'endosser le rôle d'entités nommées.

A côté de ces expressions que nous pouvons qualifier de « multi-types », d'autres présentent une physionomie quelque peu différente. *Lund University, Department of Medicine*, ainsi que *Laboratory of Immunology, Georges-Pompidou European Hospital* ou encore *Istanbul University, gdincsimsek@yahoo.com* peuvent chacun faire l'objet d'un découpage en plusieurs unités, que nous schématisons dans la figure 4.2.

En conséquence, il est légitime de se demander si l'expression dans son entier est une entité nommée, si au contraire elle ne peut accéder à ce statut, ou bien encore si un traitement préalable doit être effectué sur la forme brute : contient-elle du bruit à éliminer totalement ? Un découpage doit-il être réalisé entre les différents éléments afin de dégager la ou les entités nommées en présence ?

Nous avons établi dans le chapitre 2 page 25 que l'objectif de ces travaux est de concevoir le modèle d'un système dans lequel l'utilisateur sera en mesure de naviguer, voire de s'immerger au sein d'un ensemble documentaire. Cette navigation devra lui permettre d'accéder à l'information pertinente en fonction d'un besoin donné, et non par le biais d'une structure rigide et imposée a priori. Les informations liées à la situation de communication de chaque document produit fournissent les éléments fondamentaux qui permettront le cheminement de l'utilisateur

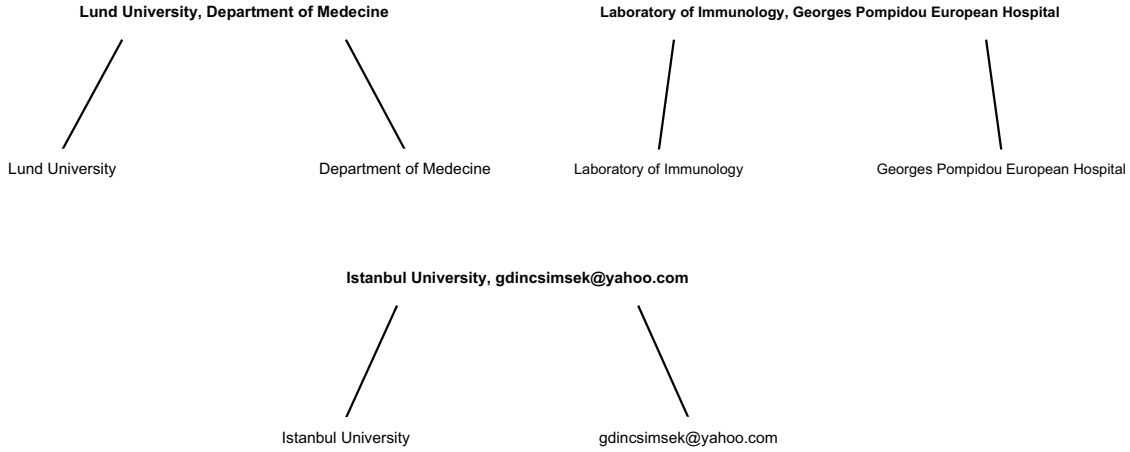

FIGURE 4.2 – Exemples de découpage en plusieurs unités pour trois expressions

en tant qu'interprétant, qui retrouvera dès lors un certain libre arbitre vis-à-vis du système, et déterminera, dans une situation de réception donnée, son information pertinente.

Il est donc nécessaire de modéliser ces informations contextuelles au sein d'une ressource de manière à ce qu'elles soient pleinement exploitables par le système de navigation. Puisque ces dernières sont préférentiellement exprimées par des entités nommées, ce sont ces entités nommées qu'il faut intégrer à la ressource.

Enfin, pour assurer la flexibilité du système, nous avons pris le parti de construire une ressource termino-ontologique multi-plans : c'est la combinaison à l'envi de ces plans par l'utilisateur qui permettra d'obtenir un grand nombre de points de vue différents sur un même ensemble documentaire.

Par conséquent, les entités nommées modélisées dans la ressource doivent être classées en fonction d'une typologie correspondant aux besoins d'exploitation, mais aussi intégrées au sein d'un ensemble structuré.

Dès lors, deux questions doivent être résolues : quels sont les types d'unités répondant aux besoins d'exploitation ? Et quel est le niveau de granularité nécessaire à une exploitation optimale ?

Types d'unités répondant aux besoins d'exploitation La plupart des unités présentées en exemple correspondent aux paramètres de la situation d'énonciation et ont un contenu informationnel fort. Nous avons établi en 3.2 page 79 que cette situation d'énonciation et ces indices sur le contenu de l'énoncé sont pertinents pour la navigation au sein d'un ensemble documentaire avec des objectifs de veille et de conseil stratégique en innovation. Les catégories nécessaires à l'adéquation du système avec les besoins exprimés se posent alors d'elles-mêmes :

nous avons besoin des auteurs de chaque document, des organisations publiantes, des dates et des lieux de publication. Enfin, nous devons dégager le thème des documents.

Or, pour que ces éléments soient exploitables, ils doivent être étiquetés et répertoriés en fonction de leur catégorie. En premier lieu, les entités nommées doivent être distinguées des suites de mots graphiques issues des résumés. D'autre part, les entités nommées elles-mêmes doivent être classées en fonction du type de leur référent. De fait, chaque unité sera intégrée à un plan déterminé de la ressource termino-ontologique sur laquelle s'appuiera le système de navigation, en fonction de son type. Sans surprise, un nom de personne sera intégré au plan des auteurs, un nom d'organisation au plan des organisations publiantes, une suite graphique issue du résumé d'un document au plan thématique, etc. En somme, les types d'unités à prendre en considération correspondent donc aux cinq plans différents de la ressource termino-ontologique, ainsi que représenté dans la figure 4.3.

Plan de la RTO	Unité correspondante
Auteurs	Noms de personnes
Organisations	Noms d'organisations
Lieux	Noms de lieux
Dates	Descriptions définies temporelles abrégées
Thèmes	Collocations brutes des résumés

TABLE 4.3 – Les cinq plans de la ressource termino-ontologique (RTO) et unités correspondantes

Nous avons vu que les unités révélant le thème des documents, dans notre cadre applicatif, ne relèvent pas du rôle des entités nommées, mais sont des suites de mots graphiques potentiellement répétés.

En revanche, les quatre autres types d'unités rentrent parfaitement dans la définition des entités nommées, du point de vue de leur forme et de leurs propriétés référentielles comme du point de vue de leur rôle au sein de l'application.

En miroir de ce constat, un autre s'impose : une unité qui ne relève pas de ces catégories ne peut pas, dans notre cadre applicatif, endosser le rôle d'entité nommée ni être considérée comme vecteur de thème. Par conséquent, tout ne doit pas être pris en compte dans l'expression : *Istanbul University, gdincsimsek@yahoo.com*. Nous l'avons mentionné, elle peut être prédécoupée intuitivement en deux séquences, soit *Istanbul University* d'une part et *gdincsimsek@yahoo.com* d'autre part. Au regard de la catégorisation qui vient d'être établie, la première séquence relève clairement d'une entité nommée, et plus précisément d'une entité nommée d'organisation. Par

contre, une séquence telle que *gdincsimsek@yahoo.com* ne s'intègre dans aucune des catégories et est considérée comme du bruit, puisqu'elle n'est pas exploitée.

Nous avons défini quatre types distincts d'entités nommées. Or, nous avons relevé dans les métadonnées des expressions telles que *Nippon Paper Industries Ltd, Japan*, où deux types différents d'entités nommées sont en présence. *Nippon Paper Industries Ltd* est une entité nommée d'organisation tandis que *Japan* est une entité nommée de lieu. Pour respecter le modèle, ces entités doivent être séparées, de manière à avoir une entité organisation et une entité de lieu. De la même façon, ce qui relève au sein d'une même expression de deux entités indépendantes du même type doit être séparé. Ainsi, dans *Univ Toulouse CNRS*, il est important de distinguer l'université de Toulouse du CNRS, sans quoi le caractère unitaire et unique de l'entité nommée ne serait pas respecté.

Degré de granularité nécessaire. Le degré de granularité d'un modèle doit être fixé en fonction des résultats attendus, eux-mêmes devant correspondre aux besoins des utilisateurs. En l'occurrence, l'objectif est de naviguer au sein de documents pour répondre à des questions stratégiques industrielles et/ou scientifiques, et les questions peuvent être de diverses natures. La diversité des demandes doit entraîner une diversité des accès aux documents interrogés. Cette variété dans les modes d'accès doit se faire à la fois du point de vue de la nature des unités, et d'autre part par le niveau de ces unités, lorsque cela est pertinent. Comme nous l'avons vu, les différents plans, ou facettes, de la ressource termino-ontologique remplissent ce rôle en ce qui concerne la nature des entités avec un type d'unité par facette. La diversité des besoins relativement au niveau des unités interrogées pour accéder aux documents motive l'existence d'une structure hiérarchique entre unités d'un même type, lorsque les unités s'y prêtent.

Il est donc fondamental de maintenir le lien d'appartenance qui existe dans une expression comme *Lund University, Department of Medicine* entre l'université de Lund et son département de médecine. Cela vaut également pour *Laboratory of Immunology, Georges-Pompidou European Hospital*, mais aussi pour *Paris* et *France* dans *Danone SA, Paris, France*. Concrètement, dans le modèle, cela passe par l'isolation de chaque entité distincte, mais également par la restitution de cette relation d'inclusion. En effet, sans la conservation de ce lien, il serait impossible d'assurer l'autonomie d'une expression telle que *Department of Medicine*, puisqu'il serait impossible d'identifier le référent auquel elle renvoie : un grand nombre d'universités ont un département de médecine, et rien ne permettrait alors de distinguer un département d'un autre. De cette manière, l'utilisateur aura le choix du niveau de granularité auquel il souhaite travailler. En vertu de ces besoins, nous considérons donc, dans notre cadre applicatif, que *Lund University, Department of Medicine* contient deux entités nommées distinctes, bien qu'incluses l'une dans l'autre. Le degré de détail pour une entité nommée doit aller aussi loin qu'il est nécessaire pour restituer l'ensemble des composantes pertinentes présentes dans une expression brute donnée.

Ainsi, quatre de nos cinq types d'unités relèvent de la définition des entités nommées d'Ehrmann. La seule catégorie ne rentrant pas dans ce champ est celle des suites graphiques relevées dans les résumés des documents, révélatrices de leur thème. Ceci est à corréler selon nous avec le fait qu'alors que les entités nommées viennent directement de métadonnées externes des documents tirées des en-têtes, les termes viennent eux de métadonnées « internes », dans le sens où le résumé constitue en lui-même une forme de discours, reprenant des éléments du texte du document, bien qu'il soit séparé visuellement et structurellement du corps du texte. De plus, alors que les entités nommées sont toutes liées à la situation d'énonciation, les termes ont à voir, eux, avec le contenu de l'énoncé lui-même. Il n'est donc pas surprenant que dans notre modèle, nous ayons été amenée à exclure les termes de l'ensemble entités nommées. Les autres unités relèvent classiquement de ce qui a été identifié comme des entités nommées dès l'apparition de la tâche d'extraction des entités nommées, par leur forme comme par leur contenu informationnel. De plus, dans notre cadre applicatif, chacune d'entre elles réfère à une entité unique du modèle, et ce de manière autonome.

Entité nommée : un rôle pour une expression linguistique ? Cependant, au vu de notre cheminement pour distinguer les entités nommées au sein des expressions linguistiques en présence, nous nous interrogeons sur la pertinence du statut d'objet du TAL accordé aux entités nommées. Le fait qu'elles relèvent du TAL ne fait pas de doute : elles relèvent clairement de la langue, et sont par ailleurs intimement liées à un cadre applicatif donné, à un modèle et à un corpus qui seront amenés à être déployés informatiquement. En revanche, leur nature d'objet peut soulever des questions : qu'entendons-nous généralement par « objet » ? Les entités nommées, telles que définies par Ehrmann, rentrent-elles dans ce cadre ? Si ce n'est pas le cas, quel est alors le statut des entités nommées ?

D'une manière générale, il peut être considéré que les objets d'un domaine scientifique donné ont une existence « stable » [34]. Il est à noter que nous ne postulons pas nécessairement une existence *a priori* de ces objets : en accord avec les théories constructivistes, nous partons du principe que les objets mêmes des sciences se construisent en interaction avec le sujet pensant, « l'agent de la science »[Juignet, 2008]. Cependant, même dans le cadre du constructivisme, ces objets acquièrent une certaine stabilité, à travers une intersubjectivité [von Glasersfeld, 1994]. Ainsi, même si la théorie sous-tendant un objet peut être discutable, même si sa définition peut être débattue, il existe en quelque sorte « une fois pour toutes », à l'intérieur d'un système de pensée donné si ce n'est dans le réel. Par exemple, l'objet linguistique le plus proche de l'entité nommée, le nom propre, existe en tant que tel, même si sa définition pose problème et si l'ensemble de la communauté linguistique ne s'accorde pas sur les unités de la langue rentrant dans ce champ.

34. En tout cas jusqu'à ce qu'une nouvelle connaissance naissant d'une avancée scientifique remette en cause cette stabilité, voire l'invalide.

Or, ce ne semble pas être le cas des entités nommées : une expression linguistique ne peut être intrinsèquement une entité nommée. Ehrmann affirme elle-même qu'elles « n'existent pas dans l'absolu » [Ehrmann, 2008], même si certaines unités de la langue accèdent à ce statut de manière plus « naturelle » que d'autres. Cet état de fait est dû à la nature applicative de la discipline du TAL. Cela n'exclut pas l'existence d'objets relevant du TAL. En revanche, nous estimons que le concept d'entité nommée, tel qu'il est défini dans une approche globale et théorique, relève moins de l'objet que d'un rôle potentiellement endossé par certaines expressions linguistiques, cette prise de rôle dépendant d'un cadre applicatif donné. En revanche, dans le cadre bien circonscrit d'un modèle applicatif et d'un corpus donnés, les entités nommées, c'est-à-dire les expressions linguistiques endossant ce rôle, peuvent être considérées comme des objets, dans le sens courant attribué au terme *objet* par le Petit Robert :

> « Chose solide ayant unité et indépendance et répondant à une certaine destination. »

Nous ne sommes certes pas dans le cadre de choses « solides », puisque la matérialité des entités nommées est langagière. Cependant, nous sommes face à des éléments qui ont une existence « tangible » dans le cadre d'une application, et qu'il est possible de ce fait de mesurer, de quantifier, de qualifier. Ils répondent alors à une destination, ils sont dotés d'une finalité, là encore en accord avec les axiomes constructivistes et en particulier avec l'axiome téléologique (voir à ce sujet 2.3.2 page 54). En somme, il devient alors possible de caractériser les entités nommées en présence, ainsi qu'il peut être fait d'un objet scientifique dans le cadre d'une théorie donnée.

C'est ainsi qu'une adresse mail ou une url comme `gdincsimsek@yahoo.com` ou `www.cnrs.fr` pourront être des entités nommées dans un modèle donné, et pas dans un autre. En l'occurrence, ces unités ne sont pas pertinentes pour nos travaux, et nous ne leur accordons donc pas ce rôle dans notre modèle.

Finalement, nous reprenons à notre compte la définition donnée par Ehrmann aux entités nommées, mais nous préférons les considérer comme un rôle du TAL, et non comme un objet. Quoi qu'il en soit, il apparaît que certaines expressions brutes nécessitent un traitement important de modélisation pour en dégager les entités nommées en présence. Les entités nommées de lieu et d'organisation, en particulier, doivent être extraites des mêmes expressions. D'autre part, certaines expressions contiennent plusieurs entités nommées, qu'elles soient indépendantes ou incluses l'une dans l'autre d'un point de vue référentiel. Cette modélisation passe donc par un découpage et une normalisation de ces entités, qui pourront alors être intégrées aux différents plans de la ressource.

4.2.2 La formalisation des entités nommées par la normalisation

Les formes tirées des métadonnées de nos documents sont riches en information, et leur exploitation fournira des points de vue potentiels pertinents pour une navigation interdocumentaire. Nous avons cependant constaté que telles quelles, elles sont inadéquates à une exploitation efficace, ne serait-ce que par la présence d'entités nommées relevant de types que nous tenons à distinguer, ou par l'accumulation de plusieurs entités nommées de même type dans une même expression. D'autres écueils justifient un traitement et une transformation de ces données avant qu'elles puissent accéder à un statut d'unités informationnelles cohérentes, et par là avant leur intégration à une structure de représentation des connaissances. A côté des problèmes soulevés précédemment (voir 4.2.1.2), le phénomène linguistique de synonymie, présent même au sein des entités nommées [Ehrmann & Jacquet, 2006], entraîne des difficultés d'interprétation qu'il convient de résoudre. Ce phénomène est particulièrement présent dans les métadonnées concernant les organisations. Les expressions présentées dans le tableau 4.4 en sont représentatives [35].

Forme souhaitée	Variante 1	Variante 2
Univ Toulouse 2 Le Mirail	Université de Toulouse 2	Univ Mirail
Alcatel	Alcatel SA	Acatel SA
Toshiba	Toshiba	Toshiba Corp.

TABLE 4.4 – Exemples de couples d'expressions synonymes issues des métadonnées des documents

Dans ces exemples, six expressions différentes réfèrent à trois entités seulement. *Université de Toulouse 2* et *Univ Mirail* réfèrent tous deux à l'objet du monde *Université de Toulouse 2 - Le Mirail*, mais le font de manière différente, en vertu de sens non équivalents, leur donnant la possibilité de désigner différemment un même référent (voir à ce sujet [Frege, 1892]). En outre, l'une contient la version abrégée du mot *Université*, et pas l'autre. *Toshiba* et *Toshiba Corp.* ne diffèrent que par la présence ou l'absence de la chaîne *Corp*, abréviation du mot *Corporation*, typique des noms d'entreprise. Enfin, *Acatel SA* contient une erreur typographique, ce qui n'est pas le cas de *Alcatel SA*.

Différence de désignation, erreurs de typographie ou variation des conventions de notation, tous ces cas peuvent être rassemblés, d'un point de vue pratique, au sein d'une synonymie de fait, qui ne peut perdurer dans un modèle de représentation des connaissances sans impliquer une incohérence du modèle d'un point de vue théorique, mais également des biais dans les traitements ultérieurs. Ainsi, au même titre que les phénomènes décrits précédemment, il est fondamental de régler ces difficultés.

35. Ces exemples sont attestés, issus des métadonnées rattachées aux documents de notre corpus.

Globalement, la formalisation des entités permettra à l'ensemble du système d'être fondé non plus sur des chaînes de caractère, mais bien sur des unités sémantico-référentielles cohérentes. Cette modélisation passe par une normalisation de ces entités, étape cruciale sans laquelle la cohérence du modèle ne serait pas assurée.

4.2.2.1 Qu'est-ce que normaliser ?

De manière générale, nous définissons la normalisation comme suit :

Définition 2. *La normalisation est la standardisation, en fonction d'une norme, de données à différents degré de granularité, allant de l'ensemble documentaire jusqu'au mot. Dans tous les cas, et quel que soit le niveau d'appréhension, cette standardisation vise à aplanir les problèmes posés par des différences dans les formes et / ou formats utilisés pour l'exploitation des données textuelles. A l'origine de ces différences se trouvent souvent des conventions de formatage ou d'écriture disparates et non coordonnées entre elles, et variant donc fortement en fonction de leur source.*

Ainsi, au niveau de granularité le plus haut, la Text Encoding Initiative (TEI) permet de normaliser des documents ou ensembles documentaires, de manière à fournir un format d'échange à visée universelle, grâce à des « directives de codage de texte pour l'élaboration et l'échange de documents électroniques » [Bonhomme et al., 1996]. Les objets de cette normalisation étant loin de nos entités nommées, nous ne développerons pas plus avant cet aspect [36].

En ingénierie des connaissances, et en particulier pour la conception d'ontologies ou de ressources termino-ontologiques, [Bachimont, 2000] parle de normalisation sémantique (voir 3.2.1 page 79). Celle-ci permet de contraindre l'interprétation des termes intégrés à une ressource, de manière à en expliciter la sémantique pour qu'ils désignent précisément les concepts en jeu.

L'important ici est de saisir le caractère crucial de la normalisation, qui est le seul moyen d'obtenir, quel que soit le palier auquel les unités à traiter sont considérées, des données exploitables informatiquement.

Dans la littérature de recherche académique, la normalisation des entités nommées a été relativement peu traitée pour elle-même. Elle est souvent décrite comme une étape préalable, et surtout annexe à la tâche principale dans laquelle elle s'insère. Pourtant, nombre de chercheurs sont conscients de l'importance de ce processus. [Alphonse et al., 2004] ont démontré, pour le projet Caderige, que la normalisation des entités nommées permettait d'optimiser largement les résultats, dans le cadre d'une tâche d'extraction d'information dans des textes bio-médicaux. [Poibeau, 2003], de son côté, qualifie cette phase de normalisation d'étape préalable essentielle pour l'extraction d'information financière.

36. Voir à ce sujet [Ide & Véronis, 1995].

Les milieux industriels s'intéressent à la normalisation des entités nommées, mais, pour des raisons évidentes de confidentialité, il est difficile d'obtenir des informations précises sur les méthodes appliquées. La société TEMIS [37] par exemple, spécialisée dans la conception et le développement d'outils d'intelligence économique fondés sur des approches de TAL, a développé un système de reconnaissance des entités nommées pour l'extraction d'information dans des domaines spécialisés comme les finances ou le domaine juridique dès sa création en 2000. Depuis quelques années, la normalisation des entités devient d'autant plus nécessaire que pour répondre à une demande nouvelle, il est nécessaire « de conceptualiser au maximum l'information extraite en s'affranchissant de la façon dont elle est exprimée dans la phrase » [Guillemin-Lanne & Six, 2006], de manière à structurer au mieux cette information. Les auteurs définissent la normalisation comme le fait « de s'abstraire de la forme de surface pour restituer avec régularité une même information, un même type d'information sous un même format ».

La normalisation « globale » au niveau du lexique, par contre, a donné lieu à un grand nombre de recherches, même si elle porte rarement le nom de normalisation et se cache derrière des termes tels que *normes de saisie*, *dépouillement*, *prétraitement*, etc. Dominique Labbé notamment [Labbé, 1990b], [Labbé, 2006], [Labbé & Labbé, 2006] prône une normalisation systématique des unités lexicales et a publié des travaux consacrés à cette pratique, selon lui inévitable dans le contexte du traitement de textes politiques, et plus largement dans le domaine de la lexicologie et celui, connexe, de la lexicographie. Dans le cadre de ses travaux sur le discours politique français contemporain, Labbé a mis en place pendant sept ans une chaîne de traitement informatisée complète de textes issus de discours de divers hommes politiques [38]. Le « journal de bord » tenu lors de ce travail a été publié sous forme de note dans [Labbé, 1990b]. Celle-ci récapitule les normes de saisie et de dépouillement des textes qui ont été définies et utilisées, de manière à rendre ces textes exploitables pour la recherche. Il part d'un double constat : d'une part, chaque chercheur, en fonction de ses objectifs, doit choisir entre plusieurs types de normes ; d'autre part, pour un traitement informatique fiable et intéressant des textes politiques, la saisie des textes doit obéir à des règles rigoureuses, et une lemmatisation doit être réalisée préalablement à tout traitement statistique.

Au final, ces « normes de saisie et de dépouillement » relèvent bien d'une normalisation des textes telle que nous l'avons posée plus haut, puisqu'il s'agit bien de standardiser les données de manière à « lisser » les différences empêchant un traitement « correct » des données. Le chercheur doit choisir entre une pluralité de normes, « parce que les philosophies et les buts qui motivent les dépouillements lexicographiques sont trop divers » [Labbé, 1990b]. Si cela est valable au sein même du domaine de la lexicographie, la disparité des besoins entre deux disciplines nécessitant

37. Site web de la société TEMIS : `http://www.temis.com/`

38. Le résultat d'une partie de ces travaux a donné lieu à la publication d'un ouvrage sur le discours politique de François Mitterrand : [Labbé, 1990a].

une normalisation est encore plus importante.

Les systèmes de normalisation dédiés aux entités nommées s'intègrent pour leur part à des applications souvent bien différentes de l'analyse lexicologique ou lexicographique. Il est à noter que paradoxalement, les travaux se consacrant entièrement à la normalisation des entités nommées relèvent plus souvent de la discipline de l'informatique, voire des disciplines liées au biomédical, que de celle du TAL proprement dit. La conférence BioCreative de 2004, par exemple, dédiée aux méthodes de fouille de textes en biologie moléculaire, a consacré une tâche entière à la normalisation des noms de gènes et de protéines [Hirschman et al., 2005]. Du côté de l'informatique, les travaux du laboratoire ISLA (Intelligent Systems Lab Amsterdam), en particulier, se consacrent entre autres à la normalisation des entités nommées [Khalid et al., 2008], [Jijkoun et al., 2008]. Leur étude porte sur la normalisation pour la recherche d'information, dans le cadre de systèmes Question-Réponse ou d'analyse de médias. Dans tous les cas, les systèmes traitent des articles d'information généraliste sur internet ou des contenus générés par des utilisateurs (*user generated content*), c'est-à-dire généralement des commentaires faisant suite à des articles journalistiques électroniques ou à des billets de blog.

Les problèmes mis en exergue pour une normalisation correcte concernent les deux difficultés majeures de ce type de tâche : la résolution des cas de synonymie et d'ambiguïté de certaines entités nommées, qui lorsqu'ils ne sont pas traités, affaiblissent nettement les performances des systèmes [Alphonse et al., 2004]. Résoudre ces deux problèmes est finalement la raison même pour laquelle la normalisation est nécessaire. En effet, les entités nommées sont soumises aux mêmes variations que les autres unités lexicales en discours [Ehrmann & Jacquet, 2006], et à ce titre, leur normalisation est loin d'être triviale.

L'impact de cette normalisation sur les résultats de la tâche globale dont elle est le préalable est important. De plus, elle se caractérise entre autres par son aspect transversal à plusieurs applications et domaines : une approche de normalisation non pas universelle, mais du moins partiellement transposable, présenterait un intérêt non négligeable. Nous postulons donc, à l'image de [Hirschman et al., 2005], qu'il est crucial de considérer la tâche de normalisation comme une tâche à part entière, qui doit être autonome, bien que dépendante, du processus global dans lequel elle s'insère. En effet, si le groupe TIA se penche sur le problème de la réutilisabilité des outils de constitution d'ontologies ou de ressources termino-ontologiques, il paraît également pertinent de s'intéresser à l'étape de normalisation des entités nommées, qui permet parfois la constitution de ces ressources.

Le caractère transposable peut sembler difficile à atteindre, vu la nature de rôle que nous avons attribuée aux entités nommées. Puisqu'un grand nombre d'expressions linguistiques peuvent devenir des entités nommées, il est difficile d'établir une méthode construite une fois pour toutes. Néanmoins, nous pouvons partir du postulat qu'un système de normalisation peut s'adapter à plusieurs applications pour peu que les types d'entités nommées soient partagés par ces applica-

tions, et que la norme utilisée soit adaptable en fonction des besoins.

Nous avons donné une définition générale de la normalisation au début de la sous-section 4.2.2.1 page 128. A présent, et pour le cas particulier des entités nommées, nous définissons la normalisation de la manière suivante :

Définition 3. *La normalisation des entités nommées est un processus permettant de ramener plusieurs occurrences d'une même entité nommée à une forme standard, établie par une norme et appelée type, et dont la finalité est de reconnaître et de regrouper toutes les occurrences d'une même entité malgré les phénomènes de variation inhérents aux unités de langue.*

La relation entre une occurrence d'entité nommée et son type est comparable à celle qui existe entre une forme graphique et son lemme : le type permet d'éliminer les problèmes de variation présents dans les occurrences en contexte.

La norme selon laquelle les entités sont normalisées est établie en fonction des besoins de l'application. Elle sera donc variable d'une application à l'autre, mais, et cela est crucial, stable au sein d'un même système.

Concrètement, le résultat de la normalisation est le regroupement des occurrences sous un même type d'entité. Il ne s'agit en aucun cas de supprimer purement et simplement ces variantes : il est en effet important de conserver les diverses occurrences rencontrées en contexte car cela permet de capitaliser l'information concernant les variantes existantes pour un même type.

4.2.2.2 Comment normaliser ?

La normalisation du lexique général : lemmatisation et racinisation. Dans les applications faisant appel au lexique « général », c'est-à-dire prenant en compte l'ensemble des formes lexicales en présence dans les données à traiter, la normalisation passe le plus souvent par des procédés visant à attribuer à chaque unité graphique une forme canonique.

[Labbé, 1990b] fait appel au principe de la norme formelle du laboratoire de lexicologie politique de Saint-Cloud, qui traite les formes graphiques en tant que telles, puis à la norme qu'il appelle « norme Charles Muller », qui consiste à lemmatiser l'ensemble des formes graphiques selon des règles strictes. La « norme Saint-Cloud » [Lafon et al., 1985], [Geffroy et al., 1973] « désigne les règles régissant la saisie des textes sur support informatique et le traitement des fichiers qui résultent de cette première opération » [Labbé, 1990b]. Labbé modifie légèrement ces règles pour les adapter à ses données. A cet égard, le point de vue qu'il adopte quant aux noms propres et autres formes nominales que nous intégrerions à notre ensemble d'entités nommées est révélateur des divergences de traitement résultant des divergences d'objectifs.

L'auteur ne s'intéresse pas au rôle d'entité nommée dans le cadre de son traitement. Cependant, il normalise des unités relevant, du point de vue de notre application, de ce rôle. Tout

d'abord, il considère les noms propres dans leur sens le plus restrictif, et n'inclut dans cette catégorie que les noms de personnes, de peuples et de lieux. Il est à noter que cette liste se rapproche peu ou prou des premières définitions des entités nommées, censées inclure en premier chef les noms de personnes, de lieux et d'organisations [Chinchor, 1998], [Grishman & Sundheim, 1996]. Il met cependant de côté les noms d'organisation, qu'il considère comme des « noms communs à majuscules » et dont la majuscule est ramenée à une minuscule en guise de normalisation. Il note à leur propos que leur inclusion dans la liste des noms propres n'est pas pertinente, entre autres parce qu'ils sont soumis à une certaine variabilité, comme *Fond monétaire*, *FMI* et *Fond monétaire international* [39], toutes ces expressions désignant la même entité.

Or, nous avons vu en 4.2.1.2 page 115 qu'il s'agit très exactement du type de phénomènes que nous devons traiter par la normalisation dans nos travaux. Il en va de même pour le cas des acronymes, où il est décidé que les versions développées ne leur seront pas assimilées. Ainsi, *PS* et *Parti Socialiste* resteront dans leur forme originale, puisque du point de vue de l'analyste de discours politique, « il y a une nuance entre les deux emplois et que le choix de l'une contre l'autre [forme] est porteur de sens » [Labbé, 1990b]. Là encore, dans nos propres besoins de normalisation, nous ne pouvons adopter une telle règle, puisque cela reviendrait à maintenir de la synonymie dans nos données, ce qui irait à l'encontre de nos objectifs, et ne présenterait de toute manière aucun intérêt notable. Cependant, la suppression des points entre les lettres d'un acronyme effectuée par Labbé va dans notre sens, puisqu'il s'agit en l'occurrence de réduire le nombre de variantes pour une même forme. De manière plus générale, un certain nombre de mots à graphies multiples n'ont pas été réduites à une seule forme. C'est par exemple le cas du verbe payer dans *paye* vs. *paie*. Encore une fois, il est flagrant que des objectifs différents amènent à des décisions différentes, voire contraires, puisque cette règle aurait été contre-productive pour nos travaux.

La deuxième phase de normalisation, que Labbé appelle le dépouillement, consiste à lemmatiser, selon des règles précises, et en accord avec les prescriptions de [Muller, 1967], les formes graphiques ainsi harmonisées. Il n'est pas fait de mention particulière des noms propres dans la note de Labbé, ce qui s'explique par le fait que lorsque ces derniers sont soumis aux variations flexionnelles, ce qui en tout état de cause n'est pas la majorité, ils le sont selon les mêmes règles que les noms communs. C'est notamment le cas des noms de peuples, comme *les Américains*. Là encore, une lemmatisation ne serait pas adaptée à nos objectifs de normalisation, mais cette fois plus en raison de la nature de nos données, soit des noms propres et descriptions définies, et de leur rôle d'entités nommées. En effet, une lemmatisation, en plus d'être inutile sur un nom propre comme « Toshiba », pourrait créer des incohérences sur des expressions comme *Laboratoire de Photonique et de Nanostructures*, où *Nanostructures* perdrait sa forme plurielle. En l'occurrence,

39. Les travaux dont est issu cet exemple, [Labbé, 1990b], ainsi que la rédaction du présent chapitre, sont antérieurs à l'éclatement de « l'affaire DSK ».

le nom officiel de l'entité désignée contient *Nanostructures* au pluriel et non au singulier, et cette désignation doit être respectée.

De même, des techniques de normalisation comme le stemming, ou racinisation, seraient de peu d'utilité pour nos données. Le stemming est « un processus qui consiste à retrouver la racine d'un mot par troncature, permettant de relier des mots ayant la même racine [...] » [Ligozat, 2006]. Le terme de « racine » n'est pas à considérer ici dans son sens linguistique, c'est-à-dire comme un « élément irréductible récurrent dans les formes lexicales apparentées par le sens et considéré en linguistique comme la forme la plus ancienne expliquant tous les dérivés ultérieurs » (*ibid.*). En effet, ces techniques sont fondées sur des algorithmes visant à réduire chaque mot d'une même « famille » à une suite de caractères commune, sans pour autant que cette suite de caractères soit motivée ou justifiée du point de vue de la morphologie en linguistique.

L'algorithme de Porter [Porter, 1980] est le plus connu et le plus utilisé à l'heure actuelle, et a bien à voir avec la normalisation. Selon son auteur, cet algorithme est « a process for removing the commoner morphological and inflexional endings from words [...]. Its main use is as part of a term normalisation process that is usually done when setting up Information Retrieval systems » [Porter, 2006]. Le principe de l'algorithme de racinisation de Lancaster [Paice, 1990], [Hooper & Paice, 2005], ou *Paice/Husk stemmer*, est le même, à cela près que l'algorithme est ici plus robuste. Le pendant négatif en est une sur-racinisation des unités lexicales. Quoi qu'il en soit, cette méthode de normalisation est à rapprocher de celle de la lemmatisation. Ces deux procédés sont utilisés couramment en extraction et en recherche d'information [Jacquemin, 2000], [Ligozat, 2006], [Heitz, 2006], [Korenius et al., 2004], en particulier durant la phase d'indexation, ou en extraction de termes pour la constitution de glossaires [Lebarbé, 2007]. Cependant, ils montreraient peu d'efficacité sur des entités nommées, c'est-à-dire sur des expressions linguistiques qui ne sont pas ou très peu affectées par les flexions et dérivations, et dans le cadre d'un besoin de modélisation de ces entités.

Ainsi, les objectifs comme les données jouent un rôle déterminant dans le choix des règles à appliquer aux unités textuelles pour les normaliser. Il est donc fondamental d'effectuer des traitements adaptés, et ce qui vaut pour un cas de figure ne vaut pas nécessairement pour un autre.

La normalisation des entités nommées : correction, regroupement, découpage et hiérarchisation. Les systèmes de normalisation dédiés aux entités nommées ont par conséquent des approches sensiblement différentes. Le premier point différenciant porte sur le fait que le traitement est appliqué spécifiquement à ces entités, à l'exclusion des autres unités lexicales contenues dans les données.

Ainsi que nous l'avons mentionné en 4.2.2.1 page 128, les champs disciplinaires s'intéressant particulièrement à la normalisation des entités nommées relèvent le plus souvent de l'informa-

tique ou du domaine du biomédical. L'ISLA [Khalid et al., 2008], [Jijkoun et al., 2008] est un exemple pertinent de cette activité de recherche autour de la normalisation dans le domaine de l'informatique.

Dans [Jijkoun et al., 2008], les chercheurs commencent par appliquer un système de reconnaissance d'entités nommées sur les textes et recensent cinq types de problèmes dans les résultats obtenus, dont certains sont récurrents dans la tâche d'extraction d'entités nommées, et qui sont à l'origine de ces phénomènes de variation :

- les erreurs de reconnaissance, et plus précisément des frontières mal placées autour des entités ;
- les multi-références, c'est-à-dire la juxtaposition de plusieurs entités nommées différentes, sans indices permettant de les distinguer les unes des autres ; elles sont en général dues à l'absence de ponctuation, en particulier dans les contenus générés par les utilisateurs ;
- les variantes pour une même entité nommée, dues à des erreurs de typographie, à l'utilisation de surnoms ou d'une seule partie d'un nom complet d'entité ;
- les entités non repérées, surtout en raison de l'absence de majuscules lors de la saisie par les commentateurs d'articles ou de blogs ;
- les entités nommées inconnues incomplètes.

Si les erreurs relatives aux frontières des entités nommées peuvent être imputées au système de reconnaissance lui-même, les autres sont inhérentes aux données et à la saisie qui en a été faite. Toutes sont dues aux phénomènes problématiques évoqués plus haut (voir la sous-section précédente), à savoir la synonymie et l'ambiguïté. Même dans le cas d'unités non repérées, leur non reconnaissance est due à une « synonymie » de fait puisque la présence ou l'absence de majuscule est un cas de variation menant deux unités de langue à désigner une même entité. [Jijkoun et al., 2008] résolvent ces difficultés grâce à l'utilisation de ressources externes, et plus précisément des informations fournies par l'encyclopédie en ligne Wikipédia [40] : les liens de redirection sont utilisés pour gérer les cas de synonymie, tandis que l'ambiguïté est résolue par l'exploitation des pages de désambiguïsation. Ils notent une nette amélioration des résultats de la tâche globale à effectuer par les systèmes mis en place lorsque la normalisation des entités nommées a été effectuée en amont. De plus, leur méthode a l'avantage d'être adaptable à plusieurs langues différentes, pourvu que les entités citées soient répertoriées dans Wikipédia. [Cucerzan, 2007] utilise lui aussi Wikipédia pour la désambiguïsation des entités nommées. Il procède pour cela à la comparaison de la représentation vectorielle du document traité avec celle de chaque entité potentielle faisant l'objet d'un article dans l'encyclopédie collaborative en ligne.

Pour des corpus d'ordre généraliste, l'apprentissage automatique peut également être exploité. [Magdy et al., 2007] procèdent à une normalisation inter-documentaire des entités nommées de personnes dans des corpus de presse. Pour cela, ils font appel à des méthodes d'apprentissage

40. `http://www.wikipedia.org/`

automatique appuyé sur un ensemble de traits lexicaux, orthographiques, phonétiques et morphologiques. Ces traits sont calculés en grande partie à partir de mesures de distance d'édition, de taux de recoupement entre deux noms en nombre de graphies ; l'utilisation très commune ou au contraire très rare de mots communs à deux entités nommées est également prise en compte pour décider d'apparier ou non deux noms potentiellement synonymes.

Ces méthodes sont efficaces sur des textes dits généralistes, comme c'est le cas pour ces corpus de presse ou de blogs. Nous pouvons raisonnablement supposer que les entités nommées présentes dans ces textes relèvent de connaissances du monde d'ordre général, et c'est pourquoi Wikipédia est très bien adapté dans ce contexte. Cependant, dès lors que les domaines abordés sont très spécialisés, et / ou que les entités nommées sont des noms de petites entreprises par exemple, il est très probable que ce type de ressource et les méthodes dans lesquelles il s'insère soient insuffisants.

Dans le cas des textes hautement spécialisés, les expressions linguistiques endossant le rôle d'entités nommées peuvent être aussi variées qu'il y a d'applications et de domaines en présence. Dans le domaine très étudié du bio-médical, [Alphonse et al., 2004] font appel à des méthodes fondées sur des patrons de reconnaissance pour gérer la synonymie en diachronie comme en synchronie, sur les noms de gènes et de protéines. [Cohen, 2005], quant à lui, génère totalement automatiquement des dictionnaires ayant la forme de thésaurus à partir de bases de données de gènes et protéines en ligne. De cette manière, les différents termes référant à la même entité rencontrés dans les textes sont ramenés au concept les englobant. Dans le même ordre d'idée, [Zhou et al., 2007] utilisent des bases de connaissances et des listes d'abréviation biomédicales pour apparier les synonymes ou les variantes lexicales d'un même gène ou d'une même protéine, et pour repérer les hyperonymes et leurs hyponymes. Dans le cadre de la conférence BioCreative de 2004, le système de [Hanisch et al., 2005] obtient la meilleure performance en termes de normalisation des entités nommées de gènes et protéines grâce à la combinaison de plusieurs méthodes : ils combinent des dictionnaires, l'appariement de chaînes proches à l'aide de mesures de proximité, et une liste d'abréviations du domaine biomédical. De plus, ils prennent en compte les fréquences des unités lexicales et règlent leur système en fonction du corpus d'entraînement. Ils font donc intervenir à la fois des mesures statistiques et des méthodes symboliques à base de dictionnaires.

Ainsi, en fonction du degré de spécialisation, les mêmes méthodes ne peuvent être employées. En effet, en ce qui concerne les textes spécialisés, il est difficile de trouver des ressources externes généralistes ayant une couverture suffisante, d'autant plus que de nouvelles entités nommées apparaissent à un rythme difficilement tenable pour la maintenance de telles ressources. Du côté des textes spécialisés, dont les textes biomédicaux sont les représentants les plus largement étudiés à l'égard de la normalisation, il existe pléthore de méthodes, d'approches, de processus, de systèmes. De plus, certains systèmes combinent un nombre conséquent d'approches, piochant

tantôt dans les méthodes statistiques, tantôt dans les méthodes symboliques. Et, si les expressions linguistiques ayant le rôle d'entités nommées sont massivement des noms de gènes et de protéines, il n'en demeure pas moins que certaines approches, au moins dans leur principe général, peuvent être exploitées dans des domaines variés.

Réciproquement, un système de normalisation traitant des textes plus généralistes, tels des corpus de presse, où les entités nommées à traiter appartiennent aux catégories plus « classiques » des noms de personnes, d'organisations ou de lieux, n'aurait que faire de dictionnaires de protéines et de gènes.

De fait, en plus de l'objectif final de l'application, le degré de spécialisation des textes semble conditionner le choix des expressions linguistiques prenant le rôle d'entité nommée et devant être normalisées, ce qui influence considérablement la méthode de normalisation et les ressources choisies.

Enfin, le degré de structuration des données à traiter joue lui aussi un rôle important dans les méthodes et approches utilisées. En effet, une base de données, par exemple, ne sera pas traitée de la même manière que des textes dans lesquels aucune structure, hors la ponctuation et la « présentation » du texte, ne permet d'identifier le type d'information et d'isoler les données. La première appartient à la catégorie des données structurées [Tannier, 2006], et est constituée par définition de champs isolés contenant des informations dont le type est déjà identifié, et ce pour chaque entrée. Le second, au contraire, s'intègre à la catégorie des documents non structurés, ou « documents plats » (*ibid.*). A mi-chemin entre données structurées et non structurées, les données semi-structurées comportent, en plus d'un texte « libre » et non structuré, des données généralement placées dans des balises et fournissant des informations variées sur le texte. Typiquement, certains documents au format XML relèvent de cette catégorie. Dans ce dernier cas, les traitements dépendront de quelle partie du document est à traiter : les données placées à l'intérieur des balises pourront être traitées par des procédés semblables à ceux utilisés pour les données structurées, tandis que le texte libre ne diffère pas de celui de données non structurées. Les méthodes décrites jusqu'ici sont toutes consacrées au texte libre, et donc aux données non structurées.

La différence majeure des traitements des données structurées par rapport à ceux des données non structurées réside dans le fait que pour la plupart, les informations pertinentes pour une application donnée sont déjà identifiées, et plus ou moins circonscrites et délimitées. Cependant, des données structurées ne sont pas nécessairement des données non bruitées : les bases de données posent un certain nombre de difficultés, et en particulier celui de l'harmonisation des champs communs à plusieurs entrées. [Elmagarmid et al., 2007] font un recensement très complet des méthodes existantes pour résoudre l'un des problèmes majeurs posés par les bases de données, à savoir la présence de doublons. Ces doublons sont dus à des divergences dans les conventions de notation des diverses sources, à des erreurs de transcription, ou encore à des informations

incomplètes.

Par conséquent, pour éliminer des doublons, il est nécessaire de détecter les variantes textuelles pour un même contenu de champ dans différentes entrées. Les auteurs répertorient un ensemble de prétraitements des données avant de pouvoir mettre en œuvre la détection de doublons proprement dite. La première étape est l'identification et le découpage des éléments individuels de données dans les fichiers source lors de l'import dans la base de données, ce qui permettra de s'affranchir des problèmes posés par des chaînes de caractères et de comparer des composants individuels. La standardisation des données est elle aussi évacuée de la procédure de détection des doublons en elle-même. Il s'agit de convertir en une représentation unique l'information issue de différentes sources où les standards diffèrent. Les auteurs prennent en exemple des adresses, des dates et des heures, ainsi que des noms de personne. Nous sommes donc bien face à des expressions linguistiques privilégiées pour jouer le rôle d'entités nommées. Pour les adresses par exemple, l'objectif est d'harmoniser des entrées telles que *44 W. 4th St.* et *44 West Fourth Street*, en choisissant un nouveau standard unique pour toutes les entrées.

Selon [Elmagarmid et al., 2007], « data standardization is a rather inexpensive step that can lead to fast identification of duplicates ». Il est à noter que dans les exemples cités ne figurent pas les noms d'organisation, qui peuvent quant à eux atteindre une grande complexité. Une fois les données standardisées, la détection des doublons elle-même est lancée. Puisque la plus grande source de variation est à ce stade la variation typographique, ce sont les techniques de comparaison de chaînes qui sont employées en priorité. Des mesures de similarité peuvent être calculées à la fois sur les caractères, les tokens ou les phonèmes, en fonction des problèmes posés par la variation lexicale.

Pour comparer entre eux plusieurs champs de deux entrées différentes, et non plus un seul, d'autres méthodes plus complexes sont mises en place. Elles se divisent en deux catégories majeures :

- les méthodes d'apprentissage, avec des techniques d'apprentissage supervisé et/ou des approches probabilistes ;
- les approches s'appuyant sur des connaissances du domaine concerné et des mesures génériques de distance : des langages déclaratifs sont utilisés pour l'appariement, et des mesures de distance adaptées à la tâche de détection de doublons.

Les méthodes, qu'elles s'appliquent à la comparaison de champs isolés ou d'ensembles de champs, sont donc pour la plupart fondées sur des mesures de similarité. Elles permettent de répondre à la difficulté majeure posée par l'hétérogénéité lexicale à l'origine des doublons, qui peut être considérée comme de la synonymie.

[Elmagarmid et al., 2007] présentent des méthodes de dédoublonnage, ce qui implique de fait un besoin de normalisation : normaliser les contenus des champs permet d'éliminer les entrées superflues. Cependant, l'autre problème récurrent dans la normalisation des entités nommées,

l'ambiguïté, n'est pas abordé, même s'il est vrai que ce phénomène est moins présent dans des données structurées : le classement par type des champs de bases de données réduit en principe ce problème à une ambiguïté intratypique. Nous verrons néanmoins, dans la sous-section 7.1.2.2 page 257 du chapitre 7, que ce type d'ambiguïté pose des difficultés qu'il convient de résoudre, en particulier pour les entités nommées de villes.

En somme, les approches décrites dans la littérature diffèrent principalement en fonction de quatre critères, par ailleurs souvent liés par des relations d'implication :

1. la nature ou le rôle des expressions linguistiques à normaliser ;
2. l'application finale à laquelle les données langagières sont destinées ;
3. le degré de spécialisation des textes ;
4. le degré de structuration des documents.

Il est à noter que l'un des problèmes que nous rencontrons dans nos données n'est abordé dans aucune de ces approches. En effet, nous avons expliqué en 4.2.1.2 page 115 que notre application nécessitait de maintenir le lien entre le département et son université de rattachement dans le nom *Lund University, Department of Medicine* par exemple. Le maintien de ce lien implique d'une part que les deux parties de cette entité devaient être clairement délimitées, et d'autre part qu'un lien hiérarchique soit établi entre les deux une fois la délimitation effectuée. Or, aucun des travaux que nous avons examinés ne traite cet aspect, qui nous semble pourtant bien appartenir à la tâche de normalisation, en tout cas pour des applications comme la nôtre : structurer hiérarchiquement les entités nommées qui s'y prêtent permet en effet de réduire l'ambiguïté, en ce sens que conserver tel quel un nom d'organisation comme *Department of Medicine* par exemple, sans lien le rattachant à son université d'origine, serait une source d'erreurs importante pour un système d'immersion tel que le nôtre.

4.2.2.3 Normalisation et extraction d'entités nommées : besoins de normalisation par type d'entités

Nous avons vu en 4.2.1.2 page 115 que les entités nommées que nous devons traiter sont loin d'être placées dans des données « propres », bien qu'il s'agisse de données structurées, et plus précisément du contenu de champs de bases de données.

Rappelons les types d'entités nommées en présence dans nos données, accompagnés de quelques exemples :

- organisations : *Toshiba Corp.* ; *Lund University, Department of Medicine*
- personnes : *Anna M. Blom* ; *A.M. Blom*
- lieux : *France* ; *Japan* ; *USA* ; *United States* ; *Chicago* ; *Paris*
- dates : *2006* ; *02-27-2005*

La nature ainsi que la forme de ces entités variant d'un type à l'autre, les stratégies de normalisation devront être adaptées à la norme particulière à chaque type. En effet, la normalisation participe grandement à l'engagement sémantique pris pour les unités de la ressource termino-ontologique conçue. Rappelons que l'engagement sémantique correspond au niveau de spécification formelle permettant de restreindre l'interprétation de chaque concept et ainsi d'en donner la sémantique [Bachimont, 2000]. Dans notre cas, une normalisation plus ou moins fine des entités nommées en présence influe sur la qualité de la désambiguïsation des unités contenues dans certains des plans de la ressource (voir 3.2.1 page 79). Il convient donc de s'assurer que chaque type d'information est normalisé en adéquation avec sa nature d'une part, et avec l'utilisation qui en sera faite d'autre part.

Les dates font l'objet du traitement le moins complexe : elles sont déjà fortement standardisées dans les données brutes importées dans la base de données. Puisque les informations sont tirées de documents dont les genres sont extrêmement contraints (rappelons qu'il s'agit de brevets et d'articles scientifiques), il existe relativement peu de variation interdocumentaire. Ainsi, une date « brute » sera formée soit d'une année de publication, soit d'un jour précis comportant les éléments *année*, *mois* et *jour*. Quand le jour est spécifié, il l'est toujours sur le modèle : *aaaammjj*, comme dans *20050227*. De fait, le travail de normalisation consiste à séparer dans des champs distincts, le cas échéant, le jour, le mois et l'année.

Les auteurs sont plus sujets à variation, et leur traitement atteint donc un degré de complexité supérieur. La variation la plus problématique concerne notamment l'accumulation des noms de plusieurs personnes au sein d'une seule chaîne de caractères, ainsi que la distinction entre noms et prénoms pour une même entité. Pour limiter le problème, des règles simples de standardisation et de découpage ont été mises en place par les ingénieurs du service R&D. Ces règles sont antérieures au début de nos travaux, et interviennent au moment de l'import des documents. Elles permettent de standardiser ces noms non de manière optimale, mais du moins de façon relativement fiable. Elles agissent en particulier sur l'identification des noms « individuels », et sur la distinction entre nom et prénom(s) pour une même personne.

Les lieux se déclinent en réalité en trois sous-types d'entités nommées : les noms de pays, les noms de ville et les adresses. Chaque type doit faire l'objet de traitements adaptés.

Les noms de pays peuvent prendre dans les données brutes la forme de noms développés ou de codes pays ISO. Les codes pays sont « des chaînes de caractères alphabétiques ou numériques, utilisées pour coder les pays à des fins de traitement de l'information » [41]. Les codes pays ISO

41. source : Wikipedia. `http://fr.wikipedia.org/wiki/Code_pays`

sont issus de la norme ISO 3166 [42], et sont constitués de deux lettres. Pour standardiser les noms de pays, une liste de correspondance entre codes ISO et noms développés a été établie. C'est la forme du code ISO qui a été sélectionnée pour être le modèle d'entité après normalisation, de manière à éliminer tout risque d'ambiguïté.

Les noms de ville sont identifiés dans un nom d'organisation grâce à une liste d'environ 2 700 000 villes du monde [43]. De plus, chaque ville est associée au pays correspondant, ce qui permet une désambiguïsation. Les coordonnées géographiques sont associées à chaque ville, ce qui permet par la suite un positionnement sur une carte géographique.

Les adresses sont repérées grâce à des lexiques contenant des amorces d'adresse. Une amorce est un mot fréquemment associé à un type donné d'entité. Par exemple, dans du texte libre, *Monsieur* sera couramment associé à des noms de personne. En l'occurrence, les amorces permettent de savoir, de manière fiable, que l'entité est une adresse, et peuvent être des formes développées ou abrégées, comme *Street/St.*, *rue*, *Cedex*, etc. Ces lexiques sont couplés à des règles de repérage lexico-syntaxiques. Rappelons que les entités de lieux se trouvent dans le champ « nom d'organisation » rattaché à chaque document. Les traitements dévolus aux noms de lieux sont donc fortement associés aux traitements appliqués aux entités organisations. Cela est d'autant plus vrai que si dans certains cas, les noms de lieu sont syntaxiquement séparés des noms d'organisation, ils sont parfois à aller chercher directement dans l'entité nommée d'organisation. Dans cette configuration, il est donc nécessaire de travailler à un double niveau de détail. Les deux noms suivants illustrent cette différence :

- *Nippon Paper Industries Ltd, Japan*
- *Université Paris 7*

Le cas particulier des noms d'organisations Reprenons quelques exemples :

- *Lund University, Department of Medicine*
- *CNRS - Universite Nancy I*
- *University of Messina, Italy. www.unime.it*
- *Nippon Paper Industries Ltd, Japan*

Le premier constat à tirer de ces cas par rapport aux autres types est que les entités les plus complexes à normaliser sont les noms d'organisation. Sans même évoquer les problèmes de synonymie et d'ambiguïté entre plusieurs entrées de la base de données, les séquences identifiées comme « organisations » dans cette base peuvent contenir, en plus des noms d'organisations, des entités d'un autre type, et en particulier des noms de lieux. D'autre part, plusieurs entités nommées, indépendantes ou liées par une relation d'inclusion, peuvent apparaître dans une seule séquence. Enfin, des éléments d'information non pertinents pour notre application peuvent être

42. Page de la norme ISO 3166 sur le site de l'Organisation Internationale de Normalisation : `http://www.iso.org/iso/fr/country_codes.htm` (consultée le 3 avril 2011).

43. Liste disponible gratuitement sur le site : `http://www.maxmind.com/app/worldcities`

présents et parasiter les noms d'entités nommées (voir la sous-section 4.2.1.2 page 115). Dans ces cas-là, la normalisation passe obligatoirement par une étape d'extraction d'entités nommées, ce qui revient, selon les cas :

- à découper correctement les différentes entités nommées en présence ;
- et/ou à éliminer le bruit de manière à conserver seulement le nom de chaque entité nommée isolée.

Ainsi, la normalisation passe aussi par une phase d'extraction des entités nommées. Selon [Poibeau, 2001], citant la classification de [Sekine & Eriguchi, 2000], il existe trois types de systèmes de reconnaissance d'entités nommées. Les systèmes fondés sur une base de règles écrites à la main définissent les patrons permettant d'extraire les entités nommées. Bien que cette méthode soit la plus ancienne, elle est encore largement utilisée aujourd'hui. Des règles peuvent également être obtenues par des systèmes fondés sur de l'apprentissage, utilisant des techniques pour développer un modèle à partir d'un corpus annoté. Ce modèle permet alors d'étiqueter correctement les entités nommées dans les textes. Les systèmes par apprentissage peuvent aussi donner lieu à un modèle numérique ou à un arbre de décision. Poibeau note que l'inconvénient majeur de ces systèmes est qu'ils ne rendent pas toujours possible une intervention humaine sur le modèle. Enfin, les approches mixtes mêlent système de règles et apprentissage automatique ou semi-automatique à partir de ces règles. L'utilisateur expert intervient en amont ou en aval de l'apprentissage, afin de le contrôler. Ces approches nécessitent souvent d'avoir des dictionnaires en entrée des traitements.

Quoi qu'il en soit, exécuter cette tâche sur nos données diffère des tâches classiques d'extraction sur deux points, qui sont la conséquence directe du fait que nos données sont structurées, et ne se placent donc pas en discours au sens habituellement entendu. Tout d'abord, le nombre de types distincts d'entités est limité à un ou deux par champ. En l'occurrence, le champ « nom d'organisation » ne peut contenir que des noms d'organisations, comme son nom l'indique, et éventuellement - de manière récurrente - un ou des noms de lieux. D'autre part, la syntaxe dans laquelle s'insère l'entité est extrêmement réduite. Ce dernier point présente un avantage majeur, puisque cela induit un nombre de schémas d'occurrence relativement limité. La contrepartie de cet état de fait est que le contexte pouvant permettre d'identifier une entité fournit beaucoup moins d'indices que si les entités étaient placées en discours. Ainsi, des indices lexico-syntaxiques peuvent s'avérer insuffisants pour découper correctement deux entités cooccurrentes ou pour éliminer les éléments parasitaires.

D'autre part, normaliser des données structurées ne revient pas (obligatoirement) à dédoublonner des entrées. A la différence des systèmes décrits par [Elmagarmid et al., 2007], l'objectif de notre système n'est pas de dédoublonner les entrées de la base de données. Nos besoins portent sur des champs isolés et non sur des entrées entières, et les méthodes citées pour la comparaison de champs multiples, très utiles pour des bases de données commerciales par exemple, ne

s'appliquent pas ici. En effet, en raison du haut degré de standardisation induit par le genre très contraint des brevets et articles scientifiques, le risque de récupérer des entrées en doublons est très faible. Les titres des documents, en particulier, ou les numéros de dépôt pour les brevets, servent d'identifiants et limitent les possibilités d'erreur. L'objectif n'est donc pas de repérer les documents présents en plusieurs exemplaires dans la base, mais bien d'harmoniser leurs méta-données pour pouvoir exploiter ces dernières dans l'immersion, et accéder aux documents qu'elles caractérisent. C'est l'immersion qui permettra ensuite de procéder à des analyses qualitatives et quantitatives. A ce titre, il est fondamental de savoir par exemple qu'un document publié par l'*Université de Toulouse 2* a été publié par le même référent que celui publié par l'*Université du Mirail*.

Les problèmes liés à l'extraction de la forme correcte des entités en présence jouent à un niveau individuel, que nous qualifions d'intra-entité. Pour le contenu d'un champ donné, et pour une entrée, la ou les entités doivent être correctement délimitées. Soit plusieurs entités sont présentes dans un même champ, relevant d'un seul ou de deux types, soit la seule entité contenue dans le champ doit être « nettoyée » du bruit qui l'entoure. D'autre part, dans le cas particulier de la co-présence de deux ou plusieurs entités unies par des relations d'inclusion, le sens de l'inclusion doit être restitué sous la forme d'une hiérarchie. A cet égard, nous touchons ici clairement à la formalisation de ces entités nommées. Concernant la hiérarchie des entités repérées, nous verrons dans la deuxième partie de ce document que deux types de variation posent problème : non seulement la variation des entités nommées dans la langue, mais également celle qui peut exister dans le « monde réel ». En effet, il est impossible de tirer une structuration « universelle » pour l'ensemble des entités nommées de catégorie « université » par exemple, puisque la structure réelle des universités varie grandement d'un pays à l'autre, voire d'une université à l'autre, et même au sein d'une même université dans le temps.

Le reste du travail de normalisation se place plutôt au niveau de la comparaison du contenu d'un champ pour une entrée donnée avec tous les autres champs de même type. A ce titre, nous qualifions ce niveau d'inter-entités. C'est avant tout la présence d'une multitude d'autres entrées qui nécessite une harmonisation des standards de notation à des fins de comparaison. Il s'agit alors de corriger la variation typographique, et d'harmoniser la notation de certaines entités ou de composants d'entités. La notation varie soit en fonction de conventions établies dans les différentes sources, soit en raison du caractère multilingue des données. Par exemple, le composant d'entité *Université* pourra être noté *University*, *Université*, *Univ*, etc. Ce dernier travail correspond à la phase de standardisation chez [Elmagarmid et al., 2007].

Chapitre 5

Les processus en jeu dans la modélisation, la conception et l'exploitation d'un système d'immersion

Sommaire

Nous avons présenté dans les premiers chapitres les différents éléments rentrant en jeu dans l'élaboration d'un modèle pour un système d'immersion, et ce dans le cadre du traitement d'ensembles documentaires scientifiques et techniques pour le conseil en stratégie de l'innovation. Ces éléments, c'est-à-dire les entités nommées (EN), la ressource termino-ontologique (RTO) multi-plans et le système d'immersion documentaire (SID) dans son ensemble, se situent les uns par rapport aux autres, selon un ordre d'inclusion, avec un niveau de granularité allant du grain le plus fin au grain le plus gros :

Entités nommées

$\subset$ *Ressource termino-ontologique multi-plans*

$\subset$ *Système d'immersion*

En effet, les entités nommées (et les suites de mots graphiques) sont les éléments constituant la ressource termino-ontologique, cette dernière étant la ressource utilisée par le système d'immersion. Chacun de ces différents niveaux de granularité est traité par différents processus : les entités nommées doivent être normalisées, la RTO doit être constituée, puis enrichie au fil des utilisations, et le système d'immersion doit être construit d'une part, et donner les moyens de son utilisation d'autre part.

Globalement, ces processus peuvent être rattachés à trois catégories distinctes, en fonction de leur nature :

- les processus endogènes font appel aux informations tirées du corpus lui-même pour traiter ce dernier ;
- les processus exogènes utilisent des ressources externes au corpus ;
- les processus anthropogènes font appel à l'utilisateur comme ressource pour traiter les données.

Ainsi, en combinant niveaux de granularité et types de processus, nous obtenons la grille formalisée en figure 5.1 page suivante.

Cette grille peut être lue de deux manières : en partant des unités et en allant vers les processus pour chacune d'entre elles, ou bien, à l'inverse, en partant des processus et en allant vers les unités. Puisque nous avons exposé dans cette partie les unités en jeu dans nos travaux, chaque type d'unité correspondant à un niveau de granularité, nous présentons brièvement pour chacune d'elles les processus qui leur sont appliqués, suivant le parcours schématisé sur la grille en figure 5.2 page ci-contre.

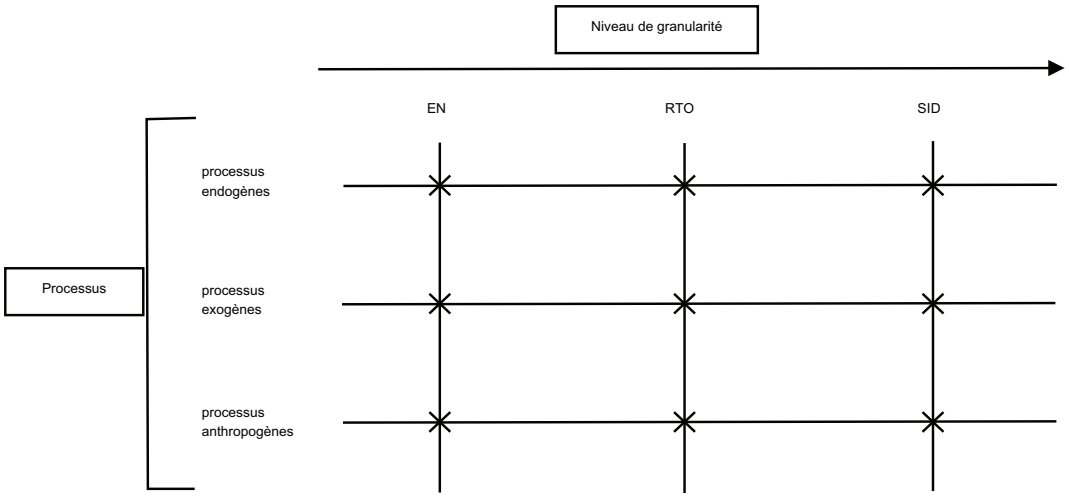

FIGURE 5.1 – Grille combinant niveaux de granularité et processus appliqués

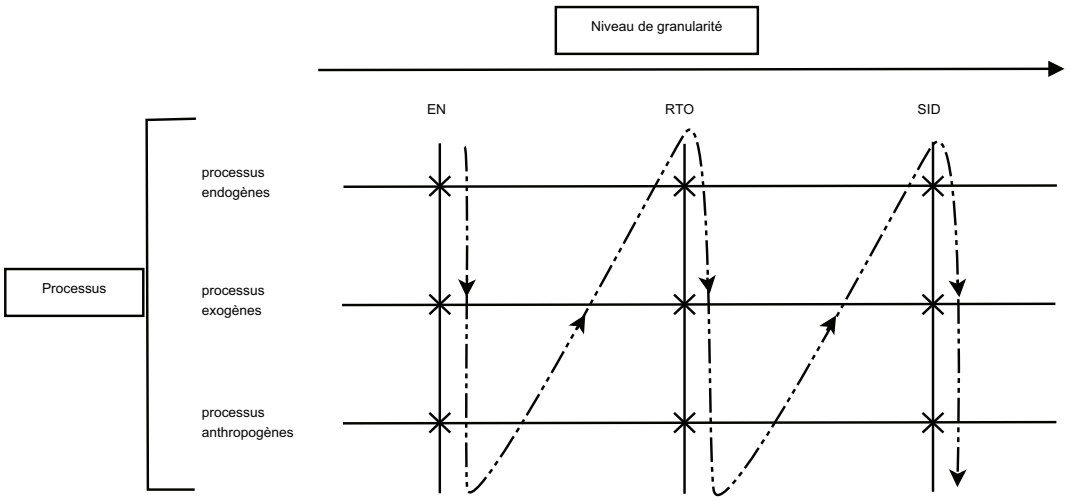

FIGURE 5.2 – Parcours de lecture par niveaux de granularité

5.1 Première lecture : approche granulaire

5.1.1 Les processus au niveau le plus fin : la normalisation des entités nommées et l'extraction des suites de mots graphiques

Le niveau de granularité le plus fin concerne deux types d'unités : les entités nommées d'une part, et les suites de mots graphiques d'autre part.

5.1.1.1 La normalisation des entités nommées

Pour les entités nommées, les traitements concernent en priorité, nous l'avons vu (voir chapitre 4 page 101), la normalisation de ces entités. Un premier repérage se fait de manière très simple, puisqu'elles sont pour la plupart stockées dans des champs distincts de la base de données. Ce repérage n'est donc pas en soi un problème à résoudre. En revanche, en raison de la disparité des formes pour une même entité nommée, en particulier pour les entités d'organisations et les entités de lieux, la normalisation est un processus à la fois crucial et complexe. Ainsi que nous l'avons définie en 4.2.2.1 page 128, la normalisation est un processus permettant de ramener plusieurs occurrences d'une même entité nommée à une forme standard, établie par une norme et appelée type, et dont la finalité est de reconnaître et de regrouper toutes les occurrences d'une même entité malgré les phénomènes de variation inhérents aux unités de langue. Les traitements que nous avons mis en place pour atteindre cet objectif relèvent de trois types d'approches : endogènes, exogènes et anthropogènes. A partir des bases de données en ligne, les documents sont tout d'abord importés dans les bases de données internes de la société TKM. C'est lors de cet import que les entités nommées sont grossièrement repérées, à partir des en-têtes des documents. Le processus de normalisation lui-même peut alors être appliqué.

Nous formalisons notre chaîne de traitement sur la figure 5.3 page suivante. Lors de l'import des documents, un processus d'extraction et réécriture est appliqué aux entités nommées. Sur ces noms pré-normalisés, deux traitements sont appliqués : les variantes sont détectées, et un découpage endogène est effectué sur les noms d'organisations à entités multiples. Les résultats de ces deux processus sont alors proposés à l'utilisateur pour correction. Ce dernier valide, corrige ou invalide les suggestions proposées et les noms pré-normalisés. Enfin, les noms normalisés sont intégrés à la base de données de TKM. Alors que le premier traitement, c'est-à-dire l'extraction et la réécriture, relève de méthodes exogènes, les suivants combinent méthodes endogènes et anthropogènes.

5.1.1.2 L'extraction des suites de mots graphiques

Les suites de mots graphiques sont extraites à partir des résumés des documents constituant l'ensemble documentaire. Cette extraction a lieu au moment de l'import, c'est-à-dire au même

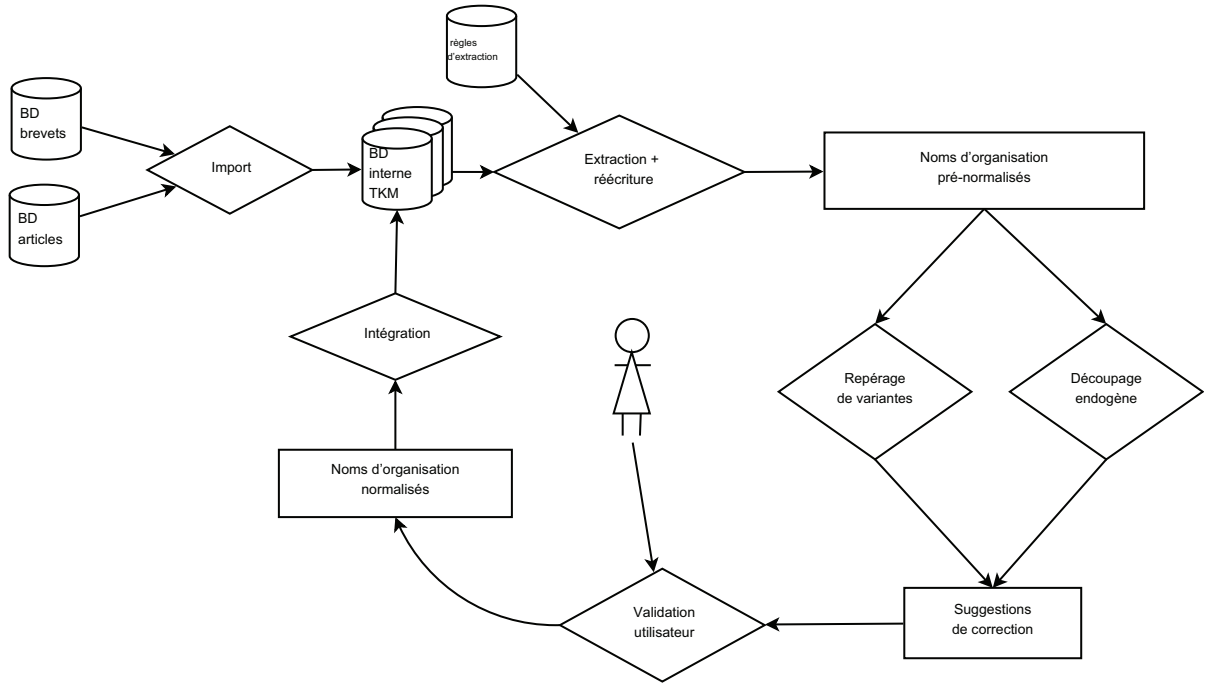

FIGURE 5.3 – Processus global de traitement des entités nommées

moment que la normalisation des entités nommées. Nous appelons *suites de mots graphiques* des mots « uniques » hors mots-outils, appelés mots simples, ou des suites de deux à cinq mots (voir en 4.1.2 page 106).

Nous schématisons la chaîne de traitement appliquée à ces suites graphiques sur la figure 5.4 page suivante. A partir de l'import, la langue du résumé est détectée. Une fois la langue connue, les suites de mots graphiques sont extraits directement des documents, avec l'aide d'une liste de mots interdits. Ces traitements relèvent donc de processus endogènes, puisque nous puisons l'information dans le corpus, et exogènes, avec les listes externes qui ont été construites.

L'ensemble de ces unités, c'est-à-dire les entités nommées et les suites de mots graphiques, sont ensuite intégrées à la ressource termino-ontologique multi-plans, qui vient se placer au niveau de granularité intermédiaire. Cette intégration s'effectue grâce à différents traitements, en fonction de la nature des informations et de l'utilisation qui en est prévue. Là encore, plusieurs types de processus sont en présence.

5.1.2 Les processus et méthodes au niveau intermédiaire : la modélisation de la ressource termino-ontologique

La ressource termino-ontologique (RTO) se place au niveau intermédiaire de granularité de nos unités. En effet, en tant que structure de représentation des connaissances, elle contient

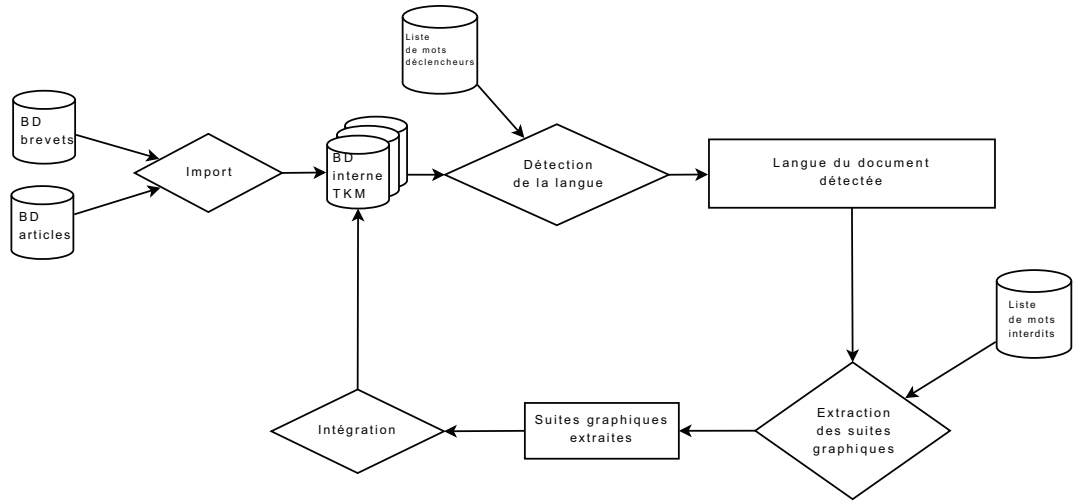

FIGURE 5.4 – Processus global d'extraction de termes

l'expression linguistique de ces connaissances, c'est-à-dire les entités nommées. En tant que ressource, elle est intégrée au système d'immersion permettant de naviguer efficacement au sein d'un ensemble documentaire.

La constitution et l'enrichissement de cette RTO passent eux aussi par un ensemble de processus relevant d'approches diverses. Nous les schématisons en figure 5.5 page ci-contre. Toutes les facettes sont constituées à partir des informations présentes dans les documents. Les facettes des lieux, des organisations et des thèmes sont construites par extraction endogène à laquelle s'ajoutent des traitements exogènes plus ou moins complexes. Les dates et les auteurs sont quant à elles remplies grâce à une extraction endogène.

Ainsi que nous l'avons mentionné en partie 3.2.2 page 84, notre RTO est une ressource multiplans, en ce sens que chaque type d'information est intégré à une facette distincte des autres. Distinguer ces informations permet en effet, outre de respecter une certaine cohérence dans le modèle, de les croiser plus simplement et plus efficacement pour répondre à des besoins des utilisateurs. Ainsi, chaque facette a une structure qui lui est propre, de façon à tirer parti au maximum des informations qu'elle contient. La méthode de constitution, et les processus utilisés pour ce faire, s'adaptent donc à chaque type d'information, et finalement, à chaque facette. C'est pourquoi une facette donnée ne sera pas forcément construite par les mêmes méthodes que les autres.

La facette des dates est construite à partir d'une extraction endogène, et a besoin de peu de formalisation. Les noms d'auteurs quant à eux sont extraits puis normalisés de manière relativement simple avant d'être intégrés à la facette correspondante. La facette des organisations, beaucoup plus complexe, est constituée grâce à la normalisation (évoquée en 5.1.1 page 148) des noms d'organisations, incluant une hiérarchisation de ces noms, et réalisée grâce à des proces-

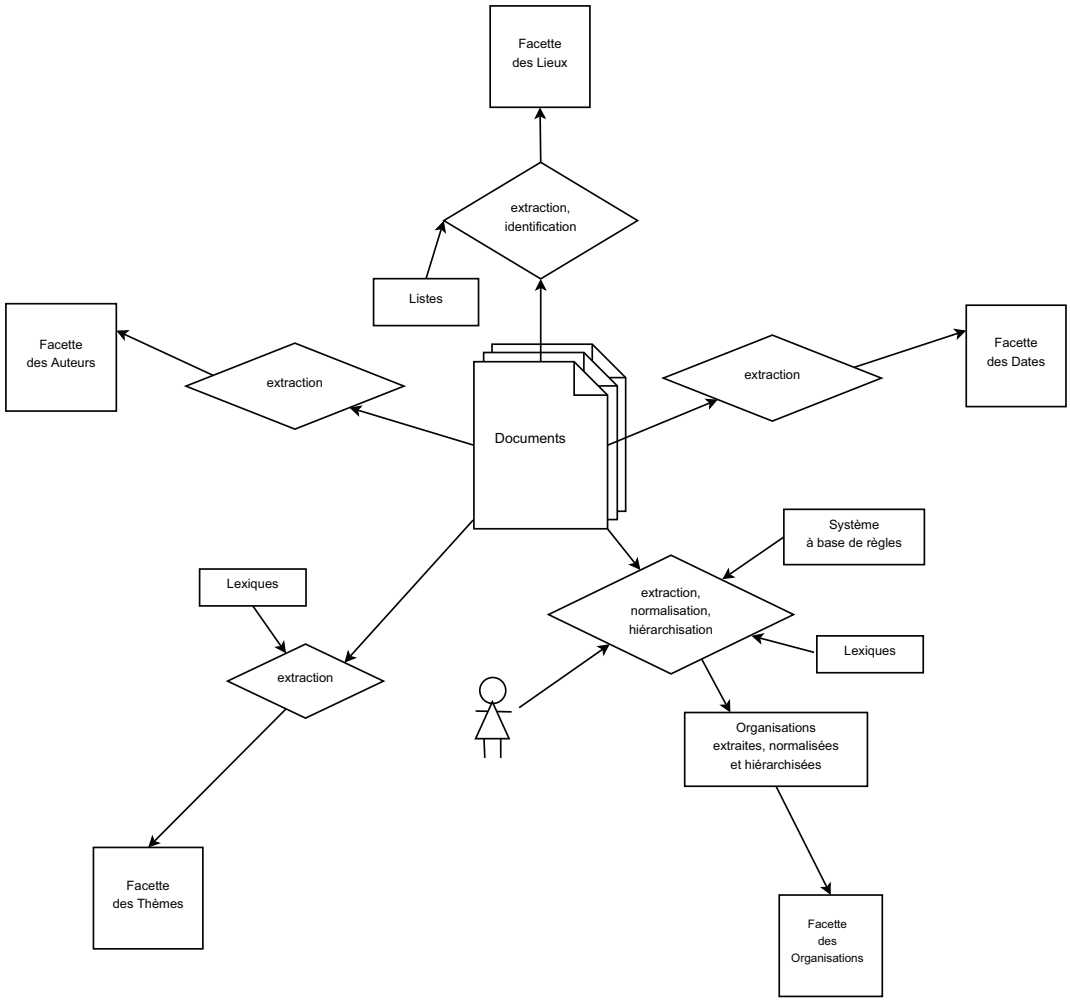

FIGURE 5.5 – Processus global de constitution de la RTO

sus anthropogènes et exogènes. La facette des lieux se construit à l'aide de méthodes endogènes et exogènes, fondées sur une identification entre éléments d'un lexique et contenu des données. Enfin, les suites de mots graphiques sont intégrées à la facette correspondante, là encore, par la combinaison de lexiques et du contenu des documents, par approches endogènes et exogènes.

5.1.3 Les processus au niveau global : l'immersion documentaire

Le système d'immersion permet à un utilisateur ayant un besoin d'information de la rechercher en fonction de critères définis par lui, et non imposés par la structure du système. Les processus permettant cette recherche relèvent, au fil des étapes de la consultation, d'approches endogènes, exogènes ou anthropogènes. Nous représentons le déroulement de la recherche dans le système d'immersion dans la figure 5.6 page suivante.

A partir d'un ensemble documentaire qu'il a lui-même constitué, l'utilisateur choisit, par

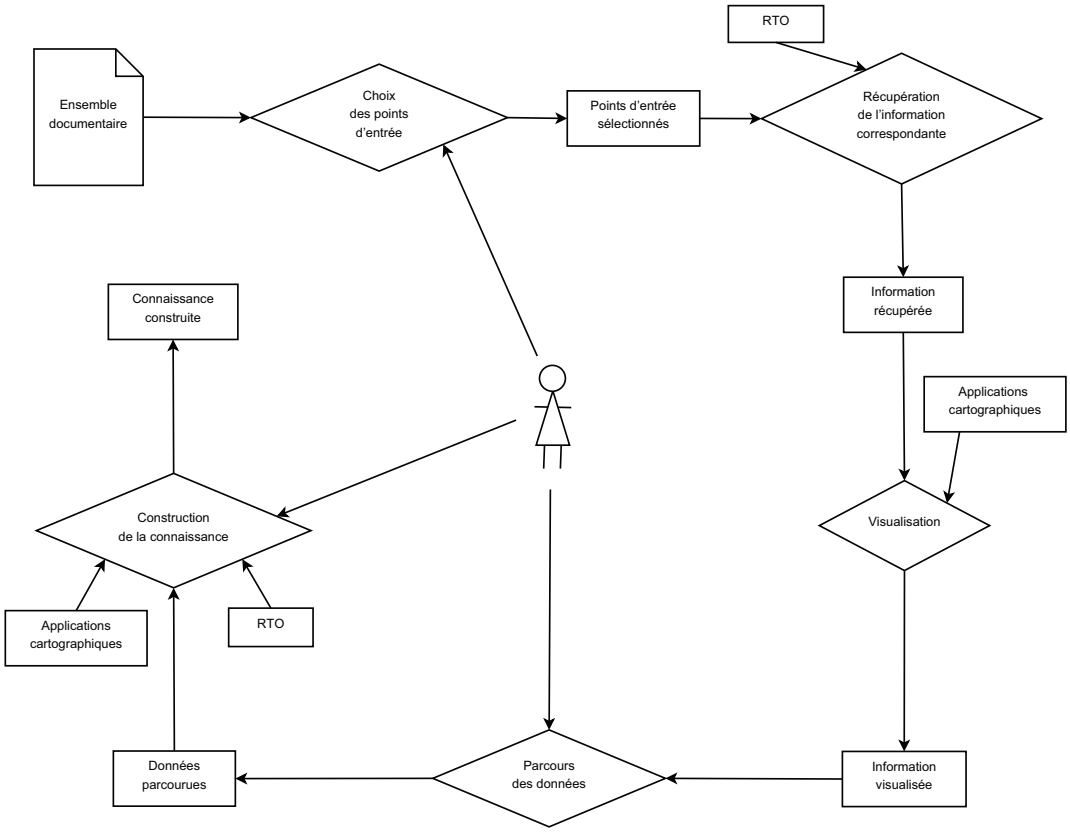

FIGURE 5.6 – Processus global de fonctionnement du système d'immersion

processus anthropogène, des points d'entrée en fonction de son but de recherche d'information. La requête construite à partir de ces points est appariée avec les informations contenues dans la RTO, ressource endogène, et permet de sélectionner les inscriptions numériques correspondantes. L'information ainsi récupérée est alors visualisée grâce à des applications cartographiques, exogènes par définition. A partir de cette visualisation, l'utilisateur peut parcourir les données, puis construire sa connaissance pertinente à partir de ce parcours. Ici, ce parcours et cette construction sont le fruit de la combinaison de ressources et processus des trois types. Le processus s'arrête lorsque l'utilisateur a construit la connaissance nécessaire à son activité, et à l'exécution de la tâche principale qui lui a été confiée.

Cette distinction entre processus se fonde avant tout sur les ressources mises en jeu pour leur exécution. Selon que ces ressources sont inhérentes au corpus, externes à ce corpus, ou encore fondées sur l'utilisateur en tant qu'agent connaissant ou interprétant, les processus fonctionnent différemment.

5.2 Deuxième lecture : approche par types de processus

Dans ce chapitre, nous avons utilisé une grille situant les unités impliquées dans nos travaux et les processus que nous avons mis en place. Nous avons présenté en section 5.1 page 148 ces unités classées en fonction de leur niveau de granularité, et pour chacune de ces catégories d'unités, les processus utilisés. Or, ce n'est qu'une des deux lectures possibles de notre grille. En effet, nous avons jusqu'à présent choisi les unités traitées, c'est-à-dire les entités nommées (EN), la ressource termino-ontologique (RTO) et le système global d'immersion documentaire (SID), comme premier niveau de segmentation. Mais il est également possible de proposer un autre parcours de lecture : le parcours fondé au premier chef sur les types de processus utilisés. Reprenant notre grille de départ, nous obtenons alors la figure 5.7.

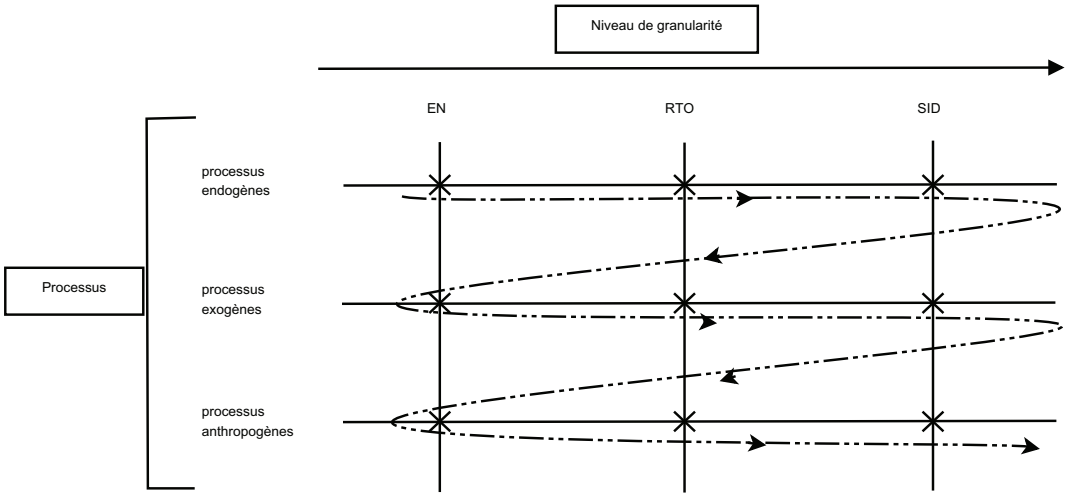

FIGURE 5.7 – Parcours de lecture par les processus appliqués

Cette distinction entre processus se fonde avant tout sur les ressources mises en jeu pour leur exécution. Selon que ces ressources sont internes au corpus, externes à ce corpus, ou encore fondées sur l'utilisateur en tant qu'agent connaissant ou interprétant, les processus fonctionnent différemment.

5.2.1 Les processus endogènes

Les processus endogènes font appel aux données contenues dans le corpus ou l'ensemble documentaire pour traiter ce dernier. Tout traitement faisant appel à ces données que nous pourrions qualifier d'internes relève donc, peu ou prou, d'un processus endogène. Ainsi, la détection de variantes pour la normalisation des entités nommées par exemple (voir figure 5.3 page 149) est un processus endogène, les informations du corpus permettant de détecter les entrées qu'il contient en double *modulo* certaines variations.

5.2.2 Les processus exogènes

Les processus exogènes utilisent des ressources externes au corpus ou à l'ensemble documentaire, comme des lexiques, des dictionnaires, etc. Reprenant notre exemple de la normalisation, ainsi que la figure 5.3 page 149, l'extraction et la réécriture des entités nommées se fait à l'aide d'un système de règles, fondé par ailleurs sur divers lexiques indépendants de l'ensemble documentaire.

5.2.3 Les processus anthropogènes

Les processus anthropogènes font appel à l'utilisateur comme ressource pour traiter les données. Considérant ce dernier comme agent interprétant et concepteur de l'information pertinente, nous lui octroyons une place à part entière dans notre système, et ce aux différents niveaux de granularité que nous avons définis. Pour la normalisation, l'utilisateur intervient à deux reprises dans le processus, pour valider ou corriger les suggestions du système.

C'est ce parcours que nous avons choisi pour structurer la suite de la présentation de nos travaux. De cette manière, nous mettons en avant la complémentarité des méthodes pour l'optimisation de l'immersion documentaire, et ce pour chaque niveau de granularité. Dans le chapitre 6, nous abordons donc les processus endogènes appliqués aux trois niveaux de granularité en présence. Le chapitre 7 est consacrée aux processus exogènes. Enfin, les processus anthropogènes sont détaillés dans le chapitre 8.

Cependant, si le lecteur souhaite lire cette grille avec pour point de départ les unités traitées en fonction de leur niveau de granularité, il peut reprendre, à partir du plan, les sections correspondantes, et réaliser une lecture rhizomatique par approches granulaires de ce travail (6.1, 7.1 et 8.1 pages 159, 241 et 320 pour le grain des entités nommées, 6.2, 7.2 et 8.2 pages 199, 284 et 343 pour le grain de la RTO, et 6.3, 7.3 et 8.3 pages 222, 294 et 349 pour le système d'immersion documentaire).

Deuxième partie

Les processus et ressources en jeu dans l'immersion documentaire : traitements endogènes, exogènes et anthropogènes

Chapitre 6

Les processus et ressources endogènes : traiter le corpus par les données intrinsèques au corpus

Sommaire

L'adjectif *endogène* désigne de manière générale "tout ce qui vient de l'intérieur, ce qui a son origine au-dedans de l'objet, de l'organisme, du système ou de l'ensemble étudié" [44]. Il s'oppose à l'adjectif *exogène*.

De manière générale, les documents textuels de nos ensembles documentaires sont riches d'informations, de plusieurs points de vue. Ils véhiculent en effet, nous l'avons expliqué dans la première partie de ce travail (voir chapitre 3 page 69), des informations relatives à leur contexte de production, que nous avons qualifiées d'informations contextuelles. D'autre part, en tant que signes, ces documents fournissent des connaissances, par le biais de leur contenu. C'est ce que nous avons nommé des informations thématiques. Enfin, ils donnent accès, à un niveau plus fin, à des structures textuelles caractéristiques à la fois de leur genre et des informations exprimées, contextuelles ou thématiques.

Pour traiter ces documents, à plusieurs niveaux de granularité, il est possible de mettre à profit ces informations intrinsèques. Pour cela, l'ensemble documentaire lui-même est exploité au sein de processus endogènes. Ceux-ci permettent de tirer de l'information nécessaire au traitement des documents dans ces mêmes documents.

Afin de poser les fondations de ces ressources, nous définissons d'ores et déjà les processus endogènes de la manière suivante :

Définition 4. *Dans le cadre de nos travaux, un processus endogène utilise les données de l'ensemble documentaire lui-même pour traiter ce même ensemble documentaire.*

Dans le domaine du traitement automatique des langues (TAL), des méthodes endogènes sont utilisées pour l'analyse syntaxique de textes et/ou l'extraction terminologique. [Vergne, 2004], par exemple, expose une méthode d'analyse grammaticale partielle sans ressources de corpus multilingues. [Bourigault, 1993] a conçu le logiciel LEXTER, un logiciel d'extraction terminologique fondé sur une analyse syntaxique robuste détectant des syntagmes nominaux. Plus tard, il crée Syntex, analyseur détectant également les syntagmes verbaux [Bourigault et al., 2005] dans le même objectif. [Vergne, 2003] s'intéresse lui aussi à l'extraction terminologique, dans un contexte multilingue.

Tous ces travaux ont en commun l'utilisation exclusive d'informations intrinsèques aux corpus pour aboutir à leur objectif.

Ces méthodes mettent en avant les avantages à tirer de ressources endogènes. En particulier, elles économisent l'utilisation, voire la constitution de ressources externes à intégrer aux processus. D'autre part, et c'est là l'une de leurs principales forces, les systèmes endogènes, quel que soit leur objet, sont adaptés par essence à la nature des données qu'ils traitent. Un traitement endogène peut donc fonctionner, en principe, sur des textes dont la langue [Vergne, 2004] et/ou le domaine [Frérot et al., 2003] ne sont pas connus à l'avance.

44. Définition tirée du Lexique des termes de la complexité : `http://www.intelligence-complexite.org/fr/documents/lexique-de-termes-de-la-complexite.html`. Page consultée le 25/01/2011.

De telles méthodes et ressources sont adaptées à nos besoins, en particulier du fait de la diversité des domaines abordés, et de l'impossibilité de constituer des ressources externes pour traiter la plupart des types de données qui nous intéressent. Par ailleurs, leur application concerne l'ensemble des niveaux de granularité en présence.

Dans un premier temps, nous décrivons les méthodes et ressources endogènes qui sont employées au niveau de granularité le plus fin, c'est-à-dire pour la normalisation des entités nommées. Puis nous présentons la manière dont ces méthodes influent sur la conception et la construction de la structure de représentation des informations des documents, nommément la ressource termino-ontologique multi-plans. Enfin, nous détaillons le processus d'immersion documentaire par endogénéité pour l'accès à l'information et la construction de connaissances par l'humain.

6.1 Les méthodes endogènes pour le traitement des unités d'information contextuelle : traiter les entités nommées par les entités nommées

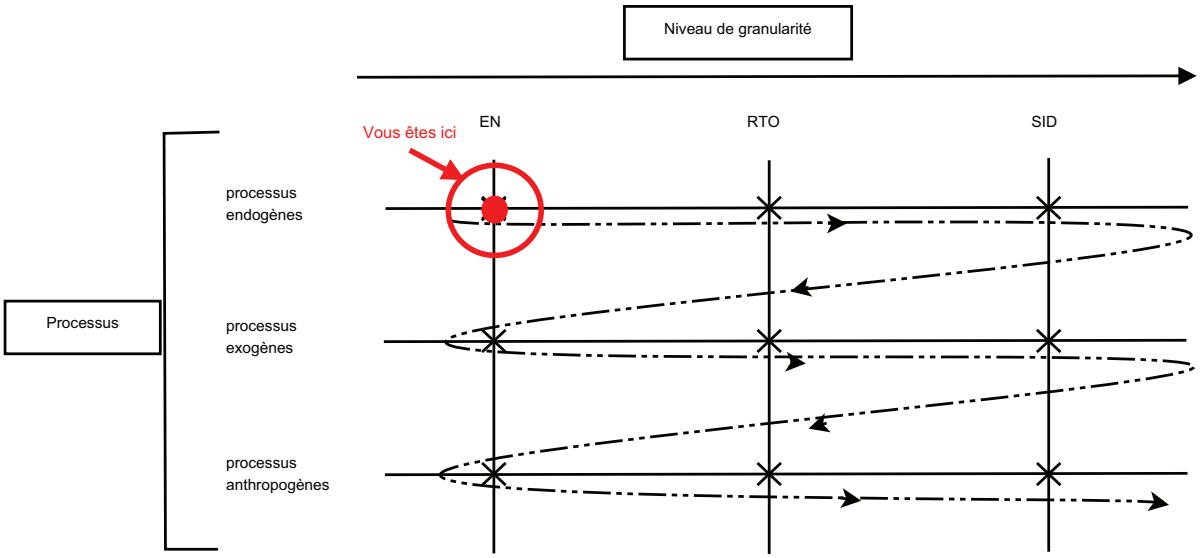

FIGURE 6.1 – Positionnement des traitements endogènes pour la normalisation dans l'ensemble des processus

Les approches endogènes pour la normalisation se placent dans un processus plus global de traitement des entités nommées. Le positionnement de ces traitements est rappelé dans la figure 6.1. Les entités nommées représentent une grande partie des unités d'informations qui permettent de caractériser et structurer les documents d'un ensemble documentaire. Or, dans leur forme brute, tirée des en-têtes des documents, elles sont inexploitables pour cette tâche de

caractérisation et de structuration.

La première étape intégrée à notre système d'immersion est donc celle de la normalisation de ces entités nommées. Pour cela, des processus et des ressources de différents types sont utilisés. Les ressources endogènes, et les processus reposant sur celles-ci, sont exploités en combinaison avec des ressources exogènes et anthropogènes. Cette combinaison est un moyen de tirer parti des avantages de chacun, et de pallier leurs limites respectives.

Nous avons exposé dans la section 4.2 page 114 une vue d'ensemble des problèmes posés par les entités nommées dans notre corpus. La plupart de ces difficultés sont présentes dans les entités d'organisations, et c'est pourquoi nous nous concentrons particulièrement, dans cette section, sur ces unités.

Pour les besoins spécifiques à la normalisation des entités nommées, nous postulons que les traitements endogènes permettent de dégager, à partir du corpus d'entités à normaliser, les informations pertinentes pour effectuer cette normalisation.

Dans la section 4.2.2 page 127 de ce document, nous avons recensé les différents types d'erreurs et de difficultés posées par les entités nommées dans leur forme brute, et en particulier les entités d'organisations. Nous reprenons et enrichissons dans le tableau 6.1 la table 4.4 page 128 présentée alors.

Forme souhaitée	Variante 1	Variante 2
Univ Toulouse 2 Le Mirail	Université de Toulouse 2	Univ Mirail
Alcatel	Alcatel SA	Acatel SA
Toshiba	Toshiba	Toshiba Corp.
Univ Kentucky	Univ Kentucky	Univ Kentucky Medical Center

TABLE 6.1 – Exemples d'expressions synonymes ou partiellement synonymes issues des métadonnées des documents

Nous avons qualifié ce phénomène de *synonymie des entités nommées*, puisque deux ou plusieurs variantes différentes peuvent désigner un même référent. D'un point de vue conceptuel, elles peuvent être considérées comme différentes instanciations d'un même concept.

Tous ces exemples illustrent le phénomène de variation, très présent dans nos données, à l'origine de cette synonymie de fait :

- Le premier type de synonymie, représenté par le premier exemple, relève d'une différence de désignation, avec deux dénominations concurrentes pour un référent identique du monde réel. Nous parlons dans ce cas de variation sémantique, puisqu'elle découle de désignations différentes malgré une équivalence de sens.
- Le deuxième type relève d'une variation orthographique puisqu'en l'occurrence, une faute de frappe empêche l'assimilation des deux noms bruts entre eux.

– Le troisième type de synonymie relève de conventions de notation différentes, puisque la deuxième variante contient un identifiant d'organisation, *Corp.*, non présent dans la première.
– Enfin, le dernière type de variation menant à la synonymie a une forme particulière : la seconde variante a un niveau de détail que n'a pas la première, puisque celle-là inclut le nom d'un centre que celle-ci ne contient pas. Il s'agit donc de variantes partiellement synonymes. Ici, non seulement la synonymie doit être résolue, mais une structure hiérarchique doit être dégagée de la seconde variante, plus riche du point de vue informationnel.

Les ressources et processus endogènes de normalisation peuvent être utilisés pour la résolution de la synonymie typo-orthographique et de la synonymie partielle. Les autres cas sont traités par d'autres types de traitements.

Pour la normalisation des entités nommées, les ressources endogènes se placent dans un contexte multi-ressources. Elles sont employées pour résoudre un certain nombre de difficultés pour lesquelles les autres types de ressources sont limités, ou engendrent des coûts trop importants. Dans cette section, nous présentons d'abord les apports d'un calcul de distance d'édition pour la normalisation des entités d'organisations. Puis nous décrivons les traitements de segmentation des noms d'organisations à entités multiples qui ont été mis en place. Enfin, nous synthétisons en présentant la manière dont l'ensemble de ces traitements endogènes permettent de dégager de l'information pertinente pour la normalisation des entités nommées.

6.1.1 Normaliser par la distance d'édition de Levenshtein

Dans le tableau ci-dessus (table 6.1 page précédente), le cas illustré par les variantes *Alcatel SA* et *Acatel SA* relève d'un « accident » lors de la saisie des données et non d'un parti pris. Cependant, il s'inscrit bien dans le phénomène de synonymie, puisqu'il crée cette dernière de fait entre deux ou plusieurs entités nommées. Les deux noms d'organisations *Alcatel SA* et *Acatel SA*, attestés dans notre corpus, ne seront pas reconnus comme une seule et même organisation, et ce en raison de la faute de frappe sur le second nom, même si en réalité, ils renvoient tous les deux à la même organisation *Alcatel SA*, entreprise implantée dans le domaine des technologies de la communication.

Ce phénomène pose de nombreux problèmes pour le traitement et l'interprétation des données par les analystes de TKM, non seulement pour les calculs statistiques visant à déterminer les organisations les plus actives pour un domaine donné, qui passent par des comptages sur les noms d'organisations, mais également pour les analyses qualitatives. Dans ce dernier cas, l'utilisateur peut passer à côté d'un document interprétativement riche simplement parce que l'organisation propriétaire de celui-ci ne porte pas le nom qu'il recherche.

Par exemple, dans un corpus de 3 653 organisations présentes dans un ensemble documentaire

relevant du domaine de l'optique, le nom *Chinese Acad Sciences* (*Chinese Academy of Sciences*) compte 68 occurrences, le nom *Univ Tsinghua* 27, et le nom *OTI Ophthalmic Technologies* 9. Cependant, il existe également d'autres variantes pour chacun de ces noms, que nous rassemblons respectivement dans les tableaux 6.2 à 6.4.

Variante	Fréquence	Pourcentage
Chinese Acad Sciences	68	64,8%
Acad Chinese Sciences	10	9,5%
Chinese Acad Science	22	21%
Chinese Acad Sciences CAS	1	0,9%
Chinese Acad Sci	4	3,8%
TOTAL	105	100%

TABLE 6.2 – Variantes du nom d'organisation *Chinese Acad Sciences*

Variante	Fréquence	Pourcentage
Univ Tsinghua	27	79,4%
Univ Tsinkhua	6	17,7%
Univ Tszinkhua	1	2,9%
TOTAL	34	100%

TABLE 6.3 – Variantes du nom d'organisation *Univ Tsinghua*

Variante	Fréquence	Pourcentage
OTI O**phth**almic Technologies	9	81,8%
OTI O**pth**almic Technologies	1	9,1%
OTI O**pht**almic Technologies	1	9,1%
TOTAL	11	100%

TABLE 6.4 – Variantes du nom d'organisation *OTI Ophthalmic Technologies*

Plusieurs types de variations apparaissent ici. Tout d'abord, des variations fondées sur l'inversion de deux mots dans un nom relèvent plutôt de la syntaxe. C'est le cas de *Acad Chinese Science* et de *Chinese Acad Sciences*[45]. Cependant, les autres variantes de ce nom sont dues à la présence d'abréviations, ou à l'alternance entre un pluriel et un singulier. Les variantes de

45. Ces inversions au niveau syntaxique sont partiellement résolues par d'autres types de traitements. Voir en particulier la section 7.1 page 241

Univ Tsinghua relèvent quant à elles probablement de variations dans la traduction et surtout la transcription d'un nom chinois. Enfin, les variantes du dernier exemple relèvent clairement de l'erreur orthographique autour du mot *ophthalmic*.

Pour le nom *Chinese Acad Sciences*, il est important de rassembler les variantes afin d'en tirer des statistiques plus justes. En effet, au sein de notre corpus de test, le cumul de toutes ces variantes font de la Chinese Academy of Sciences l'organisation ayant publié le plus de documents. Mais ce rang ne peut lui être attribué si les variantes ne sont pas détectées, et dans ce cas, d'autres organisations prennent sa place dans le classement.

Quant à l'université Tsinghua, bien qu'elle ne se place pas en tête, elle est en concurrence, du point de vue du nombre de publications, avec d'autres organisations ayant publié sensiblement le même nombre de documents. De fait, pouvoir attribuer les 34 documents à cette organisation, au lieu des 27 de départ, est un gage de fiabilité des statistiques tirées des données.

Enfin, les variantes de *OTI Ophtalmic Technologies* sont beaucoup moins nombreuses. Cependant, il peut être intéressant de les détecter malgré tout, d'un point de vue statistique ou qualitatif. Si l'organisation en question présente un intérêt particulier pour l'analyste, celui-ci doit pouvoir identifier toute sa production, qu'elle soit massive ou plus limitée, afin de l'examiner en détail s'il le souhaite.

Pour toutes ces raisons, les fautes de saisie, de frappe, d'orthographe, les variantes de traduction et certaines variations syntaxiques localisées doivent être détectées pour être corrigées. Dans les bases de données, les problèmes engendrés par ces erreurs peuvent être résolus grâce à des mesures de similarité. En fonction des besoins, différents types de mesures de similarité peuvent être appliqués.

6.1.1.1 Un tour d'horizon des mesures de similarité

Les solutions classiquement utilisées pour résoudre de tels cas sont les mesures de similarité. Elles sont particulièrement utilisées dans le cadre de la détection et de la correction de doublons d'entrées ou de champs dans les bases de données. [Elmagarmid et al., 2007] ont réalisé une revue très complète des techniques existantes dans le domaine. Nous nous appuyons sur cette étude et les différentes publications auxquelles elle fait référence pour développer cette section et nous l'approprier. Toutes les techniques partent du principe que pour parvenir à dédoublonner les entrées d'une base de données, il faut avant tout détecter les variantes textuelles existant pour un même élément contenu dans différentes entrées.

Ces auteurs distinguent les techniques visant à établir des correspondances entre champs individuels de bases de données, par exemple le champ *nom d'entreprise* d'une table de base de données contenant les contacts des clients d'une entreprise, des techniques cherchant à détecter plusieurs exemplaires d'une même entrée dans sa totalité, par exemple l'intégralité des informa-

tions sur un client dans cette même table. Dans notre cas, l'existence de doublons sur la totalité d'une entrée reste extrêmement marginale, puisque les documents déjà présents dans la base sont identifiés au moment de l'import des nouveaux documents entrants. C'est pourquoi nous ne nous intéresserons qu'aux techniques de détection de doublons dans des champs individuels.

Toutes les méthodes relevées par [Elmagarmid et al., 2007] utilisent des mesures de similarité. Ces dernières consistent à comparer deux ou plusieurs chaînes de manière à savoir si elles sont ou non similaires, et donc, à terme, pour déterminer s'il s'agit ou non de doublons. Les auteurs recensent quatre types de mesures de similarité, en fonction des unités faisant l'objet de ces mesures. Les trois premiers types s'intéressent aux chaînes de caractères alpha-numériques, tandis que le dernier est consacré aux chaînes exclusivement numériques :

1. les mesures fondées sur les caractères ;
2. les mesures fondées sur les tokens ;
3. les mesures phonétiques ;
4. les mesures de similarité de chaînes numériques.

Les mesures fondées sur les caractères sont les plus appropriées pour régler les problèmes engendrés par les erreurs typographiques. Les distances d'édition permettent de calculer une « distance » entre deux chaînes de caractères, distance fondée sur le nombre minimum d'opérations entre caractères pour passer de la première à la seconde chaîne. La distance d'édition la plus connue est celle calculée par l'algorithme de Levenshtein [Levenshtein, 1966], qui consiste à calculer le nombre d'insertions, d'effacements et de substitutions de caractères pour transformer une chaîne en une autre. Ainsi, une distance de 0 représente la correspondance parfaite entre les deux chaînes, puisqu'aucune opération n'est nécessaire. Les deux chaînes sont donc, dans ce cas, identiques. De fait, plus une distance est faible, plus les chaînes sont similaires. Pour notre exemple précédent, la distance d'édition entre *Alcatel SA* et *Acatel SA* est égale à 1, puisqu'une opération d'effacement est nécessaire pour aller de l'une à l'autre. Nous présentons la matrice correspondant au calcul dans la figure 6.2 page suivante.

D'autres distances d'édition sont des extensions de la distance de Levenshtein. Par exemple, la « affine gap distance » de [Waterman et al., 1976] ajoute deux opérations à celles qui sont utilisées dans l'algorithme de Levenshtein : *open gap* et *extend gap*. Ces opérations permettent de calculer un écart moins grand dans le cas d'abréviations, pour le prénom d'un nom de personne par exemple, où la distance de Levenshtein ne fonctionnerait pas. Ainsi, cette distance permet d'établir une similarité entre les chaînes *John Richard Smith* et *John R. Smith*.

La distance de Smith-Waterman [Smith & Waterman, 1981] étend les deux premières distances et pondère à la baisse les opérations de transformation nécessaires au début et à la fin des chaînes de caractères. De cette manière, *John R. Smith, Prof.* et *Prof. John R. Smith, University of Calgary* peuvent être rapprochées par une faible distance, puisque le « suffixe » *University of*

Alcatel SA | Acatel SA

		A	l	c	a	t	e	l		S	A
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
A	**1**	0	1	2	3	4	5	6	7	8	9
c	**2**	1	1	1	2	3	4	5	6	7	8
a	**3**	2	2	2	1	2	3	4	5	6	7
t	**4**	3	3	3	2	1	2	3	4	5	6
e	**5**	4	4	4	3	2	1	2	3	4	5
l	**6**	5	4	5	4	3	2	1	2	3	4
	7	6	5	5	5	4	3	2	1	2	3
S	**8**	7	6	6	6	5	4	3	2	1	2
A	**9**	8	7	7	7	6	5	4	3	2	**1**

FIGURE 6.2 – Matrice de calcul de distance de Levenshtein pour le couple *Alcatel SA* vs. *Acatel SA*

Calgary sera ignoré.

La distance de Jaro [Jaro, 1976] permet quant à elle, grâce à une comparaison des caractères communs aux deux chaînes comparées, de détecter des similarités particulièrement entre noms de famille ou entre prénoms. Enfin, [Ukkonen, 1992] propose d'utiliser les q-grammes pour comparer deux chaînes. Les q-grammes sont obtenus grâce à une fenêtre de longueur q, qui glisse sur les caractères de chaque chaîne. L'auteur part en effet du principe que deux chaînes sont similaires si elles partagent un grand nombre de q-grammes [46].

Les mesures fondées sur des tokens permettent de dépasser des difficultés qu'il n'est pas possible de résoudre avec les mesures sur les caractères. Les conventions de notation variant d'une source à l'autre, l'ordre des éléments saisis peut changer. Par exemple, pour les noms de personnes, une source peut noter *John Smith* et une autre *Smith, John*. L'algorithme le plus simple est celui des chaînes atomiques [Monge & Elkan, 1996], où une chaîne atomique est une chaîne de caractère entre deux signes de ponctuation. Deux chaînes atomiques sont considérées comme correspondantes si elles sont égales, ou encore si l'une est le « préfixe », c'est-à-dire le début de chaîne, de l'autre. Ainsi, la similarité de deux chaînes est mesurée sur leur nombre de chaînes communes, divisé par leur nombre moyen de chaînes atomiques.

Le système WHIRL [Cohen, 1998] combine la similarité cosine et la pondération par tf.idf

46. Le q-gramme correspond à la notion de *n-gramme* utilisée dans d'autres disciplines, et en particulier en traitement automatique des langues. Cependant, nous employons ici le terme de q-gramme, puisque c'est celui qui est utilisé par l'auteur de la méthode.

pour calculer la similarité entre deux champs d'une base de données. Parce qu'il prend en compte l'apparition des mots dans un champ et non leur positionnement, et leur fréquence relative, ce système fonctionne dans un grand nombre de cas.

Enfin, [Gravano et al., 2003] combinent le système WHIRL aux q-grammes. Il est alors possible de gérer également les erreurs de typographie, en utilisant les q-grammes plutôt que les mots comme tokens.

Les mesures de similarité phonétique permettent de rapprocher des chaînes graphiquement différentes, mais phonétiquement proches. Des algorithmes comme Soundex [Russell, 1918] ou Metaphone [Philips, 1990] peuvent être utilisés pour rapprocher des chaînes contenant, par exemple, le mot anglais *Corporation* et sa « traduction » approximative en japonais *Koporeysion* [47], que nous pouvons représenter en phonétique en [koɹpəɹeɪʃən] et [kopəɹeɪʃən], respectivement. Ces deux algorithmes convertissent des lettres différentes pouvant être prononcées de la même manière en suites de chiffres identiques.

Enfin, les mesures de similarité appliquées aux chaînes de chiffres, et non plus de lettres, sont encore peu développées. En général, sont appliquées aux chiffres les mêmes mesures que pour les lettres. Dans ce cas, les suites de chiffres sont considérées comme des chaînes de caractères classiques.

Avec pour objectif une application différente, [Lebarbé & Breese, 2001] utilisent des mesures de similarité pour comparer des noms de marques pour détecter la contrefaçon. Le système CATMInE est fondé sur une distance de Levenshtein optimisée, qui détecte, dans des positions différentes, des sous-chaînes communes aux deux noms.

Nous récapitulons dans le tableau 6.5 page ci-contre les différentes méthodes que nous avons évoquées. Chacun de ces types de mesure présente un intérêt, hormis, dans notre cas, les mesures appliquées aux mesures numériques. Au vu des différents types d'erreurs rencontrées dans nos données, nous aurions pu appliquer les unes après les autres les mesures de similarité correspondantes. Cependant, nous avons fait le choix d'utiliser uniquement une distance d'édition fondée sur les caractères. En effet, les autres mesures de similarité permettent de traiter des problèmes que nous avons nous-même traités d'une autre manière, plus cohérente avec nos objectifs, et moins coûteuse en temps de traitement.

6.1.1.2 Hypothèse : une distance de Levenshtein adaptée

En pratique, nous avons besoin de repérer des variantes d'un même nom d'organisation, dues aux erreurs que nous venons de citer, afin de pouvoir les rassembler par la suite sous un seul et même nom. L'objectif est donc bien d'établir des correspondances entre chaînes de caractères similaires. Reprenant notre exemple de *Alcatel SA* vs. *Acatel SA*, la normalisation doit pouvoir

47. Ces exemples sont attestés dans notre corpus.

	Auteur	Système / Méthode	Unité utilisée	Approche	Types de problèmes résolus
1	[Levenshtein, 1966]	Distance d'édition	caractère	Mesure des quatre opérations de modification de caractère d'une chaîne à l'autre	Fautes de frappe, erreurs typo-orthographiques
2	[Waterman *et al.*, 1976]	Affine gap distance	caractère	A partir de la distance de Levenshtein, coût moindre attribué à l'opération d'effacement	Abréviations, dans les noms de personne par exemple: *J.R. Smith* vs. *John Ronald Smith*
3	[Smith et Waterman, 1981]	Distance de Smith-Waterman	caractère	A partir de 1 et 2, coût moindre pour les différences en début ou en fin de chaîne.	Inversion de l'ordre de certaines chaînes, par exemple: *John R. Smith*, *Prof.* vs. *Prof. John R. Smith*
4	[Jaro, 1976]	Distance de Jaro	caractère	Comparaison des caractères communs de deux chaînes puis calcul des déplacements de ces caractères communs	Comparaison des nom et prénoms de personnes
5	[Ukkonen, 1992]	q-grammes	fenêtre glissante de caractères	Comparaison des q-grammes	Chaînes qui diffèrent mais ont des séquences en commun
6	[Monge et Elkan, 1996]	Chaînes atomiques	Token (caractères entre deux signes de ponctuation)	Deux chaînes atomiques C1 et C2 sont correspondantes si elles sont égales ou si C1 est le préfixe de C2	Conventions de notation différentes: *John Smith* vs. *Smith, John*
7	[Cohen, 1998]	WHIRL	token	Similarité cosine + tf.idf. C'est la présence et la fréquence d'un token qui compte, et non sa position.	Grand nombre de cas traités, mais pas les erreurs typo-orthographiques
8	[Gravano, 2003]	WHIRL + q-grammes	q-grammes	Même approche qu'en 7, mais sur les q-grammes et plus les tokens	Mêmes cas qu'en 7 + erreurs typo-orthographiques
9	[Russell, 1918]; [Phillips, 1990]	Soundex; Metaphone	phonèmes	Attribution de codes chiffrés identiques à des caractères ou groupes de caractères différents, mais pouvant avoir la même prononciation	Transcriptions approximatives, comme *Corporation* vs. *Koporeysion*
10	[Lebarbé et Breese, 2001]	CATMInE	caractères	Distance de Levenshtein optimisée: détecte des sous-chaînes communes	Contre-façons de noms de marques

TABLE 6.5 – Récapitulatif des mesures de similarité

établir, *a minima*, que ces deux chaînes sont similaires.

Par ailleurs, notre système doit traiter un nombre important de données. Par exemple, 20 000 noms d'organisations liés à une étude particulière doivent pouvoir faire l'objet d'une normalisation relativement rapidement.

Pour résumer, le système de normalisation des entités nommées doit donc permettre de rapprocher des paires de noms d'organisations similaires, et ce en un laps de temps qui soit acceptable dans un contexte industriel.

Nous avons choisi d'utiliser la distance de Levenshtein pour la détection des erreurs de typographie et des fautes d'orthographe, mais nous l'avons modifiée de manière à ce qu'elle soit mieux adaptée à nos besoins.

Nous avons vu que le principe de la distance de Levenshtein est le calcul du nombre de modifications nécessaires pour passer d'une chaîne A à une chaîne B. Ce calcul s'effectue sur chacun des caractères, et sont considérés comme modifications les insertions d'un caractère, les effacements et les substitutions. La distance est donc la somme de l'ensemble des modifications minimales opérées sur la chaîne A pour arriver à la chaîne B. Dans la version originale de l'algorithme, chaque opération de modification vaut 1.

Nous posons donc l'hypothèse suivante :

Hypothèse II.1. *La distance de Levenshtein, dans son principe, est pertinente pour détecter des variantes typographiques dans les noms d'organisations de nos données, mais elle doit en revanche être adaptée à l'objet (nom d'organisation) et à l'objectif de la tâche (normalisation).*

Sous-hypothèse 1. *Pour rendre la distance de Levenshtein plus pertinente et ce quelle que soit la longueur des chaînes testées, il convient de la relativiser, en mettant en rapport le résultat absolu du calcul avec la longueur de ces chaînes de caractères.*

Sous-hypothèse 2. *Pour pouvoir intégrer un tel calcul sur des volumes potentiellement importants de données, il est nécessaire de trier avant l'exécution du calcul de distance les données qui y seront soumises, et d'éliminer les couples qui peuvent d'ores et déjà être rejetés en tant que variantes d'un même nom d'organisation.*

Sous-hypothèse 3. *Certains mots graphiques très récurrents, tels que Univ, Lab ou Institute, ne désignent pas l'organisation mais la qualifient. Ils biaisent le calcul de distance par cette fréquence élevée, tout en apportant peu d'information pertinente pour le calcul. Dans ce contexte, ils sont donc comparables à des mots grammaticaux. Cela doit être pris en compte pour le calcul, par une pondération à la baisse de ces mots.*

Si l'algorithme dans sa globalité est pertinent pour nos travaux, son exécution pose tout de même certains problèmes qu'il convient de résoudre. D'une part, et en accord avec la sous-hypothèse 1, la distance de Levenshtein est une valeur absolue, ce qui peut être un problème si les chaînes de caractère d'un couple sont très courtes : alors qu'une distance absolue de 1 est une distance faible pour des chaînes d'une vingtaine de caractères, elle peut en revanche être révélatrice d'une non similarité quand les chaînes font 3 à 4 caractères.

D'autre part, puisque cet algorithme permet de comparer, nous l'avons vu, une chaîne de caractères à une autre, d'un point de vue pratique, cela implique que le calcul est effectué sur une paire de chaînes à la fois. Le temps de calcul est donc proportionnel au nombre des chaînes à traiter. De plus, l'algorithme initial est de complexité *O(mn)* où m et n correspondent à la taille des chaines en nombre de caractères, soit une complexité quasi quadratique, puisque m et$n$ ne varient que très peu entre 2 versions [Bey, 2009]. Par conséquent, la complexité temporelle de l'algorithme d'origine est très importante. C'est ce qui motive notre sous-hypothèse 2.

Enfin, lors du calcul de distance d'édition, certains mots graphiques très récurrents servent à qualifier, à attribuer un type aux organisations. Ces mots, comme *Univ*, *Lab* ou *Institute*, jouent donc un rôle limité quant à la désignation d'une organisation, comme l'énonce la sous-hypothèse 3. De plus, par leur fréquence très élevée, ils peuvent venir fausser le calcul, établissant des correspondances entre deux noms d'organisations là où il n'y a pas lieu.

6.1.1.3 Application

Nous venons de voir que selon nos hypothèses, l'exécution de l'algorithme de Levenshtein devait être intégrée au système de normalisation de manière raisonnée et adaptée au contexte. L'algorithme en lui-même est pertinent pour notre objectif, mais il est relativement coûteux en temps. D'un autre côté, le coût de développement doit lui aussi être limité. Ce sont les conditions nécessaires à la mise en place d'un système qui puisse être rapidement opérationnel, et non d'un prototype trop long à implanter. Cette adaptation de l'algorithme porte donc sur trois points :

1. Selon la sous-hypothèse 1, la distance absolue doit être convertie en une distance relative.
2. Selon la sous-hypothèse 2, les couples de noms d'organisations qui seront envoyés pour le calcul de distance d'édition doivent d'abord être triés ;
3. Selon la sous-hypothèse 3, certains mots très fréquents, tels que *Univ*, *Lab* ou *Institute*, biaisent le calcul de distance. Ils doivent donc être pondérés.

Chacune de ces hypothèses a été appliquée à nos données, et aux calculs qui ont été effectués.

La création d'une distance relative. La distance de Levenshtein étant une valeur absolue, elle peut ne pas être pertinente lorsque les chaînes sont très courtes. Or, il n'est pas rare que nous soyons confrontée, dans nos données, à des noms d'organisations comme *Univ Boston* et

Univ Houston, qui font référence à des organisations distinctes. Calculer une distance relative, par la mise en rapport de la mesure avec la longueur des chaînes traitées, permet de pallier cette difficulté. Nous employons pour cela la formule :

$$DL_R(x,y) = \frac{DL(x,y)}{average(len(x), len(y))} \tag{6.1}$$

D'un point de vue pratique, nous procédons donc en trois étapes : nous calculons tout d'abord la longueur moyenne à partir de la longueur de chacune des deux chaînes du couple. Puis la distance absolue est calculée grâce à l'algorithme de Levenshtein. Enfin, cette valeur est relativisée puisque ramenée à la proportion qu'elle représente par rapport à la longueur moyenne des deux chaînes. Ici encore, la précision est améliorée, puisqu'une distance absolue de 2 sur des chaînes telles que *Univ Boston* et *Univ Houston*, considérée comme faible, aurait ramené ce couple comme étant des variantes potentielles d'un même nom. Or, la distance relative est une partie de la solution permettant de les éliminer.

Le tri des couples de noms d'organisations. Afin d'optimiser l'utilisation de l'algorithme de Levenshtein et de réduire le temps de traitement, nous avons pris le parti d'effectuer un tri préalable sur les couples candidats, afin d'éliminer les couples dont les membres ont peu de chances de faire référence à la même organisation, et d'en être des variantes typo-orthographiques. Nous postulons que deux chaînes de caractères aux longueurs très différentes ne peuvent être des variantes d'un même nom, du moins au niveau typo-orthographique [48]. Dans ce cas, effectuer un calcul de distance d'édition est inutile, et génère donc des traitements superflus.

Pour éliminer les couples candidats non pertinents, nous effectuons donc, comme un prétraitement, le calcul de différence entre le nombre de caractères des deux chaînes. Ce nombre est mis en rapport avec la longueur moyenne des deux chaînes. Si le résultat excède un plafond fixé à 19%, c'est-à-dire si l'écart de longueur entre les deux chaînes est supérieur à 19% de la longueur totale moyenne des chaînes, alors le couple candidat est considéré comme ne pouvant faire référence au même nom d'organisation, et est éliminé sans passer par le calcul de distance proprement dit. La proportion des couples sélectionnés en fonction de leur écart de longueur moyen est représentée schématiquement dans la figure 6.3 page suivante. Ce prétraitement permet de réduire le temps d'exécution de l'algorithme, qu'il soit effectué à la demande pour une seule chaîne de caractère à comparer à tout le reste du corpus, ou pour une comparaison totale de chaque chaîne à toutes les autres.

La pondération à la baisse des mots fréquents. Ainsi que nous l'avons signalé ci-dessus, certains mots graphiques reviennent de manière extrêmement fréquente dans les noms d'orga-

48. D'autres types de variantes existent, et sont traitées par d'autres moyens et processus.

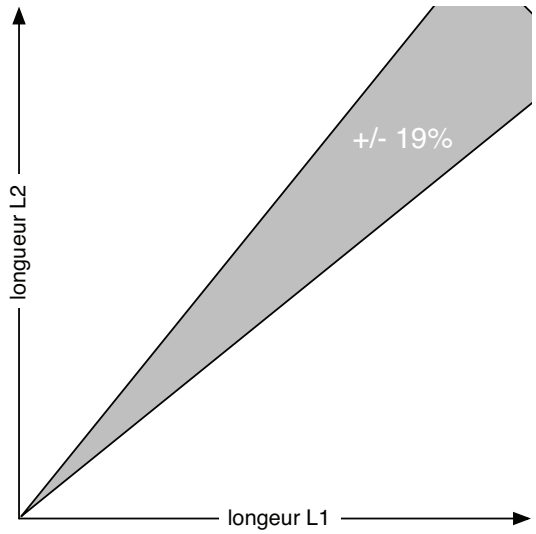

FIGURE 6.3 – Proportion des couples sélectionnés en fonction de leur écart de longueur moyen pour le calcul de distance d'édition

nisations. Des mots comme *Univ*, *Inst*, *Lab*, etc., ont des fréquences très élevées dans les noms d'organisations rattachés aux études, pour la simple raison qu'ils sont des mots génériques servant à désigner une « classe » d'organisations particulière, et non à désigner une organisation précise. Ils sont donc des identifiants d'organisations. L'inconvénient de cette fréquence importante est que deux noms d'organisations contenant tous deux *Univ* auront des chances d'être considérés comme similaires après calcul de la distance de Levenshtein, alors même qu'il s'agit d'organisations différentes. Le risque est d'autant plus grand que les chaînes sont courtes, puisque l'écart creusé par les caractères différents d'une chaîne à l'autre peut être comblé par les caractères égaux des deux exemplaires du mot *Univ* dans chacun des deux éléments de la paire. Un cas typique de cette proximité non pertinente est celui de la paire [*Univ Boston*, *Univ Houston*], qui a une distance assez faible de 2, qui pourrait laisser penser à tort que ces deux noms d'organisations renvoient en fait à une et une seule université.

Afin de diminuer l'impact négatif de ces identifiants très fréquents, que nous considérons, dans le contexte de ces données et de ce calcul, comme creux, nous avons pondéré à la baisse leur prise en compte dans le calcul de distance. Pour cela, chacun de ces mots, uniquement lorsqu'il apparaît dans les deux membres du couple de chaînes de caractères, est remplacé par un caractère unique. Ainsi, au lieu de compter comme 4 caractères égaux, *Univ* ne compte que pour un seul caractère dans le calcul. Cela vaut pour l'ensemble des désignations génériques d'organisations

que nous avons relevées dans le corpus, et rassemblées au sein d'une liste. Celle-ci est la seule ressource externe que nous utilisons dans ce calcul, et contient environ 140 items. La pondération, couplée au calcul d'une distance relative, nous permet d'augmenter la précision de l'algorithme, puisqu'elle évacue un certain nombre de couples non pertinents.

Cette pondération permet également d'améliorer le rappel, dans des configurations différentes : dans certains cas, la position des mots creux dont nous venons de parler est modifiée d'un nom à l'autre, pour désigner une même entité. Une distance non pondérée calculera souvent un écart trop grand, et ne permettra pas de rapprocher les deux membres d'un couple. En revanche, si les mots creux sont pondérés, et considérés comme un seul caractère, alors leur positionnement a moins d'impact sur le calcul. Par exemple, la distance absolue pour la paire *California Inst Technology* vs. *Inst California Technology* est de 10 sans pondération du mot creux *Inst*, contre 4 après pondération. Là encore, c'est la combinaison de la pondération et de la distance relative qui permet de détecter ces noms comme proches.

Nous résumons l'application de nos trois hypothèses sur la figure 6.6, à travers trois exemples de couples.

Hypothèse	Calcul	*Univ Houston* vs. *Univ Boston*	*Inst Stowers for Medical Research* vs. *Inst Stuwers for Medical Research*	*Thorlabs* vs. *School GW Woodruff Mechanical Engineering*
Les couples candidats doivent être triés sur leur longueur	Tri sur écart longueur <19%	8,69%	0	**134,7%**
La distance absolue doit être convertie en une distance relative	Distance → distance relative	2 → 17,40%	1 → 3,03%	Sans objet
Les mots creux doivent être pondérés	Distance relative → Distance relative pondérée sur les mots creux < 19%	**17,4% → 23,5%**	3,03% → 3,33%	Sans objet
Une distance de Levenshtein adaptée permet de repérer des proximités entre noms d'organisations	**Décision système**	**distincts**	**proches**	**distincts**

TABLE 6.6 – Application des calculs dérivés de nos hypothèses pour la distance de Levenshtein et décisions du système sur trois exemples

6.1.1.4 Validation

La mise en place de ces adaptations sur notre corpus de test nous a permis de valider nos hypothèses.

Notre première hypothèse appuyait sur l'importance de relativiser la valeur absolue de la distance de Levenshtein originale. En effet, une distance absolue, c'est-à-dire en nombre de

modifications de la chaîne, très faible peut ne pas être significative sur des chaînes courtes ; inversement, une distance plus importante n'est pas forcément le signe de dissimilarités entre deux chaînes très longues. De fait, en l'état, il était impossible d'évaluer la distance à laquelle placer la frontière entre couples de chaînes considérées comme variantes d'un même nom d'organisation, et chaînes référant probablement à deux organisations différentes. Aucun repère ne nous permettait de trancher sur ce point avec un nombre absolu. Relativiser cette distance par rapport à la longueur moyenne des chaînes du couples a permis de fixer un plafond au-delà duquel les membres d'une paire sont considérés comme non pertinents. Ce plafond a été fixé à 0,19. Des tests manuels sur échantillons ont permis de déterminer cette valeur comme étant le meilleur compromis entre précision et rappel. Nous présentons dans le tableau 6.7 les résultats de ces tests manuels, avec des plafonds fixés de 0,17 à 0,20. Le corpus de test est un échantillon d'une étude de TKM sur les caméras haut sensibilité et haute rapidité, constituée de 3 653 noms d'organisations.

Plafond	Couples détectés	Couples détectés à tort	Taux de bruit	Couples pertinents
0,17	113	26	23%	87
0,18	122	29	23,7%	93
0,19	133	32	24,1%	101
0,20	155	45	29%	110

TABLE 6.7 – Tests manuels pour fixer le plafond de distance d'édition, sur un échantillon de 3 653 organisations

Nous considérons que sont détectés à tort les couples de noms d'organisations faisant référence à deux organisations bien distinctes.

Au vu de ces résultats, nous avons pris le parti de privilégier la précision au rappel, et de fixer le plafond à 0,19. Un plafond fixé à 0,2 permet en effet de récupérer plus de couples de variantes, mais la contrepartie en est la génération de plus de bruit. Nous ne souhaitons cependant pas tomber dans l'excès inverse en éliminant trop de couples pertinents, et c'est pourquoi nous n'avons pas fixé le plafond à 0,17 ni à 0,18. Par ailleurs, puisque ces appariements entre noms d'organisations sont voués à être évalués par l'utilisateur, nous privilégions la précision de manière à ne pas lasser l'utilisateur.

Les types d'erreurs rencontrées dans certains couples ramenés pour chaque seuil sont présentés dans le tableau 6.8 page suivante.

A un plafond fixé à 0,2, des couples tels que *C T Electronics* vs. *Nec Electronics* sont ramenés. Ces erreurs sont dues à une distance faible entre les deux membres du couple, par rapport à la

Erreurs: couples détectés à tort	Seuil
Univ Trente *vs.* Univ Twente	0,17
Univ Graz Technology *vs.* Univ Brno Technology	0,18
Univ Iran Science Technology *vs.* Univ Jiangsu Science Technology	0,19
C T Electronics *vs.* Nec Electronics	0,2

TABLE 6.8 – Types d'erreurs, répartis par plafond pour la distance d'édition relative

longueur globale. En l'occurrence, la distance est de 3 pour une longueur moyenne de 15, soit une distance relative de 0,2. La présence de *Electronics*, qui n'est pas un identifiant considéré comme mot creux dans nos données mais qui est relativement générique, provoque de telles erreurs dans un certain nombre de cas.

Dans les couples proposés comme proches à un plafond de 0,19, les sources d'erreurs sont comparables, mais souvent sur des chaînes plus longues. En effet, plus les chaînes sont longues, plus la distance doit être importante pour être considérée comme significative. *Univ Iran Science Technology* vs. *Univ Jiangsu Science Technology*, par exemple, ont une distance de 18,9% après pondération des mots « creux » (16,97% sans pondération). Bien que les universités soient ici distinctes, les chaînes dans lesquelles elles s'insèrent restent proches, en particulier en raison de la présence de Science et de Technology. Ces deux mots sont relativement courants dans l'échantillon, et de manière générale dans les noms d'organisations. Ils sont donc sélectionnés bien qu'ils réfèrent à deux organisations différentes.

A un plafond de 0,18, les problèmes posés sont sensiblement les mêmes. La seule variante est la distance encore réduite, même lorsque les noms renvoient à deux entités différentes. Deux noms d'université courts, assortis de mots récurrents et relativement génériques comme *Technology*, peuvent rester sous le plafond de 0,18. Par exemple, *Univ Graz Technology* vs. *Univ Brno Technology* ont une distance absolue de 3, et une distance relative de 17,65% après pondération. La taille réduite des chaînes *Graz* et *Brno* empêchent le calcul de distance d'être performant sur ce cas.

Enfin, *Univ Trente* vs. *Univ Twente* illustrent les cas causant encore du bruit dans les résultats à un plafond de 0,17. Ces chaînes courtes diffèrent d'un seul caractère. De fait, même après la pondération des mots creux, elles atteignent une distance trop faible pour être considérées comme référant à des entités différentes.

Pour certaines entités, il est difficile de mettre en place des traitements permettant de les distinguer sans faire baisser le rappel de façon trop importante. Cependant, ces cas limites restent

minoritaires, et n'empêchent pas l'intérêt global du calcul de distance de Levenshtein.

Un plafond fixé à 0,19 permet donc un certain équilibre entre précision et rappel, sachant malgré tout que nous privilégions en dernier ressort la précision, de façon à ne pas provoquer la lassitude de l'utilisateur s'il doit corriger les noms proposés.

Notre deuxième hypothèse pose que faire un premier tri, fondé sur la différence de longueur entre les deux chaînes d'un couple, permettrait d'éliminer un certain nombre de couples candidats pour lesquels la probabilité de porter deux variantes typo-orthographiques d'un même nom d'organisation était minime. L'objectif en était de réduire le temps de traitement. Pour notre échantillon de 3 653 noms d'organisations, la combinatoire s'élève à 3 653*3 653, c'est-à-dire 13 344 409 couples calculés. En revanche, si les couples sont triés avant d'être envoyés au calcul de distance lui-même, cette combinatoire est réduite à 1 559*1 559, soit 2 430 481 calculs. Nous soulignons que ce pré-traitement n'a aucune incidence sur le taux de rappel du traitement.

Les couples ne sont éliminés que si leur différence de longueur excède 19%, ce qui équivaut finalement, en pratique, à une distance de Levenshtein relative de 19% où toutes les opérations reviendraient à des effacements ou à des insertions, selon que la chaîne A est la plus longue ou la plus courte. Le choix de ce seuil est motivé par le fait que nous avons posé, en vertu de la première hypothèse concernant l'intérêt d'une distance relative, que seuls seraient conservés les couples de chaînes dont la distance relative serait inférieure ou égale à un plafond de 19%. Par conséquent, un couple éliminé lors du pré-traitement de tri n'aurait de toute façon pas pu être proposé à l'issue du calcul de distance.

La mise en place de ce pré-traitement permet donc de réduire le temps de traitement total des données. Cependant, il n'influence pas la qualité des résultats. Il est donc nécessaire de mettre en place des traitements qui permettront d'optimiser le calcul pour nos données.

Notre dernière hypothèse, relative aux mots creux récurrents du corpus, postulait que pondérer ces mots à la baisse permettrait de gagner en précision dans les résultats de calcul de distance. Sur notre corpus de test de 3 653 organisations, nous avons donc appliqué deux versions du calcul : la première ne pondère pas les mots creux, contrairement à la seconde. Certaines paires de noms auraient été ramenées à tort sans cette pondération, tandis que d'autres auraient été mises de côté malgré une proximité. Nous en présentons quelques exemples dans le tableau 6.9.

La pondération de ces noms permet donc de travailler à la fois sur la précision et sur le rappel. Cependant, ces améliorations n'influencent pas les données dans les mêmes proportions : l'élimination de couples distincts est plus large que la récupération de couples proches non détectés sans pondération.

Nous présentons les résultats généraux de cette évaluation dans le tableau 6.10.

Ces chiffres incluent à la fois les couples éliminés grâce à la pondération, et ceux qui sont

Mesure augmentée	Exemple	Distance relative non pondérée	Distance relative pondérée	Statut des noms du couple
précision	Univ China Agricultural *vs.* Univ Jilin Agricultural	17,39%	20%	distincts
précision	Univ Hanyang *vs.* Univ Shaoyang	16%	21,05%	distincts
rappel	Massachusetts General Hospital *vs.* Hospital Massachusetts General	60%	17,39%	proches
rappel	California Inst Technology *vs.* Inst California Technology	38,46%	17,39%	proches

TABLE 6.9 – Exemples de l'effet de la pondération sur des couples de noms problématiques

	Nombre de couples	Nombre d'erreurs	Bruit (%)	Précision (%)
Version non-pondérée	142	43	30,3	69,7
Version pondérée	133	32	24,1	75,9

TABLE 6.10 – Evaluation de la méthode de pondération des mots creux pour le calcul de distance d'édition

repérés grâce à elle. De la version non pondérée à la version pondérée du calcul, la précision passe de 69,7% à 75,9%. En considérant les couples récupérés en plus grâce à la pondération, le rappel est légèrement augmenté, puisque le nombre de couples détectés comme proches à juste titre passe de 99 à 101. Cependant, l'effet sur le rappel est moindre, d'autant que la contrepartie de l'élimination des couples non proches est aussi, parfois, le rejet de couples qui auraient dû être rapprochés. Par exemple, le couple *Univ Tsinkhua* vs. *Univ Tsingnua* est rejeté en raison de la pondération du mot creux *Univ*, alors que les deux noms font référence à la même organisation. Nous privilégions ici la précision, puisque par la suite, les couples sélectionnés par le calcul de distance d'édition seront soumis aux utilisateurs. Or, pour éviter de présenter à celui-ci des données trop bruitées, il est plus important que la précision soit élevée.

Grâce à cette pondération des mots creux récurrents, la précision des résultats du calcul de distance est donc améliorée. Bien que le calcul lui-même ne change pas, les unités auxquelles il est appliqué, elles, sont modifiées. Les couples restant sous le plafond des 19% sont plus pertinents en tant que variantes potentielles d'un même nom d'organisation. Il reste du bruit dans les données, mais le traitement est relativement efficace. Pour l'aspect endogène traitant les caractéristiques typographiques de la normalisation, nous considérons donc que le traitement est satisfaisant. Ses faiblesses pourront être compensées par les autres types de processus employés par la suite.

Nous présentons dans la table 6.11 page ci-contre le déroulement procédural du traitement

de distance d'édition. Les trois étapes de traitement correspondent à nos trois hypothèses. Nous illustrons chaque étape par un exemple.

	Etape	Exemples
1	Tri des couples candidats sur comparaison de la longueur de chaque élément du couple	*Univ Salford* vs. *Univ Stanford* est conservé *Univ Salford* vs. *Univ Colorado Health Sciences Center* est éliminé.
2	Pondération des mots creux	*Univ Salford* vs. *Univ Stanford* → *X Salford* vs. *X Stanford*
3	Calcul de distance relative	$Da_{(X\ Salford,\ X\ Stanford)} = 2$ $Dr_{(X\ Salford,\ X\ Stanford)} = 0{,}21$

TABLE 6.11 – Récapitulatif des étapes de calcul de distance de Levenshtein, par ordre procédural

Dans l'ordre procédural, nous faisons intervenir la pondération des mots creux seulement après le tri sur les longueurs des chaînes. En effet, nous avons tenu à restreindre au mieux les traitements effectués sur l'ensemble des noms d'organisations, de manière à limiter le temps de traitement. C'est pourquoi la pondération n'est appliquée que sur les couples qui ont été préalablement triés.

La distance de Levenshtein, processus endogène dans l'application que nous en faisons puisque chaque entité nommée d'un ensemble documentaire est traitée à l'aide des autres entités nommées de cet ensemble, permet de détecter des couples de noms d'organisations similaires, et donc référant potentiellement à la même organisation. Notons qu'en revanche, cet algorithme ne permet en aucun cas de trancher automatiquement entre les deux variantes du couple pour sélectionner celle qui est « correcte », si tant est que l'une des deux soit correctement formulée. Par conséquent, cette méthode doit être complétée par d'autres, qui permettent de prendre des décisions à partir des suggestions du système.

6.1.2 Découper et hiérarchiser les entités nommées

Parmi les sources de problèmes causés par les données contenant les entités nommées que nous avons listées au début de ce chapitre (voir le dernier couple de variantes de la figure 6.1 page 160), nous avons cité le problème posé par le découpage de séquences contenant les noms de deux organisations distinctes, mais incluses l'une dans l'autre. Dans ces cas-là, un nom d'organisation brut donne non pas une seule entité nommée, mais peut fournir toute sa hiérarchie, particulièrement pour les entités nommées de type organisme. Les entités sont donc non plus simples, mais complexes, et structurées.

Le découpage entre les différentes organisations peut être explicite, grâce à la présence de la virgule séparant les deux organisations parentes, qui est un délimiteur fiable dans nos données. En effet, la virgule est un délimiteur d'organisations liées dans 98,14% des noms contenant ce signe. Le seul autre usage recensé est celui de la virgule permettant une énumération à l'intérieur même d'un nom d'organisation, comme dans *Department of Pharmacology, Physiology, Radiology, and Biomedical Engineering, Wayne State University School of Medicine, Detroit, MI 48201, United States*, où le département a un nom complexe.

Dans les données que nous avons traitées, la présence de la virgule comme séparateur de noms d'organisations reste majoritaire, et représente environ 72% du total des noms d'organisations sur notre corpus de test. Néanmoins, il arrive qu'il n'y ait pas d'indice typographique permettant d'identifier la frontière entre plusieurs entités liées hiérarchiquement : le découpage est implicite.

Notre corpus de test est composé de 3 653 organisations, à partir du corpus d'une étude sur l'optique, à l'origine beaucoup plus important. La proportion des types d'organisations de l'étude d'origine a été respectée concernant les organismes publics et assimilés d'une part et les entreprises d'autre part : notre extrait contient 1 335 noms d'entreprises et 2 318 noms d'organismes [49], tous pré-normalisés. A partir de ce corpus, nous avons détecté 40 cas de noms d'organisations à entités multiples, représentant en tout 54 occurrences, puisque certains cas sont répétés plusieurs fois. Ce phénomène représente donc environ 1,5% de la totalité des occurrences de noms d'organisations sur notre corpus, contre une séparation par des virgules des organisations liées dans 72,4% des cas.

Bien que ce cas de figure soit minoritaire, il est assez courant pour devenir problématique lors du traitement de grands volumes de données. Nous présentons ici un exemple de chaque configuration :

1. *Lund University, Department of Medicine*
2. *Univ Kentucky Medical Center*

Dans tous les cas, les différents noms d'organisations doivent être correctement identifiés, et segmentés sur le modèle *Lund University // Department of Medicine* pour 1, et *Univ Kentucky // Medical Center* pour 2. Cependant, si le découpage explicitement exprimé par la virgule permet une segmentation efficace, d'autres moyens doivent être trouvés pour les cas de découpage implicite. Il est pour cela possible de faire appel à des méthodes endogènes, et en particulier des techniques utilisant les fréquences des items d'une séquence donnée.

Sur nos deux exemples, seul le second est concerné par ce problème de découpage implicite. Le centre médical appartient à l'université du Kentucky, mais aucune virgule, ni autre signe typographique, n'est présent pour signaler la distinction entre centre et université. Il convient donc de fixer la limite entre les deux noms d'organisations, de façon à les découper efficacement.

49. Les noms d'organismes correspondent aux organisations publiques, et par extension, aux hôpitaux, universités et écoles privés ; voir à ce sujet la sous-section 3.2.2.9 page 94.

Pour cela, il est difficile de faire appel à des règles syntaxiques fiables. En effet, la syntaxe dans les noms d'organisations placés dans des champs de bases de données est réduite au minimum, et les indices lexicaux révélateurs de cette structure sont donc particulièrement peu présents. Cela est d'autant plus valable pour l'anglais, qui utilise moins de mots outils que le français, en particulier pour la fonction syntaxique de modifieur du nom.

Cela s'explique en partie par le fait que l'anglais, d'après [Tesnière, 1959] et sa typologie syntaxique, est une langue centripète mitigée, contrairement au français par exemple, langue centrifuge mitigée. Alors que le français par exemple utilisera fréquemment des modifieurs de nom de la forme *de + Syntagme Nominal* venant se placer après le nom modifié, l'anglais utilisera souvent comme modifieurs, non plus des syntagmes prépositionnels, mais des formes nominales adjectivales [50], placées avant le nom modifié. Ainsi, là où des patrons lexico-syntaxiques auraient permis de découper une suite brute en français comme *Centre médical de l'université du Kentucky* grâce à la présence du mot vide *de*, rien n'indique la « frontière » pour *Univ Kentucky Medical Center*, où *Univ Kentucky* est le syntagme nominal adjectival qui joue le rôle de l'un des modifieurs du nom *Center*.

Nous présentons dans le tableau 6.12 des exemples de noms d'organisations pré-normalisés typiques de cette configuration syntaxique, problématiques parce que contenant plusieurs entités, tous issus d'un même corpus de 3 653 noms d'organisations.

Nom pré-normalisé	Entités à distinguer
Univ Maryland Biotechnology Inst	Univ Maryland
	Biotechnology Inst
Cincinnati Childrens Hospital Medical Ctr	Cincinnati Childrens Hospital
	Medical Ctr
Office Science Engineering Lab	Office Science Engineering
	Lab
Univ Chicago Medical Ctr	Univ Chicago
	Medical Ctr
Univ Georgia Research Fdn	Univ Georgia
	Research Fdn

TABLE 6.12 – Exemples de noms d'organisations pré-normalisés à entités multiples

Pour remédier à cela, nous pouvons nous appuyer sur les données du corpus à traiter, en partant du postulat que la majorité des données sont suffisamment fiables, au moins sur certains aspects, afin d'en tirer des informations pertinentes pour le découpage et la structuration des

50. Nous empruntons le terme de *nom adjectival* à [Maniez, 2001].

noms d'organisations complexes. Nous faisons donc l'hypothèse globale suivante :

Hypothèse II.2. *Les données du corpus, c'est-à-dire les métadonnées saisies par les auteurs des documents en présence, sont suffisamment fiables par leur masse pour en tirer de l'information pertinente. Des traitements endogènes, principalement statistiques et fondés sur la récurrence des phénomènes dans le corpus à traiter, permettent de dégager à partir d'une séquence entière les noms d'organisations les plus probablement « corrects ».*

Dans ce qui suit, nous nommons *séquence* le nom d'organisation tel qu'il est présent dans nos données après des traitements permettant de nettoyer et pré-normaliser les noms d'organisations. Pour notre exemple, la séquence est donc le nom pré-normalisé *Univ Kentucky Medical Center*. Nous définissons dans un premier temps comme une *sous-séquence* une suite de n mots graphiques extraite d'une séquence. Pour notre exemple, *Univ Kentucky Medical* est une des sous-séquences possibles. Enfin, nous considérons qu'un identifiant est un mot indiquant la présence d'une organisation, et qui a un caractère générique. Par exemple, *Univ* et *Center* sont des identifiants. Nous synthétisons la liste des ces unités dans le tableau 6.13.

Séquence	Sous-séquences possibles		Commentaire
Univ Kentucky Medical Ctr	Univ	Kentucky Medical Ctr	Première séquence non pertinente puisqu'elle contient un identifiant *Univ* seul
	Univ Kentucky	Medical Ctr	Noms corrects
	Univ Kentucky Medical	Ctr	Séquences incorrectes
	Univ Kentucky Medical Center	∅	Séquences non pertinentes puisque la première contient deux identifiants et que la deuxième est vide

TABLE 6.13 – Exemples d'unités utilisées pour les traitements endogènes de fréquence et de surface

Dans la première paire de séquences, l'une d'elles ne contient que l'identifiant *Univ* seul. Or, nous considérons que dans le cas particulier des universités, un nom est toujours constitué au minimum d'une amorce *Univ* et d'un mot graphique complétant cet identifiant. Cette séquence n'est donc pas pertinente pour nous.

La deuxième paire de séquences contient les deux noms d'organisations corrects.

La troisième paire de séquences est incorrecte, puisque la première mêle des éléments des deux noms d'organisations présents, et que la seconde est incomplète.

Enfin, la dernière séquence n'est pas pertinente, puisqu'elle contient les deux identifiants d'organisations.

Notre objectif est donc de déterminer, grâce à nos traitements, où se trouve la frontière qui permet de départager les deux noms d'organisations restitués dans les deux sous-séquences correctes.

6.1.2.1 Fréquences

Afin de délimiter les deux noms d'organisations présents dans la même séquence *Univ Kentucky Medical Center*, nous utilisons une approche endogène fondée sur la récurrence de séquences ou sous-séquences. Nous suivons en cela [Vergne, 2004], qui expose une méthode d'analyse syntaxique partielle sans ressources de corpus multilingue ; cette méthode s'appuie entre autres sur la longueur et la fréquence des mots. Bien que notre objectif diffère de celui de l'auteur, nous reprenons à notre compte un certain nombre de principes et les adaptons à notre travail. Par ailleurs, [Déjean, 1998], dans ses travaux portant sur la recherche de structures formelles dans des langues inconnues, a lui aussi travaillé sur des critères tels que les effectifs - ce que nous nommons *fréquence* - de séquences textuelles. Selon lui, « l'effectif d'une séquence de mots [est] une indication de la mise en relation de ces mots ».

Nous faisons le choix d'utiliser le principe de fréquence (ou d'effectif), tel qu'il a été exploité par [Vergne, 2004] ou [Déjean, 1998], pour tester l'ensemble des sous-séquences d'une séquence contenant plusieurs noms d'organisations.

Nous partons en effet de l'hypothèse suivante :

Hypothèse II.3. *Un nom d'organisation correctement délimité apparaîtra plus souvent dans un ensemble documentaire qu'un nom d'organisation incohérent ou dont le référent n'existe pas en tant que tel dans le monde réel. La sous-séquence la plus fréquente dans un ensemble de noms d'organisations pré-normalisés donné est donc le nom d'organisation le plus probable, et par conséquent le plus cohérent.*

Pour notre séquence 2, la sous-séquence *Univ Kentucky* devrait donc être plus fréquente que la sous-séquence *Univ Kentucky Medical*, cette dernière étant incohérente et ne correspondant strictement à aucune organisation existante. Il en va de même pour les sous-séquences *Medical Center* et *Kentucky Medical Center*, où nous supposons que la première sera plus fréquente que la seconde, puisque le Kentucky Medical Center n'existe pas en tant que tel - même si dans l'absolu, rien dans *Kentucky Medical Center* n'est incorrect syntaxiquement ni sémantiquement. Ainsi, c'est ici la référence qui prime sur le sens, puisqu'en tant qu'entités nommées, ces sous-séquences sont censées référer à des éléments du réel. A ce titre, les sous-séquences référant effectivement à des organisations existantes devraient être, en toute logique, les plus fréquentes.

Tout d'abord, nous définissons plus précisément l'unité *sous-séquence* : une sous-séquence est une suite de n mots graphiques, ou items, issus de cette séquence, dans une fenêtre dont la taille

est incrémentée d'un item pour chaque calcul de fréquence.

Application Ce calcul de fréquence n'est appliqué qu'aux noms pré-normalisés qui contiennent au moins deux organisations au sein d'une même séquence, et uniquement si les indices syntaxiques ne permettent pas de les découper correctement lors des étapes de pré-normalisation. Il sera donc appliqué au nom *Univ Maryland Biotechnology Inst* de l'exemple 2 présenté plus haut, mais pas à l'exemple 1, soit *Lund University, Department of Medicine*.

Ainsi, pour la séquence *Univ Maryland Biotechnology Inst*, des sous-séquences seront testées jusqu'à dégager le nom d'organisation principal. En l'occurrence, nous cherchons donc à obtenir *Univ Maryland*. La sous-séquence restante sera considérée de fait comme le nom de l'organisation dépendant hiérarchiquement de la première, soit ici, *Biotechnology Inst*. Pour ce faire, nous avons réduit l'utilisation des ressources exogènes au minimum : nous utilisons notamment la même liste d'identifiants que pour le calcul de distance d'édition (voir la sous-section précédente), c'est-à-dire de mots qui permettent de détecter la présence d'une organisation, ainsi que son niveau hiérarchique. Par exemple, une université, détectée par l'identifiant *Univ*, est considérée comme hiérarchiquement supérieure à un institut, repéré par *Inst*.

Le découpage en sous-séquences s'exécute à partir du premier identifiant rencontré dans le sens de la lecture, à laquelle un mot est ajouté pour la première sous-séquence. Si des mots se trouvent à gauche du premier identifiant, ils ne sont pas pris en compte dans le calcul. En effet, puisqu'ils ne se trouvent pas placés entre les deux identifiants, nous considérons que leur rattachement n'est pas ambigu : étant donnée la taille restreinte d'un nom d'organisation, et sa simplicité syntaxique relative, nous postulons en effet que les mots se trouvant isolés à la gauche du premier identifiant lui sont rattachés. La taille de la fenêtre ainsi créée est incrémentée de 1 pour chaque nouvelle sous-séquence. Le test des sous-séquences s'arrête lorsque le système détecte le second identifiant, puisque nous savons déjà que les noms d'organisations à dégager ne peuvent contenir qu'un seul identifiant chacun. Pour *Univ Maryland Biotechnology Inst*, les sous-séquences testées sont donc les suivantes :

1. *Univ Maryland*
2. *Univ Maryland Biotechnology*

Le système ne va pas plus loin, puisqu'il rencontre, après *Biotechnology*, le second identifiant, *Inst*, qui n'appartient évidemment pas au nom d'organisation principal le plus cohérent. Pour chacune de ces sous-séquences, le système va donc calculer leur fréquence dans l'intégralité des formes pré-normalisées de l'ensemble documentaire traité.

Dans les cas où un seul mot serait présent entre les deux identifiants, deux sous-séquences sont alors prises en considération : celle contenant le premier identifiant rencontré et le mot en question, et celle contenant celui-ci et le second identifiant. Cela permet de dégager des infor-

mations permettant une comparaison. Ainsi, dans *Sherbrooke Univ Biol Lab*, les sous-séquences considérées seront :

1. *Univ Biol*
2. *Biol Lab*

Puisque l'item *Sherbrooke* se trouve à gauche de l'identifiant, il n'est pas pris en compte dans le calcul : il est rattaché à l'identifiant *Univ*.

A partir d'un ensemble documentaire contenant 3 653 noms d'organisations pré-normalisés, nous avons relevé les résultats restitués dans le tableau 6.14.

Sous-séquence	Fréquence
Univ Maryland	**13**
Univ Maryland Biotechnology	1

TABLE 6.14 – Fréquence des sous-séquences de la séquence *Univ Maryland Biotechnology Inst*

La sous-séquence la plus fréquente est donc *Univ Maryland*. C'est en effet la suite d'items la plus commune, qui a été comptabilisée dans des noms pré-normalisés comme :

1. *Univ Maryland School Medicine*
2. *Univ Maryland*

Sur les 13 occurrences de *Univ Maryland* relevées, seule l'une d'entre elles se poursuivait par l'item *Biotechnology*.

Les séquences sont incrémentées de gauche à droite, et non de droite à gauche, de façon à respecter le sens de la production linguistique du message, en tout cas dans les langues effectivement traitées dans notre corpus, et ainsi de trouver une cohérence dans les séquences testées.

La séquence contenant le seul item *Univ* n'est pas testée : elle ne présente pas d'intérêt, puisque de manière générale, les données ne présentent pas de noms d'université composés seulement de l'identifiant *Univ*. De même, nous ne comptabilisons pas les fréquences des autres items isolés. Cependant, à titre d'information, nous montrons dans le tableau 6.15 la fréquence de chacun des items de la séquence testée.

En sélectionnant la sous-séquence la plus fréquente, nous sommes bien en mesure de détecter le nom d'organisation principal le plus cohérent, soit :

1. *Univ Maryland*

Par ricochet, l'organisation dépendant de cette université est donc ramenée à la forme *Biotechnology Inst*.

Item	Fréquence
Univ	1277
Maryland	13
Biotechnology	2
Inst	292

TABLE 6.15 – Fréquence des items de la séquence *Univ Maryland Biotechnology Inst* dans notre corpus

Validation Nous avons appliqué ce calcul à l'ensemble de corpus de 3 653 organisations, qui porte sur le domaine de l'optique et des caméras à haute sensibilité.

Sur ces 3 653 organisations, 55 présentent un problème de découpage, ainsi que nous l'avons mentionné dans les deux sous-sections précédentes. 41 d'entre elles sont des noms distincts, et 7 d'entre eux cumulent en tout 21 répétitions de leurs instances. L'un de ces noms d'organisation a un découpage incertain, car les informations véhiculées, même dans le nom brut, ne permettent pas de dégager manuellement la structure du nom. Nous ramenons donc le nombre d'organisations distinctes à 40, et le nombre total d'occurrences à 54.

Si nous considérons seulement ces occurrences distinctes, le calcul des fréquences permet un découpage correct dans 26 cas sur 40, soit un taux de réussite de 65%. Si nous prenons en compte la totalité des occurrences, incluant les répétitions, ce ratio est élevé à 44/54, soit 81,4% des noms pré-normalisés.

La plupart des erreurs sont dues à deux types de séquences :

- Les séquences peu fréquentes dans le corpus : il est difficile de tirer des données des tendances fondées sur la régularité lorsque la fréquence est réduite. Par exemple, pour *Inst Optics & Precision Mechanics Chinese Acad*, si la suite *Inst Optics* se retrouve régulièrement dans le corpus, ce n'est pas le cas de la suite correcte ici, soit *Inst Optics & Precision Mechanics*, qui n'apparaît qu'une seule fois. C'est également le cas de *Office Science Engineering Lab*, pour lequel à la fois *Office Science*, la suite incomplète, et *Office Science Engineering*, la séquence correcte, n'apparaissent qu'une seule fois. Les fréquences ne permettent donc pas de trancher.
- Les séquences contenant des mots composés entre les deux identifiants : le calcul des fréquences peut générer des erreurs de découpage, ou ne pas permettre de décider où placer la frontière. Pour la séquence *Univ New York Health Science Ctr* par exemple, la fréquence de *Univ New* est supérieure à celle de *Univ New York*, puisqu'il existe dans le corpus des noms tels que *Univ New South Wales* qui font augmenter la fréquence de *Univ New*.

Ces calculs de fréquence, fondés sur les données du corpus, sont donc un bon indice pour découper des séquences à noms d'organisations multiples. Cependant, les taux de réussite obtenus

peuvent être améliorés, notamment par la prise en compte des types de cas problématiques, de manière à augmenter la proportion de segmentations adéquates. En effet, certains cas de figure ne peuvent être résolus par ce calcul. C'est pourquoi il doit être enrichi par d'autres paramètres des sous-séquences testées.

6.1.2.2 Surfaces

Le calcul de fréquence n'est pas optimal, et atteint ses limites sur des cas plus complexes que celui que nous venons de traiter, mais qui obéissent cependant au même schéma syntaxique, c'est-à-dire : *forme nominale adjectivale + syntagme nominal*. Sur des noms pré-normalisés comme *Univ New York Health Science Ctr*, les fréquences peuvent donc s'avérer inefficaces. La présence d'un nom composé, comme *New York*, couplée à la qualification nominale de droite à gauche en anglais, peut amener à des résultats ne permettant pas un découpage correct. Testées sur le même ensemble documentaire que précédemment, les fréquences pour les sous-séquences de *Univ New York Health Science Ctr* sont recensées dans le tableau 6.16.

Sous-séquence	Fréquence
Univ New	**14**
Univ New York	10
Univ New York Health	1
Univ New York Health Science	1

TABLE 6.16 – Fréquence des sous-séquences pour la séquence *Univ New York Health Science Ctr*

Si nous suivons les résultats des fréquences comme nous l'avons fait jusqu'à présent, le nom le plus probable pour l'organisation principale, celle qui englobe la seconde, est donc *Univ New*. Or, ce résultat est erroné, puisqu'en l'occurrence, le nom que nous cherchons à récupérer est *Univ New York*. Cette erreur s'explique par le fait qu'un certain nombre de noms d'université, tirés de leur lieu, comportent l'item *New*. De fait, il suffit qu'une université ait un nom commençant par la sous-séquence *Univ New*, comme *Univ New South Wales* ou *Univ New Brunswick*, pour que la fréquence de cette sous-séquence soit supérieure à celle de *Univ New York*.

Nous partons du postulat selon lequel, pour résoudre ce problème, il est possible de mettre en rapport la fréquence des séquences avec leur longueur, ce qui est une manière de pondérer ces fréquences. En effet, nous venons d'expliquer que nous fonder sur des fréquences implique que nous considérons que le plus fréquent est le plus probable. Cela peut s'avérer juste, mais pose

problème pour les noms composés. Nous formulons donc deux hypothèses :

Hypothèse II.4. *Pondérer la fréquence des sous-séquences par leur longueur permet de donner plus de poids aux séquences plus longues : une séquence longue et cohérente a des chances d'être moins fréquente qu'une séquence plus courte, pour peu que celle-ci soit cohérente. Il convient donc de prendre en compte la longueur des chaînes, en nombre de mots graphiques, pour revoir le poids de ces suites plus longues à la hausse.*

Cette pondération comporte le risque de générer des erreurs, risque toutefois limité : nous postulons qu'une suite longue mais non cohérente aura une fréquence si faible que, même pondérée à la hausse, elle ne dépassera pas celle d'une suite plus courte et cohérente.

Application Nous avons donc mis en place un calcul de *surface* des sous-séquences, qui met en rapport leur fréquence et leur longueur. Le calcul de surface pour une sous-séquence est le suivant :

$$Surface = nombre\ d'items\ de\ la\ sous\ séquence * fréquence\ de\ la\ sous\ séquence$$

Il s'applique sur les mêmes unités que le calcul de fréquence, et dans les mêmes conditions, c'est-à-dire uniquement dans le cas où un nom d'organisation pré-normalisé contient encore deux identifiants d'organisations.

Nous présentons les résultats de ces calculs sur la figure 6.4 page ci-contre pour les sous-séquences tirées de *Univ New York Health Science Ctr.*

La surface d'une sous-séquence représente donc le rapport entre sa fréquence et sa longueur, qui prend sur la figure la forme d'une aire.

Ici, le meilleur score de surface est celui de la séquence *Univ New York*, soit le nom que nous avions identifié manuellement comme le plus cohérent. Les calculs de surface nous permettent donc, dans un certain nombre de cas, de résoudre le problème des mots composés sans mots grammaticaux permettant de les découper.

Validation Sur notre corpus de test, les surfaces découpent correctement 28 noms sur 40 organisations distinctes, et 37 noms sur 54 occurrences au total. Ces noms correctement segmentés correspondent respectivement à 70% et 68,5% du total.

Nous avons vu que dans le cas de *Univ New York Health Science Ctr*, le calcul de surface permettait de découper correctement les deux organisations en présence. Ce traitement pallie donc certaines faiblesses du calcul de surface. Les noms d'organisations contenant des noms

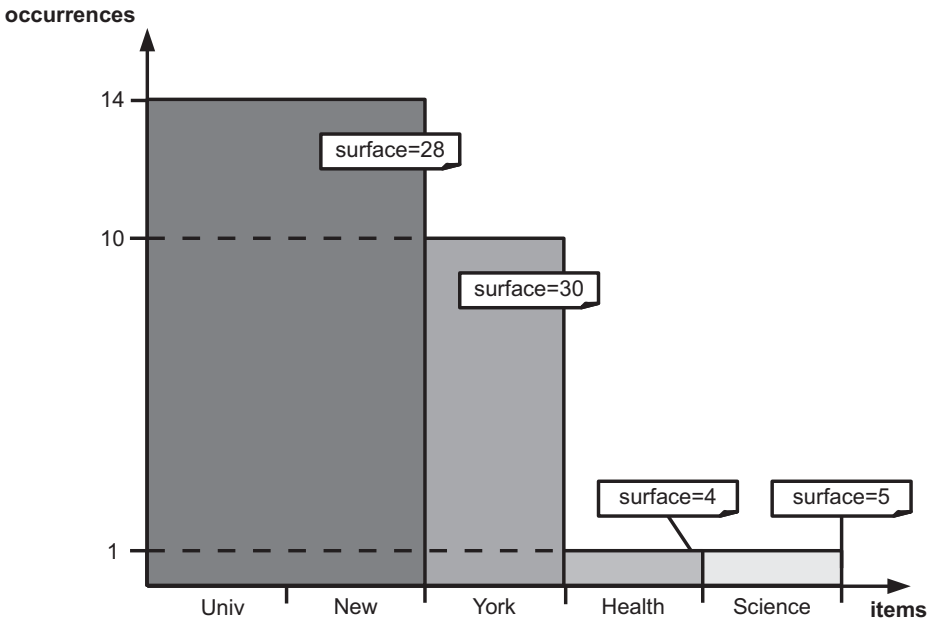

FIGURE 6.4 – Calcul des surfaces des sous-séquences de *Univ New York Health Science Ctr*

composés adjectivaux, en particulier, sont dans la majorité des cas découpés plus efficacement par ce biais. Cependant, certains noms pré-normalisés posent des problèmes récurrents, non résolus par cette méthode. Ces noms peuvent avoir l'une et/ou l'autre de ces caractéristiques :

1. ils contiennent des adjectifs (ou des noms ou groupes nominaux adjectivaux) qui revêtent un caractère générique dans l'ensemble documentaire étudié ;
2. dans un ensemble documentaire donné, soit dans les données utilisées pour le calcul, le nom de l'organisation principale apparaît très fréquemment avec une même organisation secondaire ;
3. les mots outils qui auraient dû permettre une découpage lors de la pré-normalisation ont été omis par le scripteur lors de la saisie des données.

Pour les noms pré-normalisés relevant du premier cas, des mots tels que *Catholique* ou *Applied*, sont souvent en cause. En effet, des suites telles que *Inst Applied* sont fréquentes, puisqu'un grand nombre d'instituts, entre autres entités, sont des instituts de sciences appliquées : nous trouvons dans notre corpus des organisations telles que *Inst Applied Signal Technology* ou *Xi An Inst Applied Optics*. Nous présentons dans le tableau 6.17 page suivante l'exemple de la séquence *Inst Applied Physics Russian Acad Sciences*, testée sur notre corpus de travail.

Il est à noter que cette séquence relève du premier cas que nous avons évoqué, mais également du troisième : des mots grammaticaux ou une ponctuation sont sous-entendus, et la séquence complète aurait correspondu à *Inst of Applied Physics of the Russian Acad of Sciences* ou à *Inst of Applied Physics, Russian Acad of Sciences*, et nous auraient permis d'isoler correctement, lors

sous-séquence	fréquence	surface
Inst Applied	4	8
Inst Applied Physics	1	3
Inst Applied Physics Russian	1	4

TABLE 6.17 – Fréquence et surface des sous-séquences pour la séquence *Inst Applied Physics Russian Acad Sciences*

de la pré-normalisation et avant la suppression des mots grammaticaux, les deux entités *Inst of Applied Physics* et *Russian Acad of Sciences*. Ici, le calcul de fréquence comme celui de surface sont inefficaces : la sous-séquence *Inst Applied* est trop fréquente pour que la suite *Inst Applied Physics* puisse être identifiée comme la plus pertinente. Ce problème est rencontré avec d'autres adjectifs qualifiant l'organisation, comme *Catholique* dans *Univ Catholique Louvain*.

Pour les noms relevant du second cas, puisque le nom de l'organisation principale apparaît le plus souvent avec la même organisation secondaire, les calculs de surface tels que nous les avons mis en place demeurent inefficaces. C'est par exemple le cas des séquences *Univ Colorado Health Sciences Center* et *Harvard Univ Molecular Biol Lab* dans notre corpus. Nous présentons dans le tableau 6.18 les résultats des calculs de surface et de fréquence pour ces noms sur notre corpus de 3 653 noms d'organisations pré-normalisés.

Séquence	sous-séquence	fréquence	surface
Univ Colorado Health Sciences Center	Univ Colorado	3	6
	Univ Colorado Health	2	6
	Univ Colorado Health Sciences	2	8
Harvard Univ Molecular Biol Lab	Univ Molecular	5	10
	Univ Molecular Biol	5	15

TABLE 6.18 – Fréquence et surface des sous-séquences pour les séquences *Univ Colorado Health Sciences Center* et *Harvard Univ Molecular Biol Lab*

Dans le cas de *Univ Colorado Health Sciences Center*, c'est bien *Univ Colorado* qui se dégage sur le calcul de fréquence lui-même. En revanche, c'est *Univ Colorado Health Sciences* qui obtient la surface la plus importante, puisque sur les trois occurrences de *Univ Colorado*, deux apparaissent avec l'organisation secondaire *Health Sciences Center*. Or, il est difficile de déterminer

automatiquement lequel des calculs de surface ou de fréquence est le plus probablement juste, lorsque ces derniers déterminent des frontières différentes.

Pour *Harvard Univ Molecular Biol Lab*, les fréquences de *Univ Molecular* et de *Univ Molecular Biol* sont égales, et sont dans tous les cas des noms d'organisations incorrects. Les calculs de surface donnent aussi une segmentation erronée. Cela est dû au fait que le nom de l'université, soit *Harvard*, est placé avant *Univ*, et n'est donc pas testé. Cependant, une position interne de *Harvard* n'aurait en l'occurrence eu d'impact ni sur les fréquences ni sur les surfaces, puisque les seules occurrences de *Harvard Univ* sont, dans cette étude, suivies du nom de laboratoire *Molecular Biol Lab*.

Ainsi, ces deux calculs permettent dans un grand nombre de cas d'identifier le point de rupture entre une organisation principale et une organisation secondaire. Cependant, certains schémas de noms d'organisations restent problématiques, et ne peuvent être résolus de manière certaine par ce biais. Il est toutefois possible d'envisager un dernier type de calcul statistique, qui prend en compte la relation existant entre les différents mots graphiques d'un nom d'organisation pré-normalisé.

6.1.2.3 L'information mutuelle structurée

Les fréquences permettent de déterminer la frontière entre deux organisations dans des cas où la structure hiérarchique d'organisations n'est explicitée ni par la ponctuation, ni par la structure syntaxique (voir sous-section 6.1.2.1 page 181). Les surfaces, quant à elles, sont efficaces sur certains de ces noms d'organisations lorsque apparaissent des noms composés (voir sous-section 6.1.2.2 page 185). Cependant, nous venons d'exposer les difficultés qui restent à résoudre, en particulier quand certains mots génériques fréquents influencent le calcul, et/ou quand les deux organisations d'une même séquence apparaissent majoritairement ensemble. Dans ces derniers cas, au-delà de la prise en compte de fréquences de suites de mots, il peut être pertinent d'utiliser la fréquence conjointe de mots graphiques fonctionnant préférentiellement ensemble.

Hypothèse Pour dégager la frontière entre deux noms d'organisations, nous pouvons en effet considérer le degré d'association existant entre deux ou plusieurs mots graphiques, plutôt que leur fréquence ou leur surface brute.

Notre hypothèse générale est donc la suivante :

Hypothèse II.5. *Les mots graphiques fortement associés dans les noms d'organisations d'un ensemble documentaire donné sont susceptibles de former les noms d'organisations les plus cohérents. Par conséquent, mesurer la force de cette association peut indiquer la limite entre le nom d'une organisation principale et le nom de l'organisation secondaire.*

L'information mutuelle est une mesure permettant de mesurer cette force d'association. D'après [Church & Hanks, 1990], et considérant x et y comme deux mots, l'information mutuelle se définit de la manière suivante :

> « [Mutual information] compares the probability of observing x and y together (the joint probability) with the probability of observing x and y independently (chance) ».

Formellement, l'information mutuelle est donc définie comme :

$$IM = log(\frac{p(x,y)}{p(x)p(y)}) \tag{6.2}$$

De manière générale, l'information mutuelle est utilisée pour mesurer l'association entre deux mots, par exemple pour savoir si ces deux mots forment une expression figée, un mot composé, etc. Dans le cadre de nos travaux, nous postulons donc qu'établir le degré d'association entre différents mots permettra de déterminer les deux noms d'organisations en présence dans un même nom normalisé. La question est alors de savoir comment procéder à ce calcul, et en particulier, sur quelles associations potentielles travailler en priorité. Par exemple, pour la séquence *Inst Applied Physics Russian Acad Sciences*, il convient de déterminer quels couples de mots feront l'objet du calcul. Ce choix n'est pas trivial : dès lors que nous mettons en place des tests de ce type, et puisque l'information mutuelle fonctionne par couples de mots, le choix des membres du couple influence la qualité des résultats que nous en tirons. Pour notre exemple, nous pourrions calculer l'information mutuelle du couple [*Inst, Applied*], puis celle de [*Applied, Physics*], puis celle de [*Physics, Russian*], etc., suivant en cela l'ordre linéaire de l'écriture, correspondant à l'ordre dans lequel tout énoncé écrit est produit.

Cependant, mesurer l'association de couples comme [*Physics, Russian*] peut sembler peu pertinent : chacun de ces deux mots, dans le contexte syntaxique de notre exemple, est lié à l'un des identifiants en présence, à savoir, respectivement, *Inst* et *Acad*. Or, à y regarder de plus près, c'est le cas de tous les mots placés entre deux identifiants : chacun d'entre eux dépend prioritairement de l'un ou l'autre. Le degré d'association entre ces mots placés au centre semble donc secondaire.

Les identifiants *Inst* et *Acad*, et de manière générale, les identifiants placés dans des noms pré-normalisés à entités multiples, semblent donc jouer le rôle de ce qu'[Abney, 1991] nomme, dans sa théorie des chunks, une « tête » (*head*). Selon l'auteur, un chunk est un constituant minimal et non-récursif. Chaque constituant inclut une tête et une constellation, cette dernière étant composée de mots « attirés » par la tête. Une constellation est constituée de mots grammaticaux ou d'adjectifs. Les chunks peuvent ensuite être rattachés les uns aux autres, pour former des structures de plus en plus complexes. Cette théorie permet à l'auteur de segmenter les phrases syntaxiquement.

[Abney, 1991] se fonde pour sa théorie sur la prosodie des phrases. Il débute son article par cette phrase :

> « I begin with an intuition : when I read a sentence, I read it a chunk at a time. For example, the previous sentence breaks up something like this :
>
> [I begin] [with an intuition] : [when I read] [a sentence], [I read it] [a chunk] [at a time].
>
> These chunks correspond in some way to prosodic patterns. »

Selon lui, les *chunks* syntaxiques sont donc avant tout fondés sur les pauses effectuées à l'oral dans la lecture ou la production d'une phrase. Or, si nous lisons *Inst Applied Physics Russian Acad Sciences* à voix haute, nous plaçons une pause entre *Physics* et *Russian*. Ainsi, si les mots outils qui auraient dû permettre le découpage de la séquence sont absents, c'est une pause prosodique qui permet, à l'oral, de restituer cette frontière. Les deux chunks composant ce nom seraient donc *Inst Applied Physics* d'une part, et *Russian Acad Sciences* d'autre part. De même, pour *Harvard Univ Molecular Biol Lab*, une pause existe entre *Univ* et *Molecular*.

Néanmoins, dans ces séquences, les mots placés entre les deux identifiants ne sont pas que des adjectifs, puisqu'il y a aussi des noms. Pourtant, ils jouent tous le rôle, du point de vue de la syntaxe, de modifieurs de noms, qui est une fonction syntaxique souvent occupée par des adjectifs. De plus, nous avons signalé plus haut (voir la section 6.1.2 page 177) qu'en anglais, il n'était pas rare de rencontrer des noms adjectivaux, c'est-à-dire des substantifs jouant le même rôle en anglais que des adjectifs qualificatifs pour le français [Maniez, 2001]. Le problème posé par ces noms adjectivés est donc comparable à celui du rattachement adjectival, et c'est particulièrement le cas de *Biol*, abréviation du nom *Biology*, qui joue le rôle de modifieur du nom *Lab* dans *Harvard Univ Molecular Biol Lab*. Enfin, [Abney, 1991] se place dans la perspective de structures phrastiques verbales ; or, les séquences que nous avons à traiter ici sont des structures non verbales, et les chunks peuvent ne pas avoir la même forme dans ces structures très différentes, en l'occurrence pour le traitement d'entités nommées.

Ainsi, si la théorie des chunks ne correspond pas strictement à nos données, nous partageons ses objectifs et certaines de ses caractéristiques : nous cherchons, comme [Abney, 1991], à segmenter de manière cohérente des suites de mots contiguës, même si les nôtres ne sont pas à proprement parler des phrases. De plus, dans les deux cas, des têtes se dégagent nettement, et autour d'elles sont placés des mots liés syntaxiquement à elles.

Dans le cas de *Harvard Univ Molecular Biol Lab*, puisque le mot *Harvard* est placé avant la tête *Univ*, nous considérons qu'il lui est rattaché, et donc qu'il n'est pas ambigu. C'est en effet une constante dans nos données.

De même, pour *Inst Applied Physics Russian Acad Sciences*, *Sciences* est directement rattaché à *Acad*.

Nous partons donc du principe que la théorie des chunks est applicable à nos données, et formulons pour nos travaux une deuxième hypothèse :

Hypothèse II.6. *Un mot appartenant à la constellation d'une tête donnée devrait apparaître plus souvent avec cette tête qu'avec d'autres mots, en particulier avec d'autres têtes. L'information mutuelle entre un mot de la constellation d'une tête et cette tête est donc plus forte que celle existant entre ce même mot et une autre tête.*

Selon [Church & Hanks, 1990], l'information mutuelle, en tant que « probabilité dépendant de plusieurs variables » (*joint probability*), est symétrique. L'information mutuelle du couple $[x, y]$ est donc égale à l'information mutuelle du couple $[y, x]$. Soit, formellement :

$$IM(x, y) = IM(y, x) \tag{6.3}$$

Les auteurs distinguent cette information mutuelle du ratio d'association, qui lui n'est pas symétrique, puisque l'ordre des mots est pris en compte dans ce calcul. La fenêtre moyenne qu'ils utilisent pour leur calcul est de cinq mots, car elle est assez large pour détecter des contraintes entre des verbes et leurs arguments, et à la fois suffisamment restreinte pour ne pas biaiser les contraintes de proximité de certaines associations.

Pour notre part, la totalité de la séquence, c'est-à-dire du nom pré-normalisé, constitue notre fenêtre : de manière générale, ce sont des séquences relativement courtes, et quoi qu'il en soit de la longueur, nous savons que dans la plupart des cas, tous les mots présents sont associés à l'un ou à l'autre des identifiants. Par ailleurs, nous n'imposons pas d'ordre à l'association de deux mots : notre calcul relève donc bien de l'information mutuelle et non du ratio d'association. En effet, étant donnée la variation parfois présente sur la structure syntaxique des noms d'organisations, et en particulier sur la qualification des noms en anglais, l'association entre une tête et un mot de sa constellation peut s'effectuer dans les deux sens. Par exemple, pour la séquence *Inst Applied Physics Russian Acad Sciences*, *Inst* et *Physics*, nous pouvons trouver alternativement dans un corpus *Physics Inst* ou *Inst Physics*. Or, il convient de prendre en compte dans le calcul ces deux types d'occurrences, quel que soit le sens de l'association.

Nous avons donc mis en place, pour le découpage des noms d'organisations dans lesquels la structure syntaxique ne permet pas de segmentation fiable, et en complément des calculs de fréquence et de surface, un calcul d'information mutuelle structurée à partir de chunks, ou tout au moins de ce que nous considérons comme chunks dans le cas particulier des entités nommées.

Application Ce calcul d'information mutuelle structurée est réalisé comme un calcul d'information mutuelle classique. La seule différence réside dans le choix des mots sélectionnés pour le couple à tester. Pour le nom *Inst Applied Physics Russian Acad Sciences*, les opérations de

calculs sont représentés sur la figure 6.5 page ci-contre.

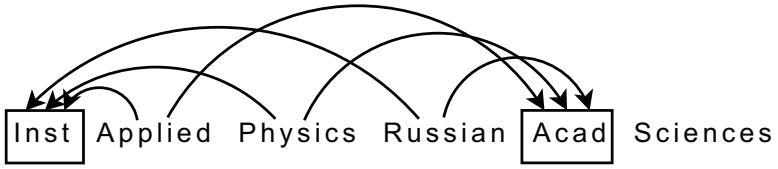

FIGURE 6.5 – Répartition des opérations de calculs d'information mutuelle structurée sur l'exemple de *Inst Applied Physics Russian Acad Sciences*

Chacun des calculs prend donc en compte un mot appartenant à une constellation, que nous ne connaissons pas à l'avance, et l'une des deux têtes auxquelles il est potentiellement rattaché. Par exemple, *Physics*, pour lequel le lien de rattachement n'a pu être établi par les règles de pré-normalisation, est utilisé pour calculer d'une part l'information mutuelle qu'il entretient avec *Inst*, et d'autre part celle qu'il entretient avec *Acad*.

La première étape de l'algorithme est de calculer la fréquence totale de chacun des membres du couple dans le corpus. Dans un second temps, les apparitions conjointes des deux éléments sont à leur tour recensées. Comme chez [Church & Hanks, 1990], les apparitions conjointes ne sont pas forcément des apparitions contigües : *Inst* et *Physics*, ou *Physics* et *Acad*, n'ont pas à être côte à côte pour être considérés comme des mots associés.

Enfin, le nombre d'apparitions conjointes des deux mots est mis en rapport avec le produit du nombre d'apparitions isolées du premier mot et celui du deuxième mot. Le log en base 2 du résultat de ce calcul est l'information mutuelle des deux mots du couple traité.

Les résultats pour la *Inst Applied Physics Russian Acad Sciences* séquence sont présentés dans le tableau 6.19.

L'information mutuelle structurée rattache chaque mot de constellation à la tête avec laquelle elle a l'information mutuelle la plus forte. Ces calculs rattachent d'une part *Applied* et *Physics* à *Inst*, et d'autre part *Russian* à *Acad*. La frontière est donc placée entre *Physics* et *Russian*, et les noms d'organisations *Inst Applied Physics* d'une part et *Russian Acad Sciences* d'autre part sont correctement distingués.

Validation Afin d'évaluer la qualité de ce calcul d'information mutuelle (IM) structurée, et plus précisément pour déterminer son utilité pour notre objectif, nous avons pris en compte notre corpus de test de 3 653 organisations.

Sur les 40 noms d'organisations distincts à entités multiples relevés, l'IM permet de placer la frontière correctement entre deux entités dans 16 cas, soit 40%. Ramené au nombre total d'occurrences, incluant donc les répétitions, ce nombre est de 21/54, c'est-à-dire 38,9%. Ce nombre est relativement faible, en particulier en comparaison avec les autres mesures que sont les fré-

Item de constellation	Tête de rattachement	Paire testée	Information mutuelle
Applied	**Inst**	**Inst, Applied**	**-8,599**
	Acad	Acad, Applied	-9,266
Physics	**Inst**	**Inst, Physics**	**-8,895**
	Acad	Acad, Physics	-9,436
Russian	Inst	Inst, Russian	-10,506
	Acad	**Acad, Russian**	**-6,266**

TABLE 6.19 – Information mutuelle structurée des items ambigus de la séquence *Inst Applied Physics Russian Acad Sciences*

quences et les surfaces. Cependant, l'information mutuelle est pertinente sur des cas non résolus par les deux premiers traitements dans 7 cas distincts, soit dans 43,75% des cas où le calcul d'IM structurée segmente correctement les noms.

Ainsi, la méthode de l'information mutuelle structurée permet de pallier certaines faiblesses des calculs de fréquence et de surface. Elle parvient par exemple à dégager des structures correctes dans des cas où les fréquences et surfaces de certaines sous-séquences non pertinentes sont trop élevées pour laisser émerger la structure pertinente, notamment lors de l'utilisation de mots très fréquents dans le corpus. Ainsi, pour *Inst Applied Physics Russian Acad Sciences*, l'information mutuelle permet de dépasser le problème causé par la fréquence élevée de *Inst Applied* par rapport à la séquence correcte, soit *Inst Applied Physics*, puisque l'information mutuelle de la paire [*Inst, Physics*] est plus forte que celle de [*Acad, Physics*]. *Physics* peut donc être rattaché à la bonne tête.

Pour le nom *Harvard Univ Molecular Biol Lab*, les fréquences et surfaces sont peu adaptées : d'une part, la récurrence des sous-séquences est trop faible dans le corpus, et d'autre part, toutes les séquences sélectionnées sont erronées. Dans ce cas, l'information mutuelle structurée permet de dépasser cette linéarité des calculs précédents. L'IM entre [*Lab, Molecular*] et [*Lab, Biol*] est en effet plus forte que celle entre [*Univ, Molecular*] et [*Univ, Biol*], et les séquences dégagées sont donc *Harvard Univ* d'une part et *Molecular Biol Lab* d'autre part.

Ainsi, en faisant varier l'unité d'analyse, et en tirant parti de cette variété, il est possible de mettre au jour des noms d'organisations issus de différents types de noms.

Parmi les cas restant problématiques, la configuration la plus représentée est celle où l'écart de fréquence entre les deux identifiants pour le corpus est important. Parmi les cas mal découpés

pour cette raison, la majorité des échecs du calcul d'IM structurée est à imputer à la fréquence massive de l'identifiant *Univ* par rapport aux autres : sur les 24 échecs de segmentation par IM des organisations distinctes, 16 contiennent un identifiant *Univ*, soit 66% des mauvais découpages ; sur les 41 échecs erronés sur la totalité des occurrences, 25 sont imputables à la présence de ce même identifiant, c'est-à-dire près de 61%.

En effet, l'identifiant *Univ* apparaît à 1 277 reprises dans le corpus des 3 653 organisations ; des identifiants comme *Inst* ou *Ctr*, pourtant relativement fréquents, n'apparaissent respectivement que 291 fois et 125 fois. Certains sont encore moins représentés, comme *FDN* qui ne compte que 19 occurrences. Ainsi, dans ces cas-là, le calcul d'IM structurée peut être déséquilibré par la sur-représentation des universités, puisque cela implique que les mots des constellations ont dès le départ un lien affaibli avec l'identifiant *Univ*, même s'ils lui sont en réalité attachés, puisque sa distribution est particulièrement importante. Par conséquent, le lien peut être plus fort avec l'autre identifiant si la fréquence de ce dernier est plus faible, sans que cela ait réellement un sens dans le corpus.

C'est pourquoi des séquences comme *Univ Colorado Health Sciences Ctr*, ou *Univ Florida Research Fdn*, ne peuvent être segmentées grâce à l'information mutuelle.

6.1.3 Synthèse : combiner des calculs ne nécessitant pas le développement de ressources pour l'aide à la décision

Nous avons exposé les calculs endogènes que nous utilisons pour traiter et normaliser les entités nommées. D'un côté, la distance d'édition de Levenshtein permet de suggérer des corrections sur des erreurs typo-orthographiques qui ont eu lieu lors de la saisie des informations. De l'autre, les calculs de fréquence et de surface de séquences, ainsi que le calcul d'information mutuelle structurée, proposent des découpages pour des noms d'organisations dont les frontières ne sont pas claires. Nous rassemblons ces trois derniers traitements sous le terme de calculs de segmentation.

Dans tous les cas, aucune ressource externe n'a eu à être développée, si ce n'est une liste des identifiants d'environ 140 items. Le calcul s'effectue sur une étude donnée, et non sur la totalité de la base, puisque l'ensemble documentaire associé à une étude est considéré comme cohérent et unitaire.

La distance de Levenshtein se suffit à elle-même, non pour la correction automatique, mais pour l'appariement de couples représentant potentiellement des variantes d'un même nom d'organisation. Elle ne peut trancher sur le nom le plus probablement correct, mais rassemble les noms variants par paires.

Par la suite, les résultats de ce calcul, et donc ces couples susceptibles d'être des variantes d'un même nom, sont proposés en suggestion de corrections à l'utilisateur, en vue de traitements

anthropogènes (chapitre 8 page 317).

En revanche, nous avons montré que chaque calcul de segmentation, pris isolément, présentait des faiblesses. C'est pourquoi nous exploitons ces trois traitements, afin d'obtenir de meilleures chances de trouver le bon découpage sur une séquence à entités nommées multiples.

Nous avons évalué les propositions obtenues sur notre corpus de 3 653 organisations. Nous avons considéré que si au moins deux sur trois des méthodes proposaient le découpage correct, alors la segmentation était correcte. Dans le cas contraire, c'est-à-dire si tous les calculs renvoient un découpage erroné d'un nom, ou que seul l'un d'entre eux permet de découper correctement la séquence, la segmentation majoritaire est incorrecte.

Nous présentons les résultats de cette évaluation manuelle dans le tableau 6.20 page suivante. Nous rappelons que le système a détecté 40 noms d'organisations pré-normalisés à entités multiples distincts, et que ramené au nombre total d'occurrences, ce nombre monte à 54 occurrences. Les calculs de proportion dans le tableau sont effectués d'après ces totaux. Les 54 occurrences représentent 1,48% de la totalité des noms d'organisations pré-normalisés du corpus.

Le cumul des traitements de découpage enregistre un taux d'erreur relativement élevé, mais toutefois suffisamment important pour que nous n'ayons pas automatisé totalement la segmentation : étant donné le taux de segmentation incorrecte de 32,5% sur les noms d'organisations distincts, la quantité des erreurs introduites aurait été trop importante sur de grands volumes de données. En envoyant ces suggestions aux utilisateurs, nous évitons une grande partie de ces erreurs, et obtenons ainsi des données fiables. Nous exploitons en cela des ressources anthropogènes.

Comme pour la distance de Levenshtein pour la correction typo-orthographique, les résultats de découpage à l'aide de ces trois calculs sont donc proposés à l'utilisateur.

Combiner ces trois calculs pour générer des propositions présentées à l'analyste permet d'accorder plus d'importance, en termes de poids, aux segmentations suggérées par le plus de calculs. Dans le cas où la même séquence serait découpée de la même manière par les trois traitements, celui-là seul sera suggéré, bien que l'utilisateur puisse saisir lui-même un nouveau découpage si celui-ci ne lui convient pas ; si deux des trois calculs tombent sur le même découpage, c'est ce dernier qui sera proposé en premier à l'utilisateur ; enfin, si chacun des trois calculs proposent une segmentation différente, les trois découpages seront suggérés.

Dans tous les cas, et même si le découpage n'est pas optimal, notre traitement détecte les séquences problématiques et les met en avant pour que l'utilisateur puisse en tenir compte et les corriger. En cela, l'approche que nous avons mise en place est utile, puisqu'elle met en avant des éléments nécessitant une correction.

Par ailleurs, nous avons testé nos calculs sur un corpus de 3 653 noms d'organisations pour des raisons pratiques. Cependant, une étude moyenne est aujourd'hui beaucoup plus importante en termes de nombre de données. Or, nos calculs de segmentation tirent parti de la récurrence ; par

Suggestions de segmentation pertinente pour un nom		Nb de noms distincts		Proportion		Nb total d'occurrences (incluant les répétitions)		Proportion	
Nb de suggestions pertinentes	Segmentation majoritaire pour un nom	Par nb de suggestions pertinentes	Par segmentation majoritaire	Par nb de suggestions pertinentes	Par segmentation majoritaire	Par nb de suggestions pertinentes	Par segmentation majoritaire	Par nb de suggestions pertinentes	Par segmentation majoritaire
0/3	incorrecte	4	13	10%	32,5%	4	18	7,4%	33,3%
1/3		9		22,5%		14		25,9%	
2/3	correcte	20	27	50%	67,5%	28	36	51,9%	66,7%
3/3		7		17,5%		8		14,8%	

TABLE 6.20 – Résultats de l'évaluation des segmentations effectuées par le système par cumul des trois méthodes de découpage

conséquent, plus les données sont nombreuses, plus les chances de tirer de l'information pertinente pour le découpage sont grandes. Nous postulons donc que plus les ensembles documentaires associés aux études sont importants, plus les calculs effectués dessus seront pertinents et efficaces.

Nous aurions pu, pourtant, tenter d'optimiser les résultats à l'aide d'autres traitements endogènes. Cependant, l'un des inconvénients majeurs en aurait été, en plus d'une moindre fiabilité des résultats, une complexité des calculs toujours croissante, puisque nous aurions dû en arriver à des traitements toujours plus fins et coûteux.

En effet, alors qu'appliquer la distance de Levenshtein reste relativement simple, la complexité algorithmique augmente avec le calcul de fréquence et surtout de surface, et encore plus avec le calcul d'information mutuelle structurée. Par conséquent, nous ne poussons pas plus loin les traitements endogènes pour la normalisation, et estimons que nous avons atteint un équilibre entre le coût des calculs et le bénéfice que nous en tirons.

6.1.4 Conclusion

Les traitements endogènes permettent de normaliser les entités nommées, en particulier sur certains types de problèmes engendrant de la variation. Un calcul de distance de Levenshtein modifiée et adaptée à notre contexte permet de détecter, de manière relativement efficace, des variantes d'un même nom d'organisation. Il est utilisé pour résoudre les problèmes de variation typo-orthographique au sens large, incluant les problèmes de fautes de frappe et d'orthographe, de traduction/transcription, certains différences dans les conventions de notation, et même, parfois, de variation syntaxique.

Des calculs endogènes de découpage fournissent des suggestions de segmentation des noms d'organisations à entités multiples. Ces derniers traitements agissent à la fois pour la détection de variantes dues à une synonymie partielle, mais permettent également de structurer hiérarchiquement ces entités multiples les unes par rapport aux autres. Ils exploitent les récurrences d'un corpus donné pour en tirer de l'information, soit par des calculs de fréquence et de surface, soit par des mesures d'information mutuelle structurée.

L'ensemble de ces traitements est donc fondé sur les informations présentes dès le départ dans les données à normaliser. Le plus souvent, plus le volume de données est important, plus les processus sont efficaces.

En revanche, la complexité algorithmique associée à ces calculs va en grandissant au fil des traitements. C'est pourquoi, pour la résolution d'autres types de problèmes, il est préférable d'utiliser d'autres types de ressources.

6.2 Les méthodes endogènes pour la modélisation des connaissances : conception et construction par le corpus de la structure de représentation

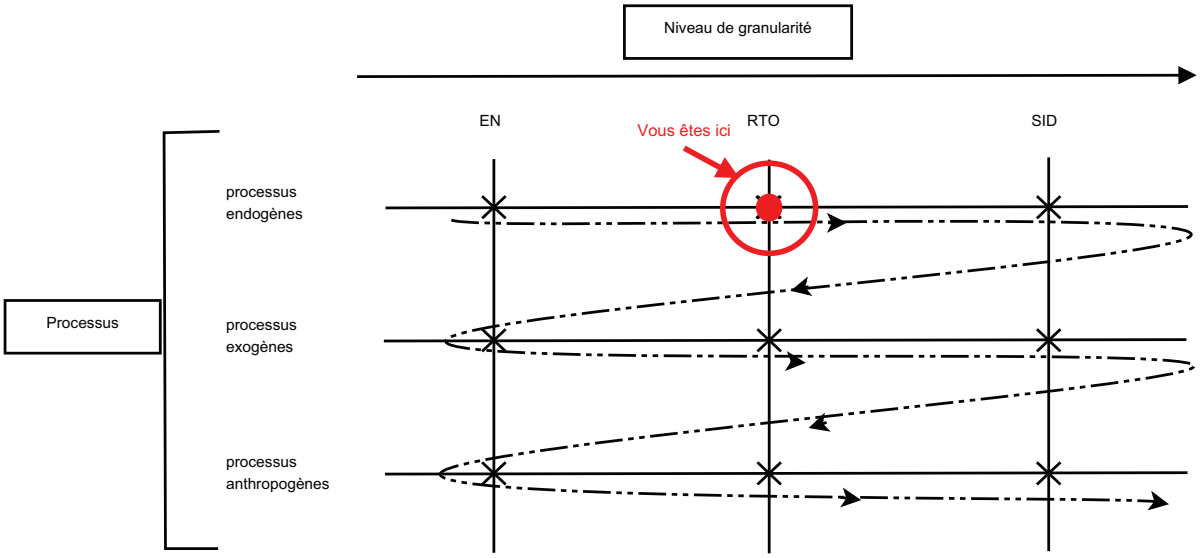

FIGURE 6.6 – Positionnement des traitements endogènes pour la ressource termino-ontologique multi-plans dans l'ensemble des processus

Nous avons établi que les entités nommées et les collocations potentielles représentaient, dans notre modèle, les unités d'informations à exploiter pour la conception d'un système d'immersion documentaire. A ce titre, elles sont les grains les plus fins qui constituent notre système, et représentent donc le premier niveau de granularité de l'analyse. Chaque unité d'information est rattachée à, ou contenue dans un document donné. En tant que vecteur des unités d'informations individuelles, le document est donc un « pont » qui permet de passer des unités d'informations à un palier d'analyse supérieur. En effet, l'accumulation des documents, au sein d'un ensemble documentaire, est une source de croisements potentiels entre les unités d'information. Or, ce sont précisément ces croisements que nous souhaitons rendre possibles, afin d'en faire émerger d'autres informations, plus complexes et plus riches que les types d'informations premiers. Ici, nous abordons les ressources et processus endogènes permettant de construire la ressource termino-ontologique (RTO) qui rassemble ces informations, et qui se place au deuxième niveau de granularité de nos travaux (voir la figure 6.6).

Cette structure de représentation est principalement constituée par rapport à un corpus donné, correspondant à une étude déterminée : la RTO en relation avec une étude est limitée par les informations contenues dans cette étude. Ne sont prises en compte que les données dont la forme apparaît dans le corpus à traiter. Par exemple, une entité d'organisation capitalisée telle

que *Saint Gobain*, qui réfère à une entreprise de matériaux, n'apparaîtra probablement pas dans une étude portant sur la pharmacologie.

Dans ces travaux, nous mettons en avant la combinaison de différents types de traitements pour la modélisation d'un système d'immersion documentaire pertinent. A ce titre, à chaque niveau de granularité, des ressources et processus de différents types rentrent en jeu pour le traitement des données : les ressources endogènes, exogènes et anthropogènes sont souvent imbriquées les unes dans les autres au fil des traitements. Cependant, ainsi que nous l'avons signalé, nous nous en tenons ici exclusivement aux ressources et processus endogènes. Nous revenons sur les autres types d'apports dans les chapitres 7 page 239 et 8 page 317.

Nous partons du postulat selon lequel l'exploitation de ressources endogènes pour la ressource termino-ontologique se montre pertinente, notamment parce qu'elles permettent de rassembler les unités d'informations destinées à faire l'objet de croisements. Nous avons exposé dans le chapitre 3 page 69 l'intérêt que présente l'utilisation de ressources endogènes pour la conception de ressources de ce type : utiliser des données attestées permet de rester au plus près des spécificités de ces données, et permet un retour au texte en tant que source directe de connaissance pour l'utilisateur.

D'autre part, nous avons expliqué qu'il était fondamental, de notre point de vue, de prendre en compte l'objectif de communication des documents à traiter pour parvenir à les traiter correctement (voir section 3.1 page 71). Or, quel meilleur moyen pour cela d'utiliser une ressource fondée sur l'exploitation de ces documents eux-mêmes ?

Dans un premier temps, nous présentons le modèle que nous avons défini pour notre ressource termino-ontologique (RTO) multi-plans. Puis nous présentons la façon dont nous avons exploité les connaissances contextuelles pour structurer et remplir la RTO. Enfin, nous exposons notre méthode endogène d'extraction des collocations des résumés pour remplir la facette des thèmes.

6.2.1 Modèle de la ressource termino-ontologique multi-plans

Nous avons présenté, dans la sous-section 3.2.2 page 84, le modèle de ressource termino-ontologique que nous avons défini. Cette structure de représentation des connaissances inclut plusieurs sous-structures, autonomes mais non indépendantes, dont les formalisations et le degré de structuration varient. Ainsi, à un type donné de connaissances correspond une « sous-structure ».

Nous avons également vu que les types de connaissances étaient représentés linguistiquement par différents types d'entités nommées, et par les collocations issues des résumés des documents permettant de véhiculer le(s) thème(s) de ces derniers (sections 4.1 page 102 et 4.2 page 114).

Tous ces types de connaissances sont intégrés à la ressource termino-ontologique. Nous présentons en figure 6.7 page 203 le modèle de la RTO multi-plans.

Sur cette figure, les cinq types de connaissances sont représentées, sous la forme de cinq plans, ou facettes. Nous avons décrit la notion de facette en 3.2.2 page 84. Chacune d'elles est liée aux documents par le biais des informations qu'elle comporte, et chaque information est donc reliée au document dont elle a été extraite, par une relation dépendant de son type (voir figure 6.7 page 203).

De plus, ces relations primaires de chaque facette aux documents permet d'établir des relations secondaires entre les facettes elles-mêmes : le document est un pont établissant des relations entre différents types d'informations. Nous présentons l'ensemble de ces relations sur la figure 6.8 page 203.

La facette des dates contient les dates clés et leur sémantique : pour les articles scientifiques, seule la date de publication est répertoriée. En revanche, pour les brevets, plusieurs dates sont à retenir : la date de dépôt de la demande du brevet, la période de validité, ainsi que la date de priorité, sont toutes restituées pour chaque brevet. En effet, toutes ces dates relèvent de l'information stratégique exploitée par les analystes. L'intérêt de cette facette réside dans l'établissement de liens entre différents types de dates, en plus des liens fixés entre une (des) date(s) et un (des) document(s). Les relations entretenues par la facette temporelle avec les autres facettes sont représentées dans la figure 6.9 page 204.

Les dates elles-mêmes sont liées les unes aux autres, dans le cas des brevets, par des relations sémantiques.

La facette des auteurs est constituée des prénom(s) et nom de chaque auteur ou inventeur d'un article ou brevet. Puisqu'une publication peut émaner de plusieurs individus, un même document peut être lié à plusieurs entités de personnes. Réciproquement, un même individu peut être auteur ou inventeur de plusieurs publications : un même nom peut donc renvoyer à plusieurs documents. Les liens de la facette des auteurs sont présentés dans la figure 6.10.

La facette des lieux, qui contient potentiellement à la fois le pays, la ville et l'adresse d'une organisation propriétaire d'un document, fonctionne sur le même modèle, au moins pour les pays et les villes : un pays et/ou une ville peuvent être liés à plusieurs documents, et plusieurs pays ou villes peuvent être liés à un même document en cas de co-dépôt. Cela peut être également vrai pour les adresses précises, bien que la probabilité de rencontrer une même adresse pour deux organisations distinctes est plus faible. Les unités d'informations dans la facette même sont donc unies par des relations hiérarchiques, entre pays et villes par exemple, mais également paradigmatiques, entre plusieurs villes d'un même pays. De plus, les lieux sont reliés eux aussi aux autres types d'informations à travers les documents (voir figure 6.11 page 205).

Les organisations sont elles aussi intégrées à leur facette dédiée. Les éléments stockés dans cette facette sont de deux types : les entités d'organisations, agencées de manière hiérarchisée et structurée, sont les concepts de la facette ; les variantes textuelles de ces entités, c'est-à-dire les formes relevées dans les données brutes du corpus, représentent les instanciations des concepts.

Chacun des concepts est lié à l'ensemble de ses variantes, ce qui permet de capitaliser les données. Les variantes entretiennent entre elles des relations de synonymie, tandis que les organisations peuvent être unies par des liens hiérarchiques. Parallèlement, chaque entité d'organisation est reliée aux documents dont elle est propriétaire, et plusieurs organisations peuvent être reliées à un même document, en cas de co-publication ou de co-dépôt. Enfin, la facette entretient des relations avec chacun des autres plans de la ressource (voir figure 6.12 page 205).

Enfin, la facette thématique contient les collocations brutes potentielles tirées des titres et des résumés des documents. Ce plan est le seul à être fondé sur des unités d'informations ne prenant pas la forme linguistique d'entités nommées. Comme pour les autres facettes, chacune des collocations potentielles est liée aux documents dont elle est issue ; réciproquement, un document renvoie à toutes les collocations potentielles qu'il contient. Enfin, ce plan est lié à tous les autres, par le biais des documents communs, par des relations représentées dans la figure 6.13.

Grâce à cette structure en facettes clairement délimitées, les types d'informations pourront être interrogés de manière indépendante, ce qui permettra de les croiser entre eux en fonction des besoins, au lieu de fournir un seul moyen d'interrogation, imposant une représentation linéaire des résultats. Isoler les informations par type permet de les rendre plus accessibles, tout en contrôlant la manière dont il est possible d'y accéder. La RTO multi-plans repose donc sur le principe selon lequel il convient de diviser pour mieux régner.

La répartition en facettes des informations permet la création de trois types de liens, tous exploitables différemment par la suite, bien que simultanément :

- les liens établis directement entre un document et l'item d'un type d'information donné ; ainsi d'un lien entre un document et sa date de publication ;
- les liens établis entre différents items d'un même type, à travers un document ; ainsi des deux auteurs d'un même article scientifique, ou des différentes collocations potentielles d'un même résumé ;
- les liens établis entre différents items de différents types, encore une fois à travers les documents ; ainsi des liens entre une organisation, une date, un lieu et un auteur donnés, représentant l'ensemble des paramètres de la situation de production d'un document.

La création, mais aussi la distinction de ces liens les uns par rapport aux autres, présente l'intérêt majeur de donner la possibilité de typer ces liens par défaut, sans intervention extérieure. La relation existant entre deux éléments est ainsi explicitée par la structure même de la ressource, et est univoque.

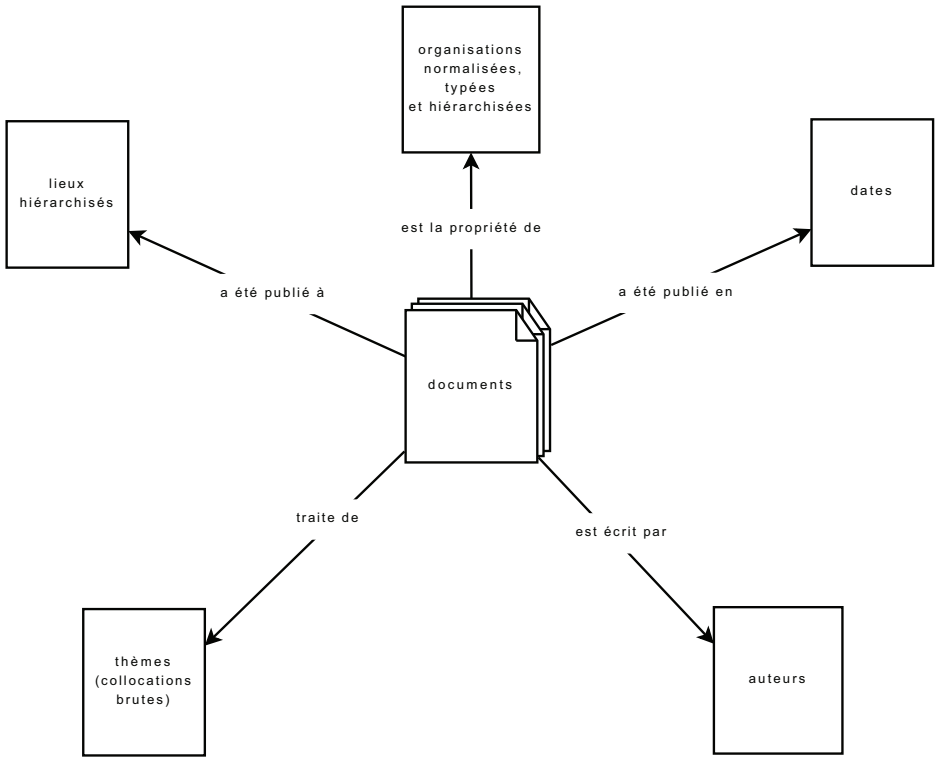

FIGURE 6.7 – Schéma de la ressource termino-ontologique multi-plans

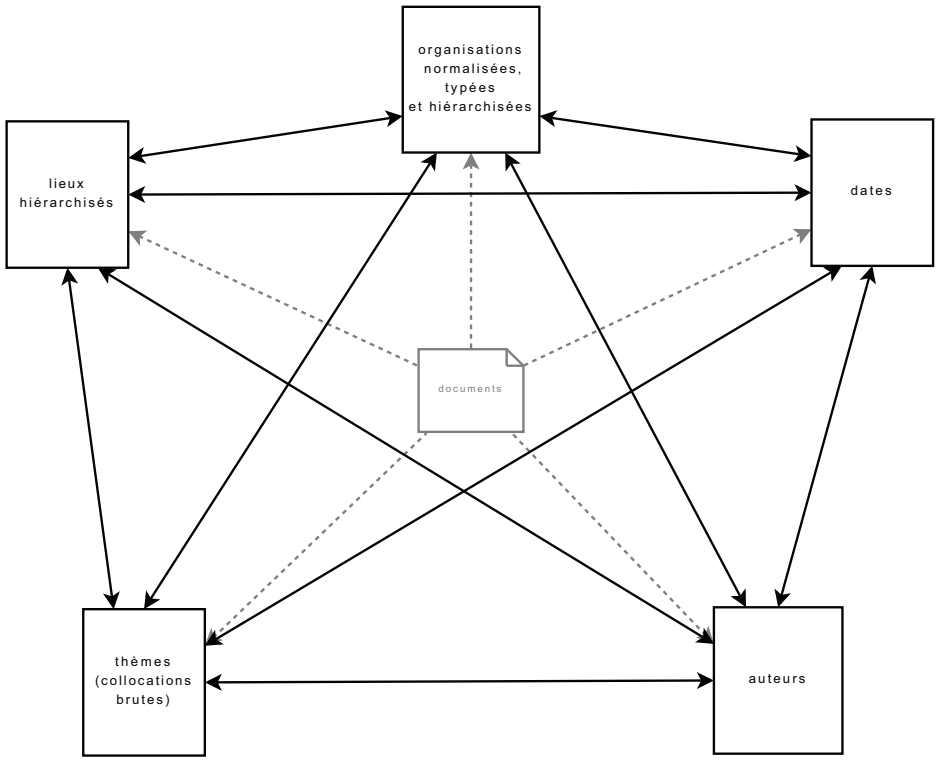

FIGURE 6.8 – Relations entre facettes par le biais des documents

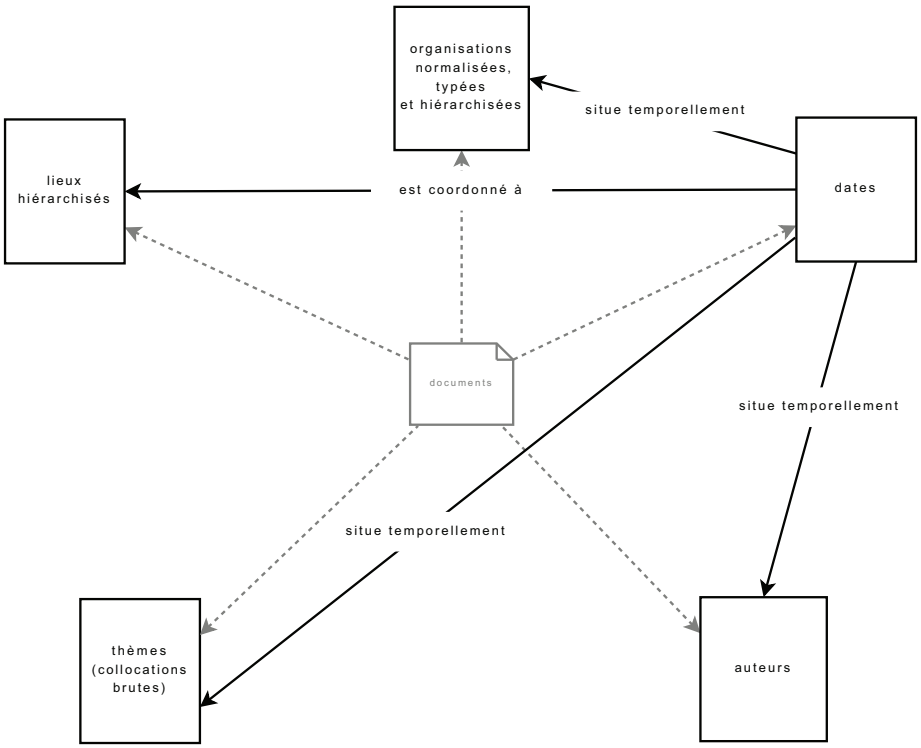

FIGURE 6.9 – Sémantique des relations entre la facette temporelle et les autres plans

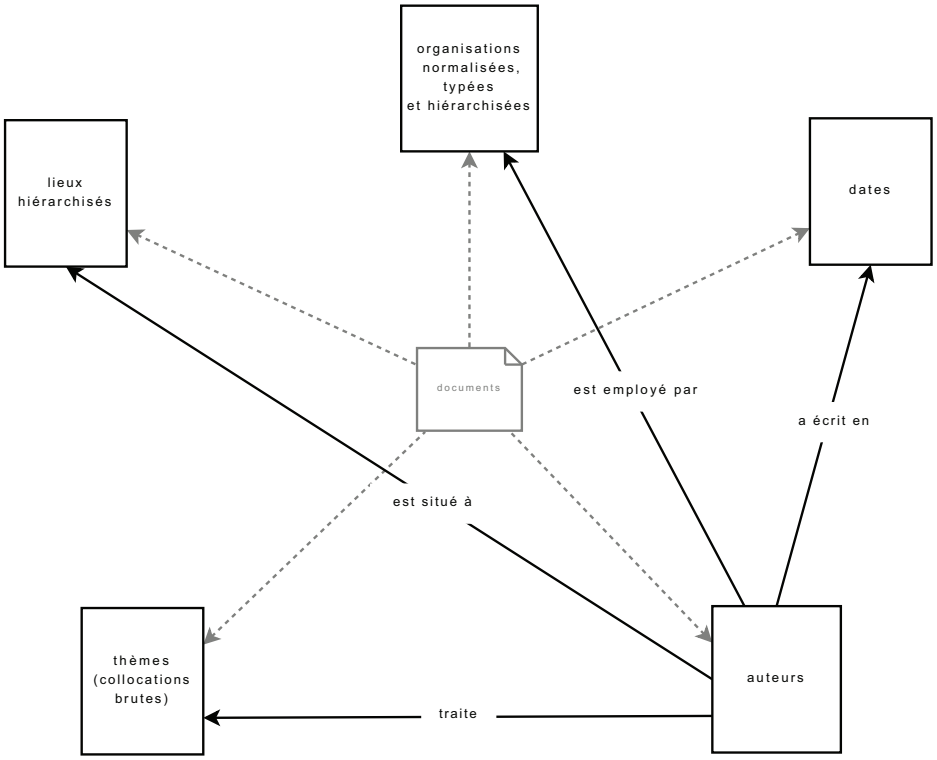

FIGURE 6.10 – Sémantique des relations entre la facette des auteurs et les autres plans

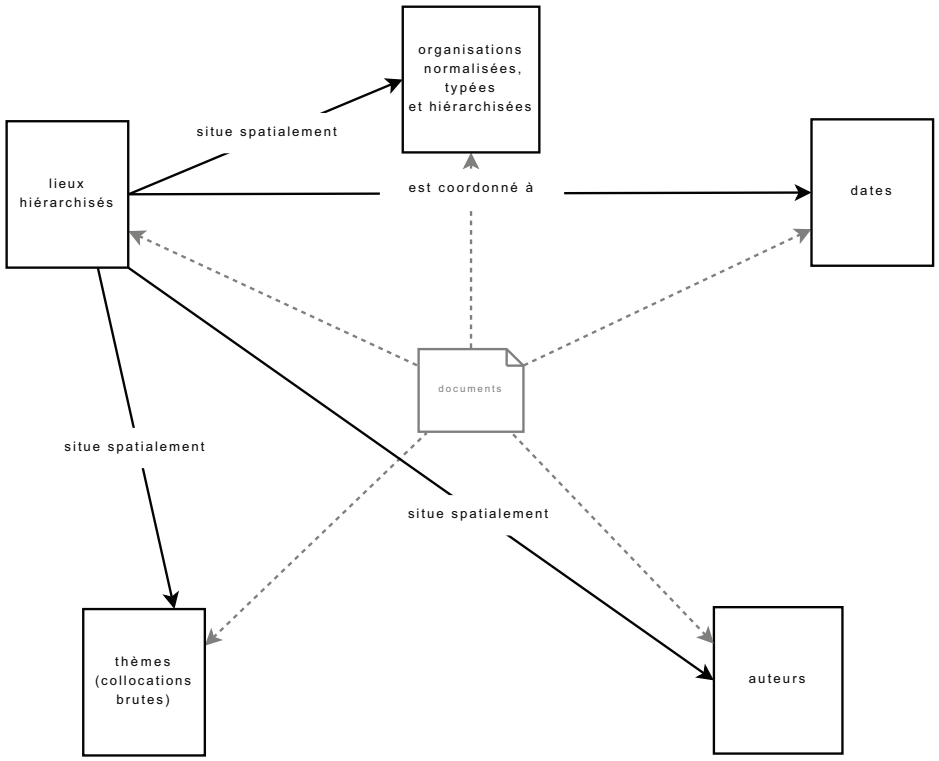

FIGURE 6.11 – Sémantique des relations entre la facette des lieux et les autres plans

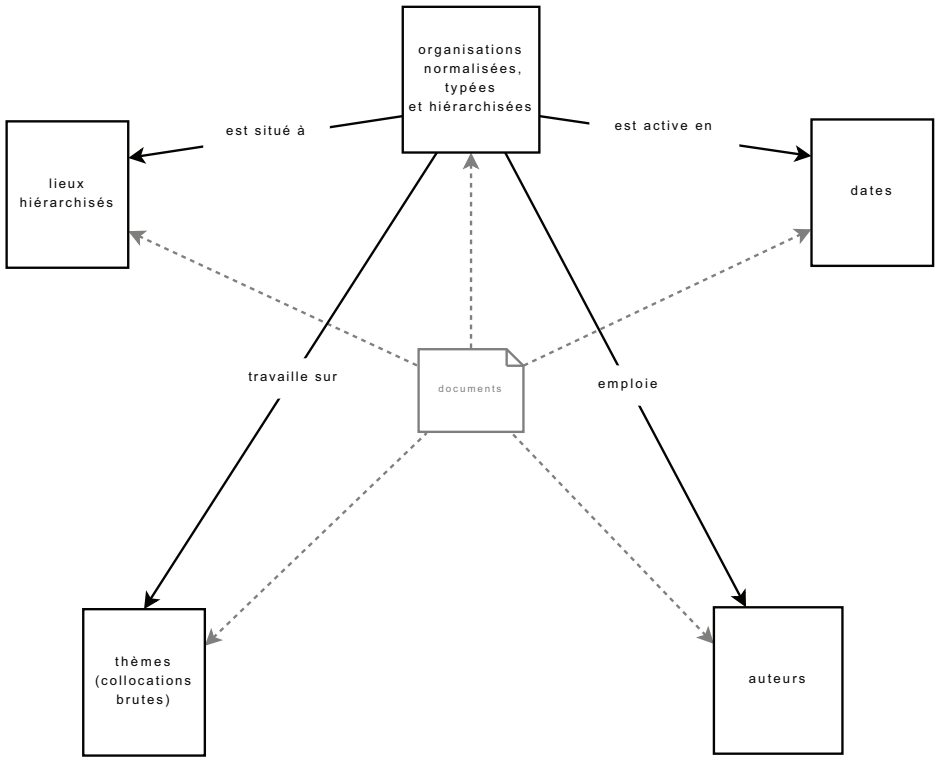

FIGURE 6.12 – Sémantique des relations entre la facette des organisations et les autres plans

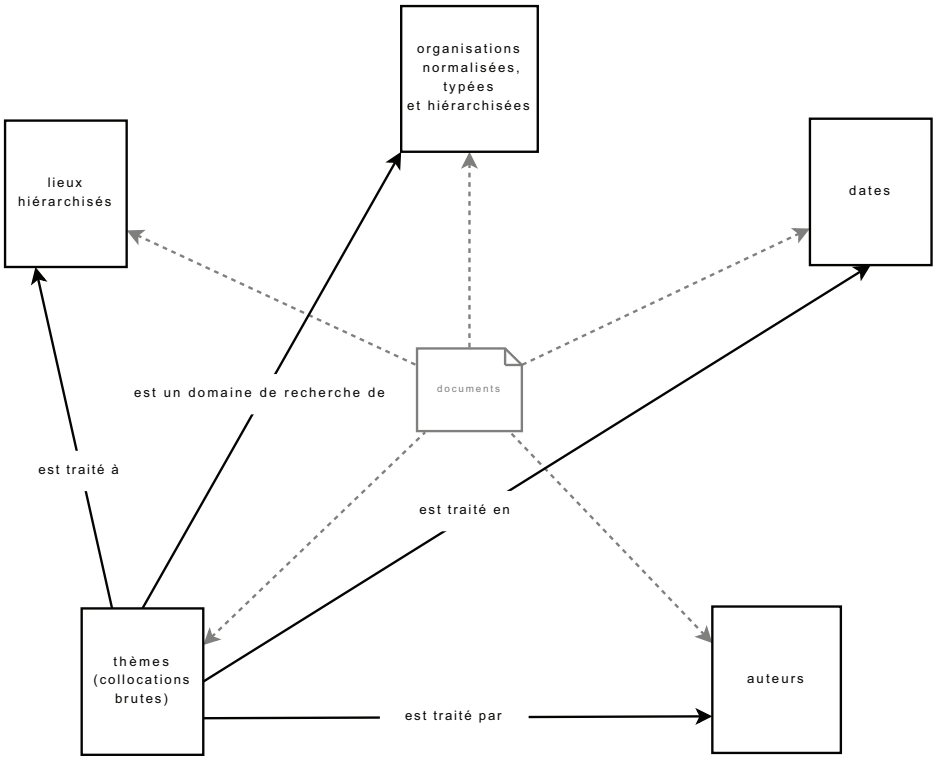

FIGURE 6.13 – Sémantique des relations entre la facette des thèmes et les autres plans

6.2.2 Utilisation des connaissances contextuelles pour la définition de la structure et de son contenu

Nous avons expliqué, dans les chapitres 3 page 69 et 4 page 101, que notre modèle de ressource termino-ontologique (RTO) était fondé en grande partie sur les connaissances contextuelles décrivant la situation de production d'un document. Ces connaissances permettent en effet d'appréhender les documents, et l'information qu'ils véhiculent, d'un point de vue adapté à l'activité de conseil qui est fondée sur l'exploitation de ces documents. En effet, dans le cadre de la société TKM, le document a un double statut de signe et de medium (voir la section 3.1.2 page 76) : son contenu même est une information scientifique et/ou technique, tandis que les informations concernant ses conditions de production ont un rôle pragmatique, permettant en particulier, dans notre cas, d'établir une preuve de propriété d'une invention ou d'un contenu scientifique.

Ces conditions de production sont exprimées textuellement dans les métadonnées associées au document, et plus précisément celles qui sont formulées dans son en-tête. Nous avons précisé (voir la sous-section 4.2.1.2 page 115) que concrètement, lors de l'import de documents dans la base interne de TKM, ces métadonnées étaient isolées de manière plus ou moins fiable en fonction de leur nature, et typées en conséquence, avant d'être intégrées dans la base de données. D'un point de vue linguistique, les métadonnées prennent la forme d'entités nommées, et plus

particulièrement des dates, des entités d'organisations, de personnes et de lieux.

Nous avons par ailleurs établi qu'intégrer ces connaissances dans différentes facettes en fonction de leur type était une manière pertinente de les représenter : cette structure donne la possibilité d'accéder à ces dernières de diverses façons en fonction de besoins découlant d'objectifs variables. Or, la souplesse de l'accès à l'information est une condition clé de l'utilité du système d'immersion.

6.2.2.1 Hypothèse

Dans notre modèle, quatre des facettes de la RTO contiennent des entités nommées : les facettes des organisations, des auteurs, des lieux et des dates. Elles sont majoritairement constituées à partir de ressources endogènes. En effet, elles font directement appel aux données du corpus pour en extraire les informations pertinentes et les intégrer à la ressource. L'utilisation de ces données ne représente toutefois qu'une partie du traitement, puisque d'autres processus peuvent être exploités (voir sections 7.2 page 284 et 8.2 page 343).

Pour trois de ces facettes, nommément celles des organisations, des dates et des auteurs, il aurait été inenvisageable de traiter les données à l'aide de lexiques, qui n'auraient pas été exhaustifs et n'auraient présenté que peu d'intérêt. Pour la facette des lieux, la situation est quelque peu différente (voir la section 7.2 page 284), mais elle reste fondée sur les informations extraites du corpus.

Pour ces facettes, nous formulons l'hypothèse suivante :

Hypothèse II.7. *Utiliser les informations du corpus pour construire et enrichir la ressource permet d'obtenir l'information potentiellement pertinente pour l'utilisateur, et seulement celle-ci.*

6.2.2.2 Application

Nous distinguons les traitements endogènes appliqués aux entités nommées, tirées des métadonnées, de ceux utilisés pour les collocations brutes, extraites des résumés. Si la ressource qui permet de les traiter est la même, c'est-à-dire le corpus, les traitements eux-mêmes sont très différents, en raison de la nature distincte des unités linguistiques en jeu d'une part, et de leur statut - métadonnées ou séquences textuelles, respectivement - d'autre part. Nous nous attachons donc à décrire ici les processus appliqués aux quatre facettes contenant des entités nommées.

Même si certains types d'entités sont, entre le corpus et l'intégration à la ressource, transformées, la source première de toutes les données est bien le corpus dont elles sont issues. Chaque facette intègre des entités, et par là des connaissances, ayant chacune leurs spécificités. C'est pourquoi les méthodes employées diffèrent parfois, en fonction des besoins nés de ces spécificités.

La facette des auteurs, contenant des entités nommées de personnes, est fondée sur l'intégration directe des noms d'auteurs tels qu'ils sont formalisés lors de l'import des données dans la base TKM. Ainsi que nous l'exposons dans le chapitre 7 (voir la section 7.1 page 241), les noms d'auteurs sont traités à partir des noms bruts récupérés sur les bases de données en ligne : ils sont découpés, et les noms de famille sont distingués des prénoms. Ces noms standardisés en amont ne sont pas analysés plus avant, et peuvent donc être intégrés tels quels dans la facette des auteurs d'articles et inventeurs de brevets.

Les dates de publication fournies par le corpus pour chaque document, après leur rapide normalisation, sont également intégrées dans la facette correspondante. Ainsi que nous l'avons signalé plus haut, chaque type de date est associé à sa sémantique, en particulier pour les brevets.

Les lieux permettant de localiser les organisations sont également extraits des données, à partir des noms d'organisations bruts qui contiennent souvent les adresses, plus ou moins précises, de ces entités. Ces éléments de localisation sont distingués en fonction de leur caractère plus ou moins global. Les pays sont donc séparés des villes, qu'ils incluent, et celles-ci sont distinguées des adresses précises. Le lien d'inclusion est restitué dans la facette, et permet de structurer cette dernière.

Enfin, les entités d'organisations sont tirées des données, avant d'être normalisées et structurées. Dans notre modèle, les résultats de cette normalisation sont intégrés à la facette des organisations en tant qu'entités, assimilées aux concepts, tandis que les données brutes représentent les variantes des entités, soit les instanciations des concepts. Par ailleurs, le lien hiérarchique est explicité entre les différentes entités d'une même « super-organisation », ce qui structure la facette de façon hiérarchique.

6.2.2.3 Validation

Ces méthodes endogènes permettent de constituer quatre des cinq facettes de la ressource termino-ontologique (RTO) multi-plans.

Les unités d'information sont ainsi restituées dans la ressource, et intégrées dans la facette qui correspond à leur type. Chaque unité d'information est liée au document dont elle a été tirée. De fait, des liens sont créés entre unités d'informations et documents, mais aussi entre les unités d'informations elles-mêmes, qu'elles soient de même type ou de types différents. C'est ce que nous avons établi ci-dessus.

Le dernier type de lien est celui qui existe entre les documents eux-mêmes, par le biais des liens entre unités d'informations. La ressource termino-ontologique ainsi constituée est donc structurante, puisqu'elle extrait les intersections existantes entre les documents, à travers les unités d'informations et leurs liens. En l'occurrence, puisque les facettes que nous venons de décrire portent sur des entités nommées extraites des métadonnées, elles décrivent la situation

de production des documents d'un corpus. Reprenant l'inventaire des liens existant dans la ressource termino-ontologique, nous y ajoutons donc ce dernier type :

- les liens établis directement entre un document et l'item d'un type d'information donné ; ainsi d'un lien entre un document et sa date de publication ;
- les liens établis entre différents items d'un même type, à travers un document ; ainsi des deux auteurs d'un même article scientifique, ou des différentes collocations potentielles d'un même résumé ;
- les liens établis entre différents items de différents types, encore une fois à travers les documents ; ainsi des liens entre une organisation, une date, un lieu et un auteur donnés, par exemple, représentant l'ensemble des paramètres de la situation de production d'un document ;
- les liens établis entre documents d'un ensemble documentaire, à travers les paramètres de la situation de production qui leur sont communs, exprimés par les unités d'informations.

L'existence de ces liens permet donc de rapprocher potentiellement, selon différents critères, des documents entre eux. Ces intersections structurantes remplissent l'objectif final de la ressource : donner la possibilité de révéler ces liens entre documents afin d'en tirer de l'information pertinente pour les analystes, à travers l'outil d'immersion documentaire.

Puisque la ressource est constituée en facettes, les liens entre documents peuvent être établis selon différents critères, et en particulier selon ceux qui sont pertinents pour l'utilisateur pour une étude donnée.

6.2.3 L'extraction des collocations brutes à partir des documents pour la structuration et l'enrichissement de la facette des thèmes

Nous avons présenté, dans le chapitre 4 page 101, les unités informationnelles prises en compte dans nos travaux. Parmi elles, les thèmes des documents forment l'un des cinq plans de la ressource termino-ontologique. Nous considérons que le thème, au sens que nous avons donné à cette notion, est véhiculé textuellement par des séquences de mots graphiques. Nous estimons que ces séquences de mots graphiques relevaient de la collocation, notion utilisée en linguistique de corpus. La collocations se caractérise notamment par sa récurrence dans une unité textuelle donnée. Certaines méthodes d'extraction ne relevant pas strictement de la collocation, pour la terminologie, la structuration thématique de texte, etc., ne placent pas cette récurrence au premier plan, mais elle existe souvent de fait dans les unités extraites.

Nous présentons plusieurs méthodes et techniques d'extraction qui peuvent être exploitées pour le repérage de ces séquences textuelles, souvent récurrentes, pouvant représenter le thème, même si les objectifs finaux peuvent parfois diverger des nôtres. Nous nous fondons en cela sur l'état de l'art constitué par [Farabet, 2010] lors d'un stage sur le sujet sous notre supervision.

[Condamines & Rebeyrolle, 1997], dans le cadre de la construction de bases de connaissances terminologiques (BCT) à partir de textes, en collaboration avec des entreprises du domaine spatial, ont une approche linguistique de la question de l'extraction terminologique. Nous avons vu, dans la sous-section 4.1.2 page 106, que l'unité terme en tant que telle ne correspondait pas strictement à notre objectif d'extraction thématique. Toutefois, la méthode employée pourrait être adaptée, dans une certaine mesure, à notre application. Les auteurs utilisent des analyses linguistiques de corpus du domaine étudié combinées à différents outils : Lexter, logiciel d'extraction de candidats termes, et Sato, logiciel d'analyse de textes, sont exploités dans la chaîne de traitement pour optimiser les résultats.

Lexter effectue une analyse syntaxique des corpus fournis en entrée grâce à des patrons morpho-syntaxiques, et en extrait des candidats termes. A partir de ces candidats, les auteurs procèdent à une analyse linguistique permettant de déterminer lesquels des candidats sont effectivement des termes. Elles fondent leur analyse sur la détection de relations, identifiées grâce à des marqueurs - marqueurs définitoires, marqueurs de relations méronymiques, etc. - permettant d'établir des relations entre les candidats. Sato est utilisé pour projeter sur des corpus de textes des structures syntactico-sémantiques contenant ces marqueurs. Ces patrons permettent de sélectionner des couples de termes ainsi que les relations qu'ils entretiennent en discours, et de les utiliser pour constituer une base de connaissances. Cette méthode est efficace, mais implique un traitement manuel relativement lourd pour trier les termes et définir les marqueurs de relations.

Un grand nombre de systèmes, à l'aide de patrons définis à la main, repèrent les termes grâce à un automate. Ces systèmes sont étudiés dans [Jacquemin, 2001], cité par [Patry & Langlais, 2005]. Ces patrons utilisent les catégories morpho-syntaxiques obtenues par étiquetage. De fait, cette méthode est soumise aux erreurs de l'étiqueteur.

[Patry & Langlais, 2005], dans le domaine de la gestion de terminologie, nécessaire à la traduction automatique, au résumé automatique ou à l'indexation, adoptent une autre approche. Leur méthode, fondée sur un apprentissage, utilise un corpus d'entraînement annoté par un utilisateur expert du domaine concerné. D'après les auteurs, ce corpus d'apprentissage permet d'éviter les erreurs générées par un étiquetage morpho-syntaxique qui ne peut être entièrement fiable. A partir de ce corpus annoté à la main, le système apprend, puis génère des règles, dans un processus entièrement automatisé. Les patrons générés se voient attribuer un score de probabilité de fiabilité. Ce score est ensuite pondéré par un ensemble de mesures statistiques, telles que la fréquence, la longueur, le tf.idf, etc. Les candidats ayant un score atteignant un seuil déterminé sont considérés comme des termes. Les faiblesses de ce système résident dans le fait que la « variation des termes » n'est pas prise en compte ; de plus, il n'identifie pas les termes simples.

[Bilhaut, 2005] travaille quant à lui sur l'acquisition automatique de connaissances pour l'analyse thématique du discours. Des applications comme la recherche d'information, la navigation

ou le résumé automatique peuvent en tirer parti. Il cherche à constituer automatiquement une ressource terminologique structurée par axes sémantiques nécessaire à l'analyse thématique. Il distingue deux types d'approches pour l'analyse thématique :

- les approches quantitatives, comme celles qui se fondent sur la notion de cohésion lexicale : [Halliday & Hasan, 1976], cités par [Bilhaut, 2005], utilisent la répétition des mots comme un indicateur d'homogénéité thématique ;
- les approches linguistiques, fondées sur la détection de marques discursives, à travers des « critères objectifs et préhensibles », détectables automatiquement.

[Bilhaut, 2005] se place du côté des modèles linguistiques, et réalise une analyse systématique des termes revêtant régulièrement des fonctions discursives caractéristiques. Le but est donc pour lui d'isoler les termes qui structurent le texte, et non d'être exhaustif sur les termes d'un domaine.

Pour la détection des termes et de leurs relations, il réalise un étiquetage morphologique, puis un marquage des syntagmes nominaux et prépositionnels grâce à une grammaire d'unification. A partir des éléments ainsi marqués, il extrait les termes structurants grâce à des critères phrastiques et discursifs, en particulier des critères de positionnement dans la phrase, le paragraphe ou le texte, de présence dans les titres, ou de saillance distributionnelle avec le tf.idf.

Les termes ainsi définis sont alors regroupés en axes sémantiques par les relations qu'ils entretiennent entre eux. Il peut s'agir de relations syntagmatiques, ou de critères phrastiques ou discursifs comme la co-énumérabilité dans toute structure énumérative.

La détection des termes structurants est efficace, mais la répartition en axes n'est pas satisfaisante.

La méthode de [Vergne, 2003] permet de constituer automatiquement des revues de presse à partir des « Une » des sites de presse sur internet. Son objectif est de répondre à la question : « de qui, de quoi est-il question dans la presse ? ».

Il utilise pour cela une méthode endogène, multilingue et sans ressources externes. Un calcul fondé sur la fréquence et la longueur des mots graphiques permettent d'identifier les mots vides dans toute langue utilisant une écriture alphabétique (par opposition aux systèmes idéographiques comme le mandarin) ; à partir de cette identification, des patrons alternant mots vides et pleins détectent les termes dans les « textes » issus du code source des hyperliens qui pointent vers les articles. Son système calcule un graphe de termes, où les nœuds sont les termes, et les arcs les relations entre termes. Les relations sont définies par la co-occurrence de deux termes dans un même « texte » d'hyperlien. L'utilisateur peut alors naviguer dans ce graphe pour accéder aux articles. Dans le cadre de l'application visée, le système est pertinent. Cependant, il ne prend en compte que des textes très courts, puisqu'il s'agit de ceux placés dans les liens de titres.

Dans une autre optique, et avec un objectif autre que la constitution de ressources terminologiques, [Lafon & Salem, 1983] travaillent sur l'inventaire des segments répétés. Nous avons brièvement décrit cette méthode dans la section 4.1.2 page 106. Cet inventaire est utilisé ini-

tialement pour des études lexicométriques. Selon [Lebart & Salem, 1994], la lexicométrie est un ensemble de « méthodes de réorganisation formelle et d'analyse statistique à partir d'une segmentation » d'un texte. Les applications de la lexicométrie sont variées, et peuvent concerner la recherche documentaire. L'inventaire des segments répétés contient tous les segments textuels de n mots graphiques, qui sont répétés un certain nombre de fois dans un corpus donné.

Nous récapitulons l'ensemble de ces approches dans le tableau 6.21 page suivante.

Certaines des approches présentent un inconvénient majeur pour nous, qui est le degré important d'intervention humaine, de la part du linguiste [Condamines & Rebeyrolle, 1997] ou de l'utilisateur [Patry & Langlais, 2005], ou les traitements linguistiques coûteux à mettre en œuvre [Bilhaut, 2005]. Dans le contexte industriel dans lequel nous nous situons, nous avons fait le choix stratégique de mettre en place des traitements légers, et nécessitant une intervention limitée de l'utilisateur au stade de la constitution de la facette thématique. Les traitements endogènes, qui ne nécessitent pas de ressources externes et qui utilisent directement les données du corpus, nous paraissent donc les plus pertinents. Les méthodes de [Vergne, 2003] et de [Lafon & Salem, 1983], en particulier, font état de traitements exploitables pour nos données et notre objectif.

Toutefois, nous mettons de côté la méthode de [Vergne, 2003], car son application à des textes plus longs, c'est-à-dire des résumés d'articles ou de brevets, impliquerait une complexité de traitement importante. En revanche, l'inventaire des segments répétés de [Lafon & Salem, 1983] est adaptable pour nos traitements, c'est-à-dire pour l'extraction des collocations brutes de ces résumés. Nous présentons donc de manière plus complète leur méthode dans la sous-section suivante.

Former une facette thématique, composée de collocations détectées à l'aide de la méthode des segments répétés, pourrait être vu comme une indexation sur le mode de celle qui est faite par les moteurs de recherche tels que Google.

Cependant, nous voyons dans cette facette et son contenu, plus qu'une simple indexation, une manière de structurer un ensemble documentaire par les thèmes qui y sont exprimés. Ce plan thématique établit, comme les plans décrivant la situation de production des documents, des liens entre les unités d'information thématiques et les documents, entre les unités thématiques elles-mêmes, et entre les unités thématiques et les autres types d'unités d'informations.

6.2.3.1 Hypothèse sur l'extraction endogène des collocations brutes

Rappelons que nous entendons la notion de collocation au sens utilisé en linguistique de corpus : la caractéristique première de la collocation est, dans ce cadre, sa récurrence, sa régularité dans un ensemble textuel donné. L'aspect statistique est donc fondamental, contrairement à l'acception usitée en phraséologie, où la collocation est avant tout une co-occurrence obéissant à une structure syntaxique significative, et lexicalement restreinte par la norme en usage. Dans

Auteur(s)	Méthode/ Système	Caractéristiques	Application	Avantages	Inconvénients
[Condamines et Reyberolle, 1997]	Extraction terminologique par analyse linguistique à partir de textes	- Utilisation de Lexter pour extraction des candidats termes par analyse syntaxique. - Analyse linguistique manuelle pour détection de marqueurs. - Utilisation de Sato pour projection de structures de marqueurs.	Constitution de bases de connaissances terminologiques pour entreprises du spatial	Extraction de termes et de relations spécifiques au domaine	Beaucoup d'intervention manuelle
[Jacquemin, 2001]	Recensement de méthodes d'extraction de termes par patrons écrits à la main	- Etiquetage morpho-syntaxique. - Ecriture de patrons fondés sur l'étiquetage.	Variées en terminologie	Problèmes dûs à la variation limités l'étiquetage	Les erreurs d'étiquetage se répercutent sur la détection des termes
[Patry et Langlais, 2005]	Apprentissage automatique	- Corpus d'entraînement annoté par expert du domaine. - Apprentissage et génération de règles. - Pondération statistique.	Gestion de terminologie pour traduction, résumé, indexation, etc.	Elimination des erreurs dues à un mauvais étiquetage morpho-syntaxique	Constitution d'un corpus annoté par l'utilisateur: coûteux
[Bilhaut, 2005]	Constitution automatique de ressource terminologique structurée par axes sémantiques	- Etiquetage morphologique - Analyse syntaxique - Détection des termes structurants et de leurs relations par critères discursifs.	Analyse thématique du discours pour la recherche d'information, navigation, résumé automatique, etc.	Accès au texte par plusieurs points de vue (axes sémantiques)	Traitements linguistiques importants
[Vergne, 2005]	Constitution endogène de graphes de termes	- Identification endogène des mots vides - Détection des termes à l'aide de combinaisons mots vides/mots pleins - Relations établies par co-occurrence	Revue de presse à partir de sites internet	Pas de ressources externes Traitement de données multilingues	Traitement coûteux
[Lafon et Salem, 1983]	Inventaire des segments répétés	Sélection des séquences de *n* mots répétées dans un corpus	Lexicométrie, exploitable en recherche documentaire	Pas de ressources externes	Mots simples non pris en compte

TABLE 6.21 – Récapitulatif des approches liées à l'extraction de connaissances ou de terminologie

cette approche, une collocation peut être acceptable ou non, ce caractère étant déterminé par introspection.

Le caractère quantitatif et contextualiste de l'approche adoptée en linguistique de corpus est plus adaptée à nos travaux que l'approche phraséologique formelle. En effet, puisque les données textuelles à traiter sont issues des documents de nos ensembles documentaires, elles sont par définition attestées. Ce sont ces données que nous devons analyser pour en tirer le thème de ces documents, et par conséquent, une approche de linguistique de corpus est la plus porteuse de sens pour nous. Nous partons en effet du principe que les collocations présentes dans un corpus ou sous-corpus donné, en tant qu'elles sont significatives pour ce corpus ou sous-corpus, sont représentatives du thème des documents en jeu.

Cependant, les unités que nous envisageons d'exploiter ne sont pas strictement des collocations. En effet, la collocation est restituée sous une forme lemmatisée, en tant que structure lexicale significative dans une langue en usage donnée. Or, les formes que nous souhaitons extraire ne sont pas lemmatisées, et représentent en quelque sorte la première étape dans l'extraction des collocations. Nous nommons donc nos unités, à partir de maintenant, des collocations brutes. En cela, elles se rapprochent des segments répétés. Cependant, nous nous écartons aussi des segments répétés tels que définis par [Lafon & Salem, 1983], en ce que certains de ces segments ne sont pas, de notre point de vue et compte tenu de nos objectifs, pertinents. Les segments commençant et/ou finissant par des mots grammaticaux, en particulier, ne présentent que peu d'intérêt pour la détection des thèmes d'un document.

Ainsi, nos unités thématiques se situent à l'intersection des collocations et des segments répétés. Nous les nommons, à partir de maintenant, des collocations brutes.

Nous partons de l'hypothèse suivante :

Hypothèse II.8. *Le repérage des collocations brutes pour un ensemble documentaire donné permet de déterminer les thèmes de cet ensemble documentaire.*

Pour extraire les collocations brutes, nous avons fait le choix de nous appuyer, nous l'avons vu, sur la méthode des segments répétés de [Lafon & Salem, 1983]. Cette approche des segments répétés est adaptée à notre objectif, moyennant quelques modifications. Elle permet en effet de dégager les collocations brutes, c'est-à-dire non lemmatisées, et par là, les potentiels thèmes et sous-thèmes des documents à traiter.

Ainsi que nous l'avons expliqué en 4.1.2.2 page 110, une collocation est, globalement, une co-occurrence de plusieurs mots graphiques récurrente dans un corpus donné. Extraire des collocations de ce corpus consiste donc :

1. à détecter les co-occurrences ;
2. à mesurer la fréquence d'une même co-occurrence au sein de l'ensemble souhaité : le texte, une partie du corpus, et/ou la totalité du corpus.

Pour ce faire, et répondre aux deux exigences de la définition des collocations, la méthode des segments répétés peut être appliquée. Cette méthode issue des statistiques textuelles, développée notamment par [Lafon & Salem, 1983], utilise la récurrence de segments pour l'analyse lexicométrique. A l'époque, elle est une alternative inédite à l'indexation et l'analyse de textes par mots graphiques. Les auteurs proposent en effet d'extraire et de considérer, sur l'ensemble d'un corpus, non plus les mots simples, mais les segments, soit toutes les suites de « formes graphiques non séparées par un séparateur de séquence » (*ibid.*).

A partir de là, seules sont prises en compte les séquences apparaissant au moins n fois dans le corpus. Cette méthode leur permet de réaliser des analyses lexicométriques plus fines et plus efficaces que celles réalisées sur des formes simples, même à l'aide de concordanciers.

Globalement, cette méthode correspond donc aux deux points nécessaires à l'extraction des collocations que nous avons mentionnés ci-dessus : elle permet de détecter les co-occurrences, et de ne prendre en compte que celles qui se répètent dans un ensemble textuel donné.

[Lafon & Salem, 1983] signalent que les segments répétés qu'ils ont étudiés sont d'une longueur comprise entre deux et cinq mots graphiques, incluant les mots outils. Cette longueur maximale de cinq a été choisie d'une part pour son aspect pratique (une longueur supérieure aurait impliqué des contraintes de programmation supplémentaires), mais aussi parce que peu de segments répétés ont une longueur supérieure, du moins dans les textes impliqués lors de l'expérience (des textes de résolution syndicale, et des numéros du journal *Le Père Duchesne*). Ils se rapprochent en cela de la collocation chez [Sinclair, 1991], qui est considérée dans un intervalle de quatre mots maximum séparant les membres d'une collocation (voir la sous-section 4.1.2.3 page 112).

Cette approche des segments répétés, couplée à celle des collocations en linguistique de corpus, nous paraît être la plus pertinente compte tenu de nos objectifs. La notion de collocation met en avant une approche interprétative de ces segments. Le segment répété, au sein d'un seul et même document ou dans plusieurs documents, est représentatif de ce document ou de cet ensemble. Il permet d'apporter une dimension concrète, « saisissable », aux collocations, qui sont d'après [Williams, 2003] « l'Arlésienne de la linguistique, tout le monde en parle, mais elles restent difficilement saisissables ». Ainsi, en tant qu'éléments représentatifs d'un document textuel ou d'un ensemble de documents textuels, nous considérons que ces unités, entre collocations et segments répétés, reflètent le ou les thème(s) de ces derniers.

Les collocations brutes, porteuses des connaissances thématiques d'un ensemble documentaire, sont donc dans notre contexte de travail des unités portant les caractéristiques des segments répétés et celles des collocations.

A titre d'exemple, nous avons trouvé ces segments et mots graphiques dans nos données, et plus précisément dans l'échantillon de 2 000 documents d'un ensemble documentaire ratta-

ché à une étude portant sur l'optique. Chacun d'entre eux est répété au moins deux fois dans l'échantillon (les nombres entre parenthèses renvoient à leur fréquence) :

- *high resolution imaging* (128)
- *refractive power* (109)
- *optical coherence tomography* (69)
- *combustion chamber* (19)
- *store the image data* (4)
- *coverage of the surveillance region* (4)
- *satellite communication robot* (3)
- *capillary Z pinch plasma* (2)
- *ultrasound producing groupware* (2)

Toutes ces séquences ont en commun de relever de domaines de spécialité particuliers. Certaines d'entre elles peuvent être considérées comme des termes au sens de [ISO, 1999] et de [Otman, 1995] (voir 4.1.2.1 page 107), puisqu'elles expriment un seul et unique concept à l'aide d'une forme nominale : *refractive power* ou *combustion chamber* par exemple correspondent en effet aux critères. En revanche, certaines formes débordent du cadre fixé pour les termes : typiquement, il est difficile de donner à *store the image data* ou *ultrasound producing groupware* le statut de terme, celui-là puisqu'il s'agit d'une forme verbale, celui-ci parce qu'il est possible de le voir comme une unité phraséologique au sens de [ISO, 1999].

En revanche, toutes ces séquences rentrent dans le cadre définitoire des collocations, en tout cas du point de vue de la forme brute, puisque toutes sont des séquences lexicales. En ce qui concerne la récurrence de ces séquences, c'est-à-dire la condition de régularité des collocations, elle est relative.

Rappelons que dans notre contexte d'application, un ensemble documentaire différent est constitué pour chaque nouvelle étude. Il n'est donc pas rare qu'un même document soit intégré à plusieurs études à la fois, et qu'il puisse être digne d'intérêt pour le lecteur selon des points de vue différents en fonction de l'étude. A ce titre, il est possible qu'au sein d'un ensemble documentaire, ou d'une partie de celui-ci, les segments répétés ne soient pas les mêmes que dans un autre. A supposer qu'un document x, portant sur une comparaison entre les énergies éolienne et hydrolienne, soit utilisé dans une étude concernant les hydroliennes, il est tout à fait envisageable que les segments répétés qui s'en dégagent soient différents de ceux qui seraient révélés pour ce même document dans un autre ensemble documentaire, cette fois pour une étude portant sur les éoliennes.

Il est donc nécessaire de tenir compte du caractère relatif du segment répété dans notre cas ; la conséquence directe en est que les segments répétés ne peuvent être extraits une fois pour toutes et dans l'absolu pour chaque document entrant dans la base, puisque ces calculs seraient non pertinents car effectués pour la totalité de la base, incluant tous les ensembles documentaires

dédiés à des études spécifiques.

Ce parti pris est à rapprocher du fait que le thème, en tant qu'objet de l'*aboutness*, est relatif et déterminé par l'objectif pour lequel le document est utilisé (voir 4.1.1.4 page 106). Potentiellement, un même document peut donc être utilisé dans des ensembles documentaires différents correspondant à autant d'objectifs.

Nous partons donc de la seconde hypothèse suivante :

Hypothèse II.9. *L'identification des collocations brutes pour un ensemble documentaire donné peut passer par l'utilisation d'une approche inspirée des segments répétés, calculés pour un ensemble documentaire donné et non pour la totalité des documents de la collection, de manière à respecter le caractère relatif que nous avons attribué à la notion de thème.*

Les segments répétés ne sont donc pas calculés pour la base documentaire dans sa totalité : étant données la grande variété des sujets d'étude, et la méthode employée par les analystes consistant à créer pour chaque étude un ensemble documentaire spécifique, effectuer ce calcul sur tous les documents intégrés manquerait de cohérence et serait peu efficace. Au contraire, afin de respecter le caractère relatif du thème tel que nous le considérons, le calcul des segments répétés doit être lancé au sein d'un seul ensemble documentaire, et donc pour une étude précise.

Nous estimons que toutes les collocations brutes relevées ne véhiculent pas les thèmes ou sous-thèmes des documents. Nous avons pris le parti, dans ce traitement, d'inclure du bruit dans les collocations brutes potentielles, plutôt que de générer du silence. En effet, le bruit sera détecté par l'utilisateur du système, qui sera capable d'identifier une collocation brute relevant d'un thème, et de la distinguer d'une autre n'en relevant pas. En revanche, le silence ne pourrait être détecté en tant que tel par l'utilisateur.

6.2.3.2 Application

D'un point de vue formel, nous avons mentionné ci-dessus que nos segments répétés ne correspondaient pas strictement à la définition qu'en font [Lafon & Salem, 1983]. En effet, ces auteurs considèrent comme segments répétés les suites de deux à cinq mots graphiques présents au moins deux fois dans une unité donnée, en l'occurrence le corpus. Sont donc écartés de cette définition les mots simples qui seraient récurrents. Or, nous partons du postulat qu'un thème ou sous-thème global peut parfois s'exprimer par un mot simple, en particulier du côté du lecteur s'il est en phase de découverte de son corpus de travail et qu'il connaît peu le domaine traité. C'est pourquoi nous intégrons les formes simples récurrentes aux segments répétés que nous repérons.

D'autre part, [Lafon & Salem, 1983] utilisent des segments répétés contenant des mots grammaticaux en début et en fin de segment. Cela correspond à leur objectif d'analyse des spécificités de certains types de discours.

Cependant, dans notre cas, la présence de telles unités en début ou en fin de segment générerait, à notre sens, un bruit important. Par conséquent, nous excluons de notre propre inventaire des segments répétés ceux dont le premier et/ou le dernier mot est un mot grammatical. Les mots grammaticaux peuvent en revanche se placer à l'intérieur d'un segment, puisqu'ils permettent parfois de structurer des formes complexes pouvant véhiculer un thème.

Ainsi, dans le cadre de nos travaux, un segment répété, et donc une collocation brute, est un mot plein simple, ou une suite de deux à cinq mots graphiques commençant et finissant par un mot plein.

Ces méthodes endogènes pour la ressource termino-ontologique, et en l'occurrence pour la facette thématique, permettent l'organisation de cette facette en fonction d'unités révélant le thème des documents. Ce sont les collocations brutes, par le biais du calcul des segments répétés, qui permettent de structurer la facette, et par là-même, structurent l'ensemble documentaire du point de vue thématique.

Les segments répétés sont donc utilisés comme structurant l'ensemble documentaire, et ce en raison de leur récurrence : plus un segment est présent, plus il est représentatif d'un document ou d'un ensemble de documents. Cette structuration se place à deux niveaux.

Un segment répété dans un même résumé de document est ainsi représentatif de ce document, et de l'un de ses thème(s). Ce thème intra-documentaire se place donc au niveau le plus fin. A un second niveau, la récurrence d'un segment répété au sein de plusieurs documents permet de dégager un thème commun à tous ses documents, soit un thème inter-documentaire.

Comme pour les autres facettes, nous l'avons expliqué, la structure du plan permet la création de plusieurs types de liens :

- les liens entre collocations brutes au sein d'un même document ;
- les liens entre les collocations brutes de différents documents ;
- les liens entre les collocations brutes et les unités d'informations d'autres facettes ;
- les liens entre les documents qui utilisent les mêmes collocations brutes.

A partir de là, il est possible de dégager des ensembles de segments répétés se retrouvant dans plusieurs documents. Ces ensembles rapprochent alors ces documents, et plus les segments répétés communs sont nombreux, plus ces documents sont proches, puisque partageant un ou plusieurs thème(s). Une répartition des documents par leurs segments répétés communs permet alors d'effectuer une segmentation thématique de l'ensemble documentaire. Les utilisateurs peuvent s'appuyer sur cette segmentation pour mener leurs analyses.

Enfin, si un segment répété est commun à tous les documents de l'ensemble documentaire, alors il est représentatif de cet ensemble : il en constitue un thème global. Cependant, il n'apporte rien à une analyse plus locale : il est trop peu discriminant, car il est trop présent pour autoriser une segmentation des documents en sous-ensembles, et donc pour structurer l'ensemble documentaire.

6.2.3.3 Validation

Sur les exemples de suites de mots que nous avons cités ci-dessus, la recherche de collocations brutes permet bien de structurer les ensembles documentaires.

Nous avons soumis notre corpus de 2 000 documents au calcul des segments répétés. Nous en avons dégagé, au total, 314 610 collocations brutes, soit des segments textuels correspondant à nos critères et apparaissant au moins deux fois dans l'ensemble documentaire. Nous rappelons que ce corpus est un échantillon d'un ensemble documentaire plus important, portant sur le domaine de l'optique.

Parmi elles, 216 828 formes simples sont dégagées, ainsi que 97 782 séquences d'au moins deux mots, et au plus cinq mots.

En moyenne, un document contient donc environ 157 collocations brutes, répétées au moins deux fois, que ce soit au sein de ce même document, ou plus largement dans l'ensemble documentaire.

Les collocations brutes les plus courantes peuvent fournir des indications sur la teneur thématique de l'ensemble documentaire dans sa globalité. Par exemple, des collocations telles que *high speed*, *high resolution*, *high speed camera* ou *image processing* sont très courantes, et permettent de savoir que de manière générale, le corpus traite d'optique, et plus précisément de caméras haute rapidité et haute résolution et de traitement d'image, ainsi que nous l'avons signalé plus haut.

Néanmoins, les plus récurrentes de ces collocations ou formes simples, nous l'avons mentionné page 217, sont peu utiles dans de telles études si l'utilisateur est expert. Un certain nombre d'entre elles sont en effet trop peu discriminantes pour que celui-ci en tire de l'information pertinente, puisqu'en tant que concepteur de l'ensemble documentaire qu'il va traiter, il connaît déjà son thème global. Des formes comme *imaging*, présentes à 2 581 reprises dans notre échantillon, seront donc considérées comme du bruit.

Cependant, ces collocations restent globalement représentatives d'un ensemble documentaire. Elles sont informatives, de manière générique, pour des individus novices du domaine traité par exemple. Il convient donc de les rendre accessibles pour ces derniers, ne serait-ce que pour qu'ils puissent vérifier la cohérence thématique des documents qu'ils ont collectés lors de la tâche de récupération. Ces utilisateurs peuvent alors faire le choix de les sélectionner dans une première approche d'un domaine qu'ils ne maîtrisent pas, lors de leur immersion dans l'ensemble documentaire qu'ils doivent analyser. Dans ce cas, ces thèmes génériques endogènes font l'objet d'une sélection anthropogène liée au contexte, et en particulier au niveau d'expertise de l'utilisateur du système d'immersion (voir section 8.3).

D'autres collocations relèvent de la structure discursive propre aux genres des documents. Par exemple, la collocation *problem to be solved*, massivement présente, est représentative de la

structure des brevets, et de leur résumé.

Ces deux types de collocations très courantes nous sont donc de peu d'aide pour l'identification spécifique des sous-thèmes des documents en présence : les thèmes généraux sont en principe déjà connus par l'utilisateur ; les collocations discursives ne sont quant à elles pas utiles dans ce contexte, puisque l'objectif n'est pas de décrire la structure des documents, mais leur contenu propositionnel.

En revanche, des collocations brutes moins fréquentes, mais restant présentes, peuvent apporter une information précieuse aux utilisateurs.

L'objectif de l'étude nécessitant l'ensemble documentaire sur l'optique était de trouver des champs d'application possible aux caméras haute rapidité ou haute sensibilité. Les collocations brutes peuvent être utilisées à cette fin, puisque sur un large corpus de documents traitant de ces caméras, il est possible de distinguer plusieurs domaines d'applications, justement par la récurrence de séquences dans les textes des résumés. Il est alors possible de rassembler des documents en fonction de leurs collocations brutes communes.

Les champs d'application possibles détectés lors de l'étude sont au nombre de 11. Pour les besoins de notre démonstration, nous n'en sélectionnons que quatre :

- l'imagerie de la combustion ;
- la vélocimétrie, c'est-à-dire la mesure de la vitesse et de la direction d'un fluide ;
- les applications militaires ;
- l'astronomie.

De l'inventaire des collocations brutes se dégagent un certain nombre de séquences clés renvoyant à ces domaines. Nous en présentons quelques-unes dans le tableau 6.22 page suivante, associés au nombre de documents qu'elles permettent de rapprocher.

Les collocations brutes permettent donc bien de distinguer plusieurs sous-thèmes au sein d'un ensemble documentaire. En l'occurrence, les collocations brutes opèrent une segmentation thématique des documents, dégageant ainsi des thèmes divers. Notons que cet ensemble documentaire a été utilisé pour trouver des applications : les collocations brutes que nous présentons sont donc en lien avec celles-ci. Cependant, avec un autre objectif, d'autres collocations auraient pu être prises en compte, et en particulier celles renvoyant à différentes techniques d'imagerie si l'objectif avait été un état de l'art technique.

Par l'extraction de l'ensemble des collocations brutes correspondant à nos critères dans un ensemble documentaire donné, nous pouvons prétendre à l'exhaustivité. Celle-ci est un moyen de valider l'existence et la répétition d'une suite de mots, sans toutefois préjuger de sa pertinence. Ainsi, elle offre à l'utilisateur des possibilités d'interactions plus larges que si les données étaient déjà filtrées au moment de notre extraction selon des critères fixés à l'avance (voir la section 8.3 page 349). Elle constitue par exemple un moyen de réaliser une première étape de reformulation, puisque lorsque l'utilisateur interroge par ces collocations le système d'immersion, il sait immé-

Champ d'application	Collocations brutes	Nombre de répétitions	Nombre global de documents associés par les collocations brutes
Imagerie de la combustion	Liquid fuel	12	21
	Gas turbine	8	
	Combustion chamber	21	
	Internal combustion engine	3	
	combustor	19	
Vélocimétrie par images de particules	Laser Doppler	6	63
	Optical coherence tomography	69	
	Fluid particles	14	
Applications militaires	Inspection system	12	25
	Surveillance camera	5	
	High resolution surveillance	2	
	Port and harbour securtiy	4	
Astronomie	Earth observation	8	35
	High resolution satellite	3	
	Mars Orbiter (*dont* Mars Orbiter camera)	10 (5)	
	High resolution spaceborne	3	

TABLE 6.22 – Exemples de collocations brutes permettant de dégager des thèmes de l'échantillon de l'ensemble documentaire « Optique »

diatement si les suites qu'il a choisies sont présentes ou non dans les documents (si tant est qu'il respecte les critères de forme de ces collocations, que nous venons d'énoncer).

L'extraction des segments répétés telle qu'elle est proposée par [Lafon & Salem, 1983] est un traitement complexe, et relativement coûteux. Cependant, si nous comparons ce coût à celui qu'auraient nécessité des analyses morpho-syntaxiques, beaucoup plus élaborées du point de vue des règles à appliquer, l'algorithme des segments répétés reste finalement intéressant en termes de coût de traitement. Or, nous avons fait le choix stratégique de minimiser les calculs, en raison du contexte d'application de nos travaux. En effet, dans le milieu industriel dans lequel nous nous plaçons, il est problématique de mettre en place des systèmes nécessitant trop de calculs.

Nous avons mentionné que notre traitement ne sélectionne que les collocations brutes potentielles commençant et finissant par un mot plein. Pour cela, nous avons besoin d'informations exogènes, et plus précisément de lexiques de mots vides, ou mots grammaticaux. Ils sont en effet un moyen de détecter les mots grammaticaux, et d'éliminer les segments indésirables. Nous présentons ces lexiques, dans le cadre de l'algorithme qui les exploite, dans le chapitre 7 page 239.

Il est à noter, cependant, que nous aurions pu utiliser, en lieu et place de méthodes exogènes, des techniques endogènes telles que celle décrite par [Vergne, 2004]. Sa méthode permet en effet de détecter les mots vides dans des textes dont la langue n'est pas connue à l'avance. Cette détection s'effectue par un calcul local, et sans ressources externes. Toutefois, si la méthode présente un intérêt évident, elle n'est pas pertinente pour nos travaux, en raison d'un coût de traitement élevé.

C'est pourquoi, dans ce cas précis, il s'avère plus avantageux de faire appel à des ressources externes, d'autant plus lorsque celles-ci représentent des listes fermées.

6.2.4 Conclusion

La ressource termino-ontologique (RTO) multi-plans a été conçue dans l'objectif de structurer les documents d'un ensemble pour mieux l'exploiter.

Nous avons vu que cette RTO était constituée en grande majorité à l'aide de ressources endogènes : les informations intrinsèques des documents sont exploitées pour leur propre représentation.

Les informations contextuelles décrivant la situation de production des documents sont issues de ceux-ci, qu'elles aient été transformées ou non dans l'intervalle. Quant aux informations thématiques, elles sont extraites à l'aide de traitements exploitant les récurrences des titres et résumés du corpus.

L'ensemble documentaire comme ressource endogène est un moyen efficace d'obtenir une RTO qui soit adaptée à ce même ensemble.

La RTO résultante crée des liens documents et informations extraites, mais également entre

les informations elles-mêmes, qu'elles soient de même type ou de types différents. Ces liens sont déterminés par intersection, *via* les documents.

Ces intersections entre éléments d'informations répartis dans les cinq facettes mènent à l'établissement de relations entre les documents eux-mêmes. Elles permettent par là d'en tirer des informations potentiellement pertinentes, à travers les recherches que permet notre système d'immersion, au niveau de granularité le plus global.

6.3 L'immersion documentaire par endogénéité

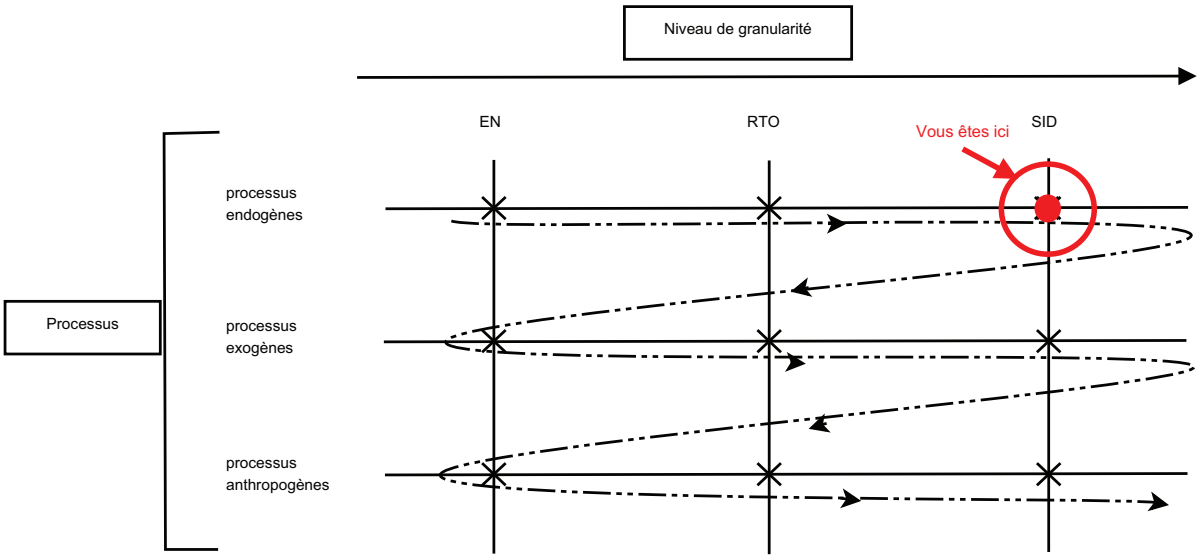

FIGURE 6.14 – Positionnement des traitements endogènes pour le système d'immersion documentaire dans l'ensemble des processus

Dans les deux sections précédentes, nous avons abordé les ressources endogènes du point de vue des niveaux de granularité fin et intermédiaire, que sont respectivement les entités nommées et la ressource termino-ontologique (RTO) que nous avons constituée. Nous traitons ici des traitements et ressources endogènes utilisées pour le troisième niveau de granularité : le système d'immersion documentaire (voir la figure 6.14).

A chaque étape, nous avons tiré parti, le plus possible, des informations véhiculées par le corpus à traiter lui-même. Ainsi, un ensemble documentaire propre à une étude fournit lui-même une grande partie des clés de son traitement, pour sa normalisation, et pour sa structuration *via* une RTO multi-plans.

Par ces traitements aux deux premiers niveaux, l'ensemble documentaire fournit donc, *in fine*, des moyens pour l'explorer.

Un ensemble documentaire est donc, comme pour les unités d'informations et pour la struc-

turation de la RTO, une ressource endogène pour une immersion au sein des documents qu'il contient.

Parler de documents comme ressource endogène pour un système d'immersion, ou plus largement de navigation documentaire, peut sembler tenir de la tautologie. Quoi de plus normal en effet que d'exploiter les documents pour naviguer parmi eux ?

Pourtant, la manière dont nous exploitons ces documents comme ressource endogène ne va pas forcément de soi. En réalité, il peut exister trois « types » de représentation des documents pour la navigation et l'immersion.

- le document peut être utilisé en tant que représentation de lui-même. C'est par exemple le cas dans les listes plates fournies en résultat de requête par les moteurs de recherche tels que Google. La ressource est alors endogène, et constituée des objets traités eux-mêmes ;
- des ressources exogènes sont également exploitables, telles que des ontologies de domaine pré-construites de manière externe aux documents qu'elles vont représenter ;
- enfin, il est possible d'utiliser des ressources constituées de manière endogène, à partir des documents, qui vont au-delà des documents eux-mêmes puisqu'elles en fournissent une représentation construite.

Nos travaux viennent se placer dans cette dernière catégorie, puisqu'ils exploitent les informations des documents pour les représenter et les structurer.

Nous sommes donc loin, finalement, de la tautologie énoncée : les documents constituent non seulement la source de l'information à rechercher, mais aussi le matériau qui permet de constituer la ressource, et les outils qui s'appuient sur cette ressource, nécessaires à leur propre exploration.

Dans la première partie de ce document, nous avons présenté ce que nous appelons l'immersion documentaire (chapitre 2 page 25). Cette dernière a pour objectif de permettre à l'utilisateur d'un système d'immersion d'accéder aux informations contenues dans un ensemble documentaire donné. Dans nos travaux, l'utilisateur est considéré non plus comme un « simple » utilisateur d'un outil, mais comme concepteur de la connaissance pertinente pour lui à un moment *t*. Pour donner cette place à l'individu comme concepteur de connaissances, il convient de le placer au centre des données, et de l'immerger au sein de l'ensemble documentaire de manière à ce qu'il puisse agir directement sur cet ensemble, et en tirer la connaissance dont il a besoin.

Pour cela, nous suivons [Tricot, 2006] [51], qui montre que la visualisation permet de maîtriser « l'espace informationnel des organisations ». Selon lui, il convient de satisfaire un ensemble de besoins des utilisateurs pour que la cartographie, qu'elle porte sur des données scientifiques ou des données abstraites, soit pertinente.

Tout d'abord, la cartographie doit permettre de naviguer selon la sémantique du domaine

51. Nous faisons référence ici à Christophe Tricot, dont les travaux prennent place dans le domaine de la gestion des connaissances et de la cartographie sémantique, et non à son homonyme André Tricot, chercheur en psychologie cognitive et ergonomie, dont nous avons notamment parlé dans la première partie de ce document.

dans un espace informationnel. C'est le moyen pour l'utilisateur de comprendre, assimiler et exploiter cet espace.

D'autre part, la visualisation doit « offrir simultanément une vision globale et synthétique » (*ibid.*) de cet espace, matérialisé pour nous sous la forme d'un ensemble documentaire. Une visualisation efficace permet donc d'embrasser un grand nombre d'informations, tout en les présentant de manière synthétique. D'autre part, elle doit permettre des allers-retours entre informations globales et informations locales : selon [Roy, 2007], un outil de navigation doit alterner représentations visuelles globales, pour permettre une appréhension synthétique des informations, et locales, pour la prise en compte des particularités d'un ensemble.

Enfin, la visualisation cartographique doit proposer à l'utilisateur des cartes qui lui sont adaptées, en fonction de son activité, de son niveau d'expertise, etc.

Ainsi, pour construire un système d'immersion efficace, et donc adapté à une situation de recherche d'information, il est nécessaire de prendre en compte les utilisateurs, dans leurs points communs comme dans leurs particularités, mais aussi de conserver systématiquement les liens existant entre niveau local et niveau global des informations de l'ensemble documentaire. Nous l'avons dit, nous rajoutons un troisième critère fondamental pour le fonctionnement de cette immersion : la place de l'utilisateur se trouve, de notre point de vue, à l'intérieur du processus, et non face à lui : il doit être immergé dans l'ensemble documentaire. Nous représentons cette conception de l'immersion pour la recherche d'information (RI) à droite dans la figure 6.15 tirée de notre première partie (chapitre 2 page 25).

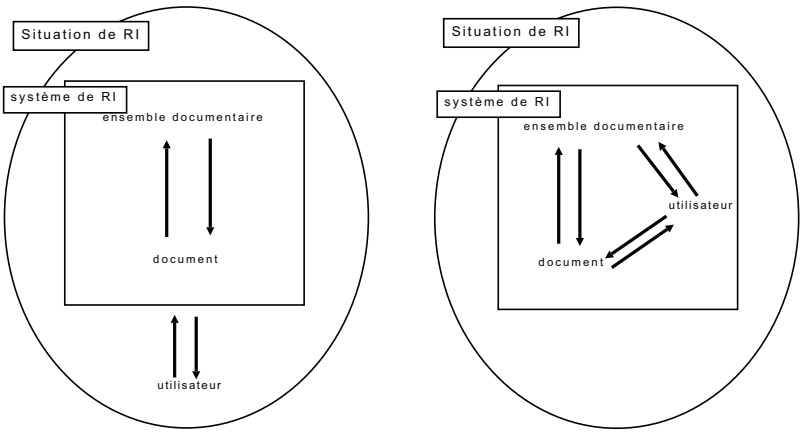

FIGURE 6.15 – Deux manières d'envisager l'utilisateur en situation de recherche d'information (RI) : l'utilisateur comme agent externe du système (à gauche), ou comme composante interne du système (à droite) (repris du chapitre 2 page 25)

Dans cette section, nous abordons plus précisément l'immersion documentaire, en particulier dans la manière dont elle est conçue et matérialisée. Tout d'abord, nous exposons les critères d'accès à l'ensemble documentaire utilisés pour l'immersion. Ils permettent à l'utilisateur d'*entrer*

en immersion. Puis nous présentons les dimensions informationnelles utilisées pour ces critères d'accès, et pour la navigation elle-même. L'utilisateur les exploite à la fois pour entrer et immersion, et lorsqu'il *est immergé* à proprement parler dans le système et dans les documents. Enfin, nous traitons des particularités propres à certaines des dimensions.

6.3.1 L'entrée en immersion documentaire par endogénéité : définition et fonctionnement

Exploiter un ensemble documentaire (ED) pour permettre l'immersion dans ce même ensemble est un processus endogène qui est, dans le cadre de nos travaux, complexe.

Cet ensemble documentaire n'est pas restitué sous sa forme brute, et de manière énumérative, dans le système d'immersion. L'immersion passe en effet par la mise en place d'un interfaçage, non entre l'humain et la machine, mais entre l'humain et le document par le truchement de la machine. Cet interfaçage consiste, pour nous, à fournir des clés d'interprétation à l'utilisateur lors des différentes étapes de consultation d'un ED. A ce titre, le système prend en compte l'ensemble des éléments participant à l'immersion, dont l'utilisateur. Nous renvoyons au schéma 6.15 ci-dessus.

Pour cela, l'utilisateur accède aux documents par le biais des informations contenues dans la RTO multi-plans. Ces informations conceptuelles et contextuelles sont restituées sous forme graphique, souvent en réseau, dont les nœuds sont des éléments d'information de la RTO, eux-mêmes issus de l'ensemble documentaire. Il peut s'agir d'éléments thématiques, d'organisations, etc., ou des documents de l'ensemble eux-mêmes, isolés ou regroupés au sein de sous-ensembles. Les arcs des réseaux représentent les relations entre nœuds, elles-mêmes issues des documents, et donc endogènes. Un arc liant deux organisations peut par exemple représenter une relation de co-dépôt d'un brevet.

Le système d'immersion offre des représentations graphiques dynamiques et interactives, qui consistent souvent en des réseaux impliquant une masse importante d'informations, et donc un grand nombre de nœuds et d'arcs. Comparées à des représentations statiques de cet ensemble d'informations, les vues dynamiques offrent plus de souplesse : d'une part, elles offrent des possibilités d'immersion par des vues plus locales du réseau, ou au contraire des vues plus globales. D'autre part, elles permettent une navigation dans la vue, grâce à des parcours de nœud en nœud par le biais des arcs, et donc de suivre un lien d'information.

Dans tous les cas, l'accès aux documents par visualisations dynamiques, et donc l'entrée en immersion, sont réalisés en fonction de trois types de critères intégrés à une requête, correspondant à trois types d'informations nécessaires à la construction d'une représentation : le point d'entrée, et les modalités d'affichage et le filtrage. Ceux-ci sont sélectionnés par l'utilisateur. Ce choix dépend de paramètres concernant son besoin d'information, sa tâche globale, le contexte

et le sujet de l'étude, qui relèvent de l'humain et de processus anthropogènes, que nous abordons dans la section 8.3.

Le système d'immersion doit avant tout permettre à l'utilisateur d'exprimer son but de recherche d'information par une requête, qui soit la plus proche possible de son besoin réel. Cette expression dépend en partie de la structuration du système de navigation utilisé, qui contraint les modalités de formulation de la requête [Tricot, 2003], et peut donc, finalement, être plus ou moins utile et plus ou moins utilisable.

La conception de la RTO et de sa structure, sur lesquelles s'appuie notre système d'immersion, a donc un impact direct sur la qualité de l'expression de la requête, puisqu'il n'est possible d'interroger des documents que sur des éléments qui ont été formalisés en amont de cette interrogation. En ce sens, c'est entre autres ici que les inscriptions numériques agencées dans la RTO, comme dans toute structure de représentation des connaissances, jouent une partie du rôle qui leur est attribué par l'ingénierie des connaissances : elles sont un moyen pour l'utilisateur, à travers la requête qui les emploie, d'agir sur la globalité des données de l'ensemble documentaire [Charlet, 2004].

La possibilité de faire appel à ces inscriptions numériques dans le système est donc sous-tendue par la RTO multi-plans et la structuration de l'ED qu'elle permet.

Concrètement, les trois critères d'immersion utilisant ces inscriptions sont considérés de manière distincte, et peuvent s'exprimer par la question générique d'un utilisateur :

- *Je veux voir quelque chose,*
- *calculé et/ou visualisé d'une certaine manière,*
- *et filtré selon certains critères.*

Chaque partie de cette question est un argument utilisable dans une requête. Dans ce qui suit, nous les désignons donc par le terme d'*arguments de la requête*. Le premier argument fait référence aux informations qui intéressent prioritairement l'utilisateur dans sa recherche. Celles-ci représentent donc le *point d'entrée* dans un ensemble documentaire.

Le deuxième argument renvoie à la façon dont les informations du point d'entrée doivent être traitées puis visualisées. Il permet donc d'établir les *modalités de visualisation* des points d'entrée.

Enfin, le dernier argument permet de filtrer les informations choisies pour le point d'entrée, en fonction de valeurs relevant d'un ou plusieurs types d'informations, et déterminées par l'utilisateur. Cet argument concerne donc les *critères de filtrage*.

Nous résumons les arguments de la requête pour l'immersion documentaire dans le tableau 6.23.

Les types de données sont ceux qui sont répartis dans la RTO, par le biais des cinq facettes. Celles-ci sont utilisées dans l'immersion comme des *dimensions informationnelles* : lorsqu'elles sont croisées grâce aux arguments de la requête, elles forment un espace multidimensionnel dans

Critères d'accès	Arguments de la requête	Informations nécessaires
Point d'entrée	*Je veux voir...*	Type(s) de données
Modalité de visualisation	*Calculé et/ou visualisé par...*	Type(s) de données/ regroupement(s)/ calcul(s)
Critère de filtrage	*Filtré selon...*	valeur(s)

TABLE 6.23 – Requête générique pour l'immersion : critères d'accès, arguments de requête et informations en jeu

lequel viennent se placer les documents de l'ensemble documentaire.

L'utilisateur peut souhaiter considérer comme points d'entrée des organisations, des individus, des mots-clés par exemple, en les visualisant par regroupements en fonction de propriétés de différents types : par année, par type de document, par mots-clés identiques, par maison mère, etc. De plus, il est possible d'effectuer des calculs de comptage, comme le nombre de documents reliés aux informations des points d'entrée, des calculs de proximité, etc. Enfin, il peut filtrer ces informations selon des empans de valeurs ou des valeurs isolées appartenant à n'importe laquelle des dimensions.

Nous pouvons alors reformuler notre requête générique de manière plus formelle, avec des arguments précisés :

- *Je veux voir* ***un type d'informations déterminé*** *[quelque chose],*
- *calculé et/ou visualisé* ***par type d'informations, par regroupement de types ou sous-types d'informations et/ou par calcul*** *[d'une certaine manière],*
- *et filtré* ***selon une valeur ou un intervalle de valeurs*** *[selon certains critères].*

Nous avons distingué les trois arguments de manière formelle, tout en étant consciente que dans l'absolu, la frontière peut parfois être mince du point de vue du sens. Par exemple, le type d'informations que l'utilisateur veut voir influence nécessairement les modalités de visualisation, puisqu'il va, justement, les voir représentées. Les modalités de visualisation dépendent donc au moins en partie du point d'entrée choisi Il en est de même pour les critères de filtrage, du fait même qu'ils peuvent s'appliquer à l'ensemble des informations : ils s'appliquent soit aux objets que l'utilisateur veut voir, soit aux éléments calculés pour la visualisation, soit aux objets liés aux objets à voir ou aux objets calculés. Ainsi, les arguments entretiennent entre eux des relations de dépendance.

Cependant, ils permettent à l'utilisateur d'élaborer, à destination de la machine, une requête complexe de manière relativement naturelle ; d'autre part, ils exploitent de manière simple les caractéristiques des cinq dimensions informationnelles. Nous maintenons donc cette séparation nette.

Les dimensions informationnelles forment le socle des valeurs qui peuvent être sélectionnées

par les trois arguments de la requête. Elles contiennent les informations potentiellement pertinentes exploitables lors de l'utilisation du système d'immersion. De plus, la requête offre une dernière possibilité : celle d'accéder directement aux documents eux-mêmes, qui constituent la facette d'origine de la RTO. Celle-ci forme une supra-dimension, transversale des cinq autres. Ainsi, l'accès aux documents peut s'effectuer indirectement, par le biais des dimensions informationnelles formées par les types d'informations issues des documents, et qui constituent les facettes calculées ; mais il peut aussi se faire directement, par la supra-dimension transversale.

6.3.2 Fonctionnement des dimensions informationnelles pour les phases d'entrée et d'immersion

Les dimensions informationnelles sont constituées à partir des cinq facettes de la ressource termino-ontologique (RTO) multi-plans. Elles contiennent donc les mêmes types d'informations et les mêmes données que cette dernière, qui est endogène puisque ses données sont issues des documents (voir chapitre 3 page 69 sur le contenu des facettes) :

- les organisations ;
- les auteurs ;
- les dates ;
- les lieux ;
- les thèmes.

Chacune des dimensions contient donc des données directement issues des documents. Chaque dimension est liée de fait aux documents de l'ensemble, et les liens entre informations et leurs documents d'origine, leur contexte, sont conservés. Ces dimensions sont par conséquent fondamentalement endogènes, n'ayant comme intermédiaires, ou plutôt comme matérialisations concrètes, que les facettes de la RTO.

Plus précisément, chacune des dimensions renvoie à un aspect informationnel particulier des documents. Par exemple, une entité d'organisation impliquée dans la situation de production d'un ou plusieurs documents sera présente dans la dimension correspondante des organisations, et sera distinguée, entre autres, d'une entité de personne en tant qu'auteur et d'une date en tant que date de publication du document.

La raison d'être des dimensions dans le système d'immersion est leur capacité à représenter les mêmes documents qui ont servi à leur constitution. En exploitant les dimensions, l'immersion exploite donc les informations contenues dans les documents eux-mêmes.

Pour permettre une représentation de ces documents, les dimensions sont utilisées grâce aux trois arguments de la requête, correspondant aux trois critères d'accès aux documents (voir la sous-section précédente). Toutes les dimensions peuvent être utilisées pour chacun des trois aspects, et peuvent éventuellement être combinées entre elles au sein d'un même critère.

Un argument utilise forcément au moins une dimension. Chaque aspect de la requête passe donc, pour une visualisation, par les documents eux-mêmes, en particulier pour établir les relations entre informations récupérées dans lesdites dimensions.

La combinaison des cinq dimensions, croisée à celle des trois arguments, offre une combinatoire très importante, qui peut sembler large. Pourtant, elle est la garantie de la liberté de l'utilisateur dans son accès aux documents, par les informations intrinsèques à ceux-ci. C'est à cette condition qu'un utilisateur peut exercer son libre arbitre sur les données, puisque c'est lui qui peut alors choisir les combinaisons adéquates à son besoin.

D'un point de vue pratique, les cinq dimensions informationnelles de l'immersion sont sur certains aspects exploitées de manière similaire par les critères d'accès aux documents. Nous l'avons énoncé, toutes les dimensions peuvent être l'objet des trois arguments de la requête : chacune d'elles est exploitable en tant que point d'entrée, de modalité de visualisation ou de critère de filtrage.

De plus, plusieurs dimensions peuvent être combinées au sein de chaque argument. Par exemple, un utilisateur peut avoir besoin de filtrer les informations à la fois sur les organisations et les unités thématiques, de façon à ne conserver que certaines organisations, et parmi elles, uniquement celles qui traitent dans leurs documents de thèmes déterminés.

Ainsi que nous l'avons mentionné, dans le critère des modalités de visualisation, toutes les dimensions peuvent faire l'objet de regroupements. De plus, des calculs permettent de traiter les unités d'information choisies. Des calculs de proximité peuvent par exemple être appliqués, ou des comptages sur le nombre de documents associés.

Enfin, lorsque l'utilisateur est entré et qu'il est immergé dans le système, les informations visualisées par le biais des dimensions, qu'elles soient regroupées ou non, filtrées ou non, mènent toujours, *in fine*, aux documents représentés par ces informations. Ainsi, derrière tout élément de visualisation se trouve un document ou un sous-ensemble de documents, permettant à l'utilisateur d'approfondir l'interprétation et de concevoir la connaissance. En cela, nous rejoignons les travaux de [Charlet, 2004], entre autres, qui appuie sur l'importance de la préservation du contexte des informations et connaissances extraites et représentées dans la RTO : le retour au texte doit toujours être possible.

La séparation des données issues de l'ensemble documentaire en plusieurs dimensions a plusieurs avantages concourant à l'efficacité du système d'immersion. Tout d'abord, elle permet une meilleure fiabilité des données : puisque chaque type d'informations est traité de manière spécifique, les informations qui en résultent sont formalisées et exploitées de façon adéquate.

D'autre part, l'utilisateur peut ainsi mieux appréhender les informations. Un accès distinct à chaque type de données lui permet de combiner à l'envi les informations de chaque dimension, et ce en connaissance de cause : il sait, pour chaque dimension, quelles informations elle contient, et comment elles sont agencées. Certaines d'entre elles peuvent être multidimensionnées en interne.

Cet aspect dépend en fait de la représentation que s'en fait l'utilisateur, et donc de leur utilisation. Par exemple, la dimension des organisations peut être considérée comme une liste plate et énumérative, et donc comme une seule dimension, si l'utilisateur ne prend en compte qu'un seul niveau hiérarchique. En revanche, à partir du moment où il exploite la structure hiérarchique, il utilise les deux sous-dimensions au sein même de la dimension des organisations.

Enfin, les calculs effectués sur les données se trouvent relativement simplifiés du fait de cette scission des informations : il peut exister plusieurs calculs entre des éléments de deux dimensions, en particulier en fonction du rôle qui est attribuée à chacune en fonction des critères d'accès dans lesquels elles sont utilisées. Cependant, la liste de ces calculs est limitée et connue.

Puisque l'ensemble de ce processus est fondé sur les dimensions informationnelles, elles-mêmes issues des documents de l'ensemble documentaire et permettant de traiter ces mêmes documents, le processus est lui aussi endogène.

6.3.3 Les différences entre dimensions

Dans la partie 1, nous avons vu que la pertinence de l'utilisateur est celle vers laquelle il convient de tendre [Mizzaro, 1998] (voir le chapitre 1 page 9 du présent document). Elle implique que les informations retournées dans les résultats d'un système de navigation, à partir d'une requête, corresponde au besoin d'information réel de l'utilisateur (*ibid.*).

Ce besoin d'information réel est mouvant d'une étude à l'autre, mais aussi au fil d'une seule et même étude. Il convient donc de concevoir un système à base de connaissances dont le modèle n'a pas à représenter le sens une fois pour toutes, mais plutôt à être un instrument pouvant intervenir dans l'interprétation de la part des utilisateurs [Charlet, 2004]. Pour cela, nous avons pris en compte les principes de l'ingénierie des connaissances, et en particulier ceux du groupe TIA, qui considère les « inscriptions numériques » (*ibid.*) intégrées à un outil comme des moyens d'arriver à l'interprétation et à la construction de connaissances.

Les inscriptions numériques sont représentées, dans notre système d'immersion, au sein de facettes regroupées dans une ressource termino-ontologique. Cette RTO contient les informations nécessaires à la construction de connaissances par l'utilisateur : la formalisation de ces informations, connaissances potentielles, est pensée justement pour que le système d'immersion fondé sur elles réponde à l'ensemble des besoins possibles.

Cette répartition en facettes est motivée par le besoin de distinguer les types d'informations : ces derniers peuvent être plus ou moins pertinents du point de vue des composantes de la pertinence que sont la tâche, le contexte, le sujet. De plus, la structuration en facettes permet de structurer chaque plan en adéquation avec les caractéristiques des informations qu'il contient. Là encore, à l'intérieur même d'une facette, certaines caractéristiques peuvent être plus pertinentes que d'autres en fonction des études menées, et donc là encore, en fonction du sujet, de la tâche

et du contexte dans lesquels se place l'analyste.

Ces facettes permettent de projeter les informations qu'elles contiennent, les inscriptions numériques, en dimensions dédiées. Ces dimensions, par leurs combinaisons et par leurs spécificités, permettent de répondre à une diversité de besoins, soit d'une étude à l'autre, soit lors de l'évolution de la représentation du but de recherche d'information. Un système de navigation efficace est en effet un système permettant de répondre alternativement, lors d'un même processus de navigation, à un besoin informationnel flou puis à un besoin précis, dans un document donné ou dans un ensemble. La combinaison de ces deux catégories de critères forment le but de recherche d'information d'un utilisateur. Nous rappelons dans le tableau 6.24 page suivante la schématisation de cette combinaison définie par [Tricot, 1993], que nous avons présenté dans la section 2.2 page 28.

		Représentation du besoin informationnel	
		Précise	Floue
Localisation de la cible	Unique, localisée	Chercher un renseignement	Explorer
	Multiple, distribuée	Collecter	Butiner

TABLE 6.24 – Les quatre buts de recherche d'information d'après [Tricot, 1993], repris de la figure 2.1 page 31

La projection distincte de ces dimensions est motivée par le fait que les types d'informations qu'elles contiennent ne sont pas tous semblables. Les dimensions explicitent la disparité des données dans leur structure, reflétant partiellement leur sémantique, et autorisent à les exploiter de manière efficace. C'est pourquoi toutes les dimensions ne permettent pas toujours les mêmes manipulations lors de l'immersion, puisque ces dernières sont adaptées aux spécificités des informations à représenter. Nous présentons ici les dimensions de l'immersion en fonction de leurs particularités lors de leur utilisation dans l'un des critères d'accès.

6.3.3.1 Les dimensions énumératives non ordonnées

La dimension des auteurs ainsi que celle des thèmes se présentent sous la forme de listes énumératives plates. Elles n'ont de valeur ontologique que parce qu'elles sont reliées aux documents.

La différence majeure entre les deux dimensions des auteurs et des thèmes est la façon dont leurs unités sont récupérées. Les noms d'auteurs sont extraits des documents, ou plutôt de leurs

en-têtes, tandis que les thèmes, représentés par les collocations brutes, sont calculés à partir des titres et résumés.

L'utilisateur peut accéder au contenu de ces listes par leurs éléments unitaires, ou bien par regroupement artificiel, rendu possible par l'équivalent d'expressions régulières. Nous soulignons que ces types d'accès représentent l'ensemble des possibles, et non ce qui est forcément souhaitable. Sur les unités thématiques, les regroupements par l'usage d'expressions régulières permettent de produire une analyse proche d'une analyse morphologique, à laquelle il est possible d'associer un sens linguistique. En revanche, effectuer des regroupements par de telles expressions sur des noms d'auteurs n'aurait pas réellement d'intérêt pour l'immersion d'un utilisateur, contrairement au cas des unités thématiques.

Du point de vue de leur visualisation, les éléments de ces deux dimensions peuvent prendre la forme de points, lorsqu'ils occupent la place de nœuds d'un réseau, ou de valeurs d'arcs, lorsqu'ils sont utilisés comme des unités de sens. Dans cette fonction, des noms d'auteurs permettent par exemple de calculer le nombre de personnes communes à deux organisations, comme Boeing et Alcatel, afin de les relier plus ou moins fortement sur la base de leur auteurs communs.

Ces unités offrent par ailleurs trois possibilités quant à leur portée sur l'ensemble documentaire. Les empans textuels sur lesquels les calculs sont effectués peuvent donc relever de trois catégories :

1. l'ensemble documentaire dans sa totalité ;
2. le sous-ensemble de documents correspondant à une requête donnée ;
3. le document pris isolément.

Si le premier empan est choisi, les informations dégagées par les unités sont d'ordre global, et peuvent établir des relations entre documents sur la base de leurs auteurs communs. Se placer au niveau du sous-ensemble permet de le caractériser pour lui-même, mais aussi de le positionner par rapport au reste des documents. Enfin, choisir le niveau du document le caractérise de manière isolée, et en tant qu'unité.

Dans le cas des auteurs, considérer ces informations en fonction d'un empan déterminé permet donc de dégager au niveau global les auteurs de l'ensemble documentaire, ou seulement les auteurs d'un sous-ensemble de documents déterminés. Enfin, en travaillant sur les documents en tant qu'unités, il est possible de dégager leurs co-auteurs.

Quant aux collocations brutes matérialisant les thèmes, rappelons qu'elles sont des séquences textuelles répétées au moins deux fois dans l'ensemble documentaire étudié. A partir de là, un utilisateur peut choisir l'empan auquel la répétition doit avoir lieu, ainsi que le nombre de répétitions minimales. Chaque empan offre une caractérisation différente des unités qu'il contient. Calculées sur l'ensemble documentaire, les collocations doivent être répétées au moins n fois dans la totalité de l'ensemble. Dans ce cas, la répétition permet d'établir des liens thématiques entre

documents.

Si l'empan considéré est le sous-ensemble, les résultats fournissent des informations quelque peu différentes. Ils permettent de caractériser ce sous-ensemble, constitué à partir de certains critères, du point de vue thématique. La présence dans ce sous-ensemble de certaines collocations calculées, mais également leur absence, peuvent être significatives.

Enfin, si l'empan choisi est le document, seuls les documents comportant au moins n fois chacun une collocation brute seront renvoyés. Ce calcul peut par exemple être effectué sur des documents ayant déjà subi une sélection, sur des critères variés, et permettre leur classement en fonction de la prégnance du ou des thèmes véhiculés par les collocations brutes, grâce au nombre d'occurrences de celles-ci. Le cumul des trois niveaux permet de combiner relations thématiques entre documents, caractérisation d'un sous-ensemble, et prééminence des thèmes au sein de documents considérés isolément.

La dernière dimension énumérative à prendre en compte est celle des documents, en tant que dimension transversale, que nous avons évoquée plus haut. Elle est exploitable de la même manière que celles des auteurs et des thèmes. Les documents eux-mêmes peuvent être utilisés pour les trois arguments de la requête. Par ailleurs, ils peuvent être représentés alternativement, sur une visualisation, en tant que nœuds ou en tant qu'arcs.

6.3.3.2 La dimension énumérative ordonnée

La dimension temporelle permet d'accéder aux dates d'émission des documents. En tant que liste énumérative, elle offre les mêmes possibilités que les dimensions des auteurs et des thèmes. Des éléments unitaires, comme des années précises, peuvent donc être sélectionnés au sein de la liste par l'utilisateur. Des regroupements sont également réalisables à partir d'expressions agissant comme des expressions régulières ; dans ce cas, plusieurs années peuvent être cumulées, qu'elles soient contigües ou non.

Cependant, en tant que liste ordonnée, cette dimension donne également d'autres moyens de traitement. Elle peut être traitée par segments, et en l'occurrence, définir les bornes d'intervalles temporels. Dans ce cas, les années sélectionnées définissent une ou plusieurs périodes, au sein desquelles viennent se répartir des informations, qui sont alors classées en fonction de ces empans temporels.

Du point de vue de leur représentation, les dates peuvent constituer, comme pour les listes non ordonnées, des points sur une visualisation, et donc être des nœuds, ou bien des liens entre points, si elles sont utilisées comme valeurs d'arcs. Par ailleurs, du fait de leur ordonnancement, elles peuvent définir des espaces distincts sur une vue, englobant alors d'autres informations.

Enfin, puisque cette dimension comporte, pour les dates de brevets, des dates allant du jour précis à l'année, elle tend vers une structure hiérarchisée. De fait, cette hiérarchie est elle aussi

exploitable pour la représentation d'informations relatives à ce type de document.

Par exemple, l'utilisateur peut souhaiter déterminer la date précise de priorité d'un brevet afin de vérifier si ce dernier n'est pas invalidé par un brevet antérieur. Dans ce cas, il considère la dimension du document en tant que medium (voir la sous-section 3.1.2 page 76), et son besoin informationnel est précis et distribué [Tricot, 1993] (voir tableau 6.24 page 232). Il aura alors besoin de visualiser une liste de brevets déterminés en fonction de leur date de priorité, au jour près.

Dans certains cas précis, la dimension linéaire que représentent les dates acquiert un caractère hiérarchisé. Les dates des brevets d'un ensemble documentaire sont en effet plus précises que celles des articles scientifiques : en plus de l'année, le mois et le jour sont accessibles à la demande de l'utilisateur (voir la figure 3.7 page 92 dans le chapitre 3). Ces derniers sont exploités lorsqu'il s'agit de déterminer précisément l'ordre chronologique dans lequel plusieurs brevets ont été déposés. Ces renseignements sont en particulier utilisés lorsqu'un analyste doit évaluer la validité d'un brevet.

6.3.3.3 Les dimensions hiérarchisées

La dimension des organisations et celle des lieux sont toutes deux fondées sur des structures hiérarchiques, puisque leurs informations, ou inscriptions numériques, sont liées entre elles par des relations d'inclusion : une organisation peut être incluse dans une autre, de même qu'un lieu peut être une partie d'un autre lieu.

A ce titre, les unités présentes dans les deux dimensions présentent des possibilités de traitement et d'exploitation similaires. Elles sont notamment utilisables en tant que nœuds ou en tant qu'arcs au sein de réseaux.

Elles peuvent être manipulées, dans certains cas, comme des listes énumératives. Ces cas sont fonction de la représentation que se fait l'utilisateur de ces dimensions. L'utilisation qui en est faite peut alors porter uniquement sur les feuilles de la hiérarchie, ou sur des unités relevant d'un seul niveau de profondeur.

Par exemple, une étude de positionnement pour une université peut se fonder uniquement sur les nœuds de type organisme au plus haut niveau hiérarchique. Ce niveau correspond à celui des universités, et/ou éventuellement des hôpitaux, et d'autres organismes de niveau 1, laissant de côté la subdivision en départements, laboratoires, etc., qui peut exister au sein de chacun d'eux. A l'inverse, une comparaison des laboratoires les plus productifs d'une organisation mère nécessitera de descendre vers les nœuds les plus bas de la hiérarchie, et de considérer les entités de laboratoire.

Nous soulignons que, même face à de tels besoins, c'est la structuration en niveaux plus ou moins fins au sein de la facette correspondante de la ressource termino-ontologique qui auto-

rise une sélection en fonction du niveau hiérarchique souhaité : sans cet agencement, rien ne permettrait la distinction entre unités de niveaux différents.

Dans les autres cas, les deux sous-dimensions de ces hiérarchies sont exploitables. Elles sont notamment utilisées dans l'argument de visualisation de la requête : elles permettent de positionner visuellement des informations sur des structures arborescentes, établissant des liens d'inclusion entre elles.

6.3.4 Conclusion

Les processus d'immersion documentaire peuvent être répartis en deux aspects, correspondant à deux types d'actions de la part de l'utilisateur :

– poser une question représentant son besoin et son but de recherche d'information ;

– s'immerger dans la réponse.

Ces deux aspects sont réalisés par le biais des informations intrinsèques aux documents, et donc de manière endogène, même si ces informations ont d'abord été extraites des documents et scindées.

L'immersion documentaire prend donc appui sur une ressource endogène, dans laquelle l'information est scindée, divisée, ce qui permet de mieux y accéder, en fonction de critères raisonnés à partir d'un besoin informationnel déterminé. L'information est accessible dans l'immersion par autant de points de vue recensés dans la ressource. Sans cette ressource, les documents qu'il contient ne seraient accessibles que grâce à des recherches très limitées et non adaptées aux tâches globales effectuées par les analystes, donnant lieu à des représentations statiques et linéaires.

Or, grâce aux dimensions endogènes, la consultation des documents devient dynamique : par le biais des informations projetées dans le système d'immersion, il est possible de passer d'une visualisation à une autre, par la sélection d'une ou plusieurs informations sur une visualisation donnée ou par transformation de la requête. Ce chemin dans les documents est rendu possible par les liens établis entre types d'informations et documents, et donc par les relations indirectes créées par ce biais entre les différents types d'information.

Grâce à ces processus, il est possible de visualiser l'intégralité de l'ensemble documentaire sur une vue, ou des sous-ensembles de taille variable, jusqu'au document lui-même. Le principe d'allers-retours entre vues globales et vues locales mis en avant par [Roy, 2007], et donc entre visualisations à un niveau macro et visualisations à un niveau micro, est donc respecté.

6.4 Conclusion

Au cours de ce chapitre, nous avons présenté la manière dont nous avons exploité les informations véhiculées par un ensemble documentaire pour en faire une ressource endogène, destinée à traiter ce même ensemble.

La ressource endogène prend plusieurs formes en fonction du niveau de granularité abordé, et de la nature des traitements à effectuer dessus. Elle est cependant toujours constituée des informations fournies par les documents, et est exploitée pour en extraire des données.

Nous présentons, dans la figure 6.16 page précédente, les données extraites par cette ressource, pour chaque type d'information et donc pour chaque facette, en fonction de leur complexité.

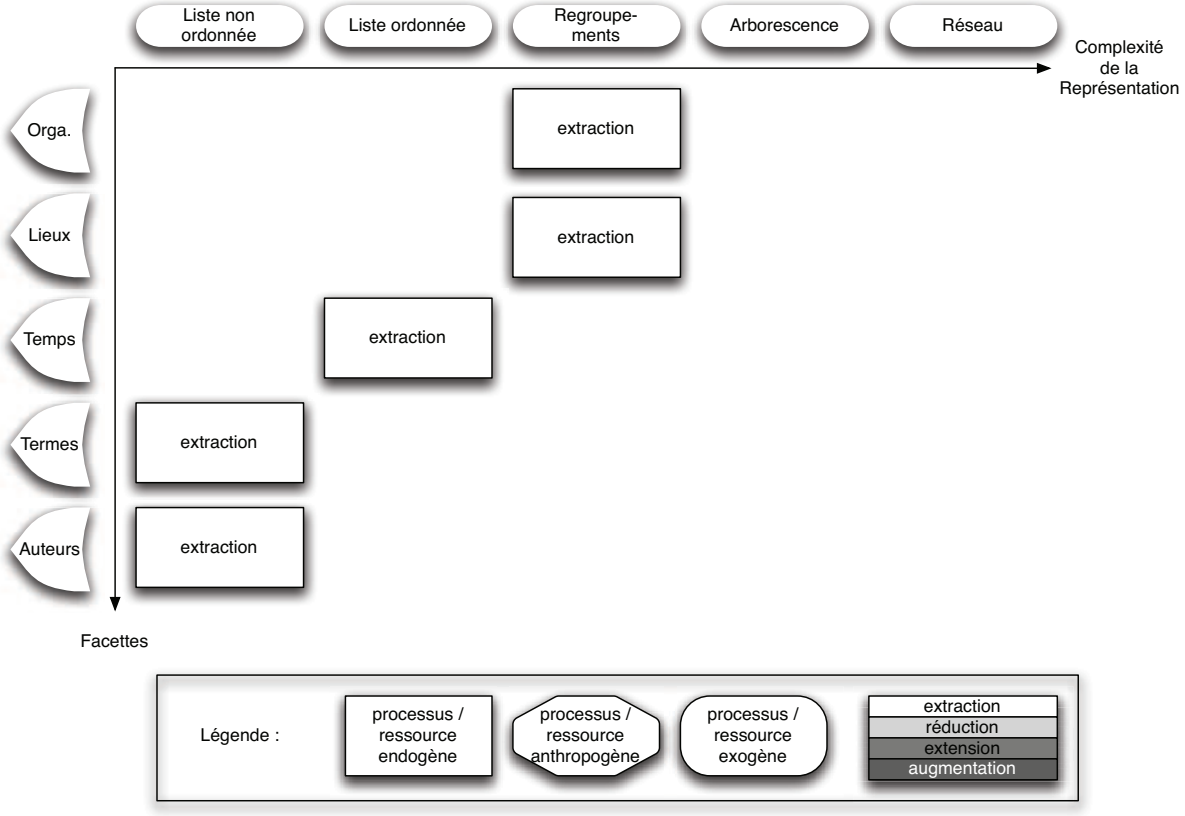

FIGURE 6.16 – Apport des ressources et processus endogènes sur chaque facette et structure résultante

L'extraction endogène permet de structurer les facettes des thèmes et des auteurs en listes non ordonnées ; le plan temporel prend la forme d'une liste ordonnée ; enfin, les organisations et les lieux sont organisés par extraction endogène en listes hiérarchisées.

Elle est une source utilisée pour le traitement des entités nommées en tant qu'expressions linguistiques à normaliser et structurer. Dans ce cas, elle fournit avant tout un matériau linguistique permettant l'exécution de mesures de similarité et de calculs statistiques fondés sur la récurrence dans un corpus donné. Elle est également la source des données extraites, qu'il s'agisse d'entités d'organisations ou des autres types d'informations prenant la forme d'entités nommées.

Elle est employée pour la constitution de la ressource termino-ontologique (RTO) multi-plans, cette fois en tant que source d'éléments d'informations exprimés linguistiquement. Elle permet

de structurer ces éléments et de les agencer au sein de la RTO. Les facettes renvoyant aux informations contextuelles utilisent la ressource endogène pour extraire le contexte de production de chaque document. Les genres particuliers des documents sur lesquels se fonde la RTO, les brevets et les articles scientifiques, permettent de délimiter de manière fiable les instances des quatre plans intégrant ces informations. Ces facettes font appel aux métadonnées, qui isolent et identifient les entités nommées par type de façon relativement sûre. Nos traitements peuvent alors tirer parti de cette délimitation inhérente aux données. La facette des thèmes, quant à elle, est constituée grâce aux récurrences de séquences textuelles dans le corpus.

Enfin, au niveau de granularité global, le système d'immersion utilise la RTO, qui est elle-même une ressource endogène puisque comme nous venons de le mentionner, elle contient des informations venant de l'ensemble documentaire à explorer. Sa structuration permet la projection de dimensions endogènes dans le système, qui ont la forme de listes ordonnées ou non, hiérarchisées ou non, en fonction de leur type. A leur tour, elles permettent une appropriation des documents en fonction de critères choisis par l'utilisateur.

L'apport des ressources endogènes se situe également à un autre niveau, dépassant celui d'un ensemble documentaire particulier. En effet, les calculs effectués par les processus endogènes, une fois validés par l'utilisateur grâce aux processus anthropogènes (voir le chapitre 8 page 317), permettent de mémoriser des choix, des décisions et donc des informations reconnues comme pertinentes. En tant que telles, ces ressources issues de traitements endogènes et anthropogènes constituent par la suite des ressources exogènes, pour les études qui seront traitées et analysées par la suite.

Elles s'ajoutent donc aux ressources exogènes constituées de manière purement externe à nos processus.

Chapitre 7

Les processus et ressources exogènes : l'exploitation de données externes à chaque palier d'analyse

Sommaire

Dans le chapitre précédent, nous avons traité les processus endogènes permettant de traiter les données et de construire et faire fonctionner le système d'immersion. A chaque degré de granularité, les processus exogènes apportent de l'information, fournissent des clés d'interprétation et aident à la décision.

Dans le présent chapitre, nous abordons les méthodes exogènes, et décrivons comment elles ont été conçues, et ce qu'elles apportent à notre travail, en complémentarité avec les deux autres types de méthodes.

Toutes concourent à la mise en place du système final, et sont exogènes par rapport aux données textuelles d'une étude déterminée. En effet, chacun des types de ressources est constitué non pas à l'aide des données de l'étude particulière en cours d'analyse, mais grâce à des traitements antérieurs, et extérieurs, à l'ensemble documentaire de l'étude elle-même.

De manière générale en effet, l'adjectif *exogène* « désigne ce qui vient de l'extérieur, ce qui a son origine en dehors de l'objet, de l'organisme, de l'ensemble ou du système étudié. Par opposition : *endogène* [52] ». Dans le contexte du traitement de l'information ou du traitement des langues naturelles, nous définissons un processus exogène de la manière suivante :

Définition 5. *Un processus exogène trouve l'information nécessaire à son fonctionnement dans des ressources externes, qui sont placées en entrée du processus. Alors que les notions d'exogène et d'endogène s'opposent en théorie, les processus endogènes et les processus exogènes peuvent être complémentaires dans leur usage.*

Nous recensons trois grands types de ressources exogènes :

- les ressources exogènes statiques : ce sont les lexiques, des structures de représentations de connaissances, mais aussi certaines ressources cartographiques, comme des fonds de carte par exemple ; elles représentent des ressources fixes, et sont des entrées aux différents traitements mis en œuvre ;
- les ressources exogènes procédurales : il s'agit dans notre cas du système à base de règles, qui décrit des procédures à appliquer sur des données ;
- les ressources exogènes dynamiques : dans nos travaux, ce sont des algorithmes de représentation cartographique ; elles effectuent des actions sur les données à traiter, souvent à l'aide de ressources statiques, endogènes ou exogènes.

Dans la première section, nous présentons d'abord le système à base de règles, fondé entre autres sur des lexiques, qui a été conçu pour la normalisation des entités nommées. Puis nous décrivons les ressources exogènes sous forme de lexiques exploitées pour la constitution de la ressource termino-ontologique multi-plans. Enfin, nous exposons la façon dont des ressources exogènes peuvent être utilisées pour le fonctionnement du système d'immersion, dans son aspect calculatoire comme dans son aspect visuel.

52. Définition tirée du Lexique des termes de la complexité : `http://www.intelligence-complexite.org/fr/documents/lexique-de-termes-de-la-complexite.html` Page consultée le 25/01/2011.

7.1 Une base de règles pour traiter les entités nommées

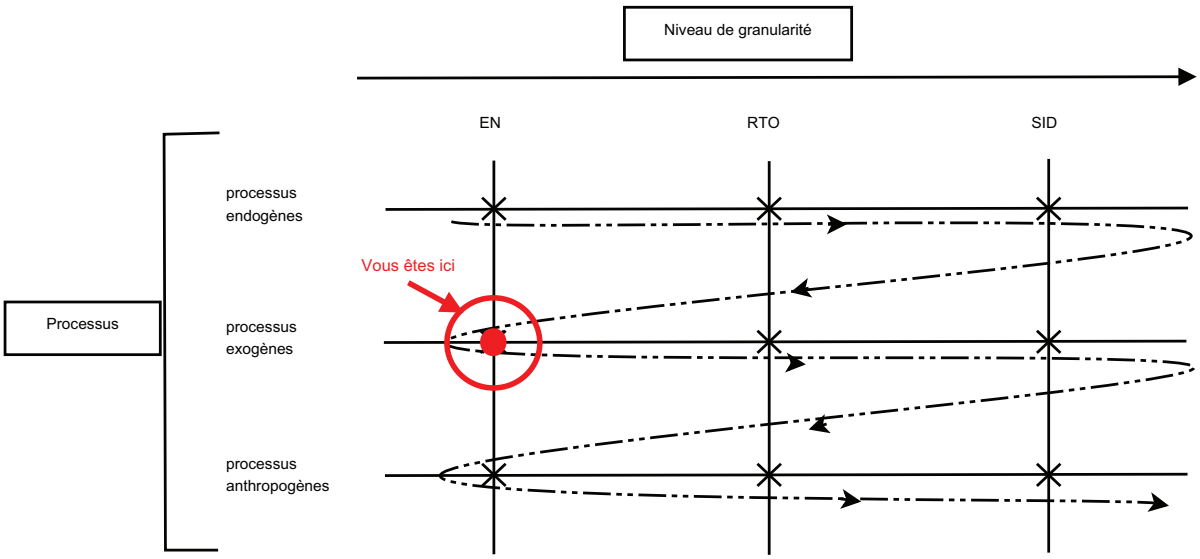

FIGURE 7.1 – Positionnement des traitements exogènes pour la normalisation dans l'ensemble des processus

Dans le chapitre précédent, au niveau de granularité le plus fin que représentent les entités nommées (voir section 6.1 page 159), nous avons présenté des méthodes exploitant les régularités du corpus pour des processus endogènes de normalisation. Dans ce cas, ces régularités étaient exploitées immédiatement pour les calculs réalisés, tirant l'information directement du corpus à traiter. Dans le cadre des processus exogènes pour la normalisation, qui nous intéressent ici (figure 7.1), c'est une autre approche des régularités du corpus qui est adoptée. Celles-ci servent à constituer une base de règles, qui représente alors une ressource exogène dynamique pour traiter le corpus.

Nous présentons d'abord les régularités du corpus pertinentes pour la normalisations des entités nommées, ainsi que l'utilisation qui en est faite. Puis nous décrivons la base de règles que nous avons conçue, et la manière dont elle est intégrée au système de normalisation appliqué aux données des ensembles documentaires.

7.1.1 L'utilisation des régularités du corpus

Nous avons présenté en 4.2 page 114 les données sur lesquelles nous avons travaillé. La normalisation porte sur les entités nommées issues des métadonnées des documents, ceux-ci étant importés depuis des bases de données en ligne vers la base documentaire interne de TKM. Chaque nouvel import de documents est effectué dans le cadre d'une étude précise, et la base documentaire totale respecte ce découpage.

Nous avons également noté que les entités nommées relevées et nécessaires à la constitution du système d'immersion sont de quatre types différents, répartis plus ou moins précisément et distinctement dans des champs dédiés de la base documentaire. Nous rappelons ici les quatre types d'entités nommées en présence :

1. les organisations : les organisations propriétaires des documents ;
2. les lieux : lieux de localisation des organisations publiant les documents ;
3. les dates : dates de publication des documents ;
4. les personnes : individus auteurs des documents.

Il convient cependant de présenter plus précisément les données sur lesquelles nous avons travaillé, ainsi que les régularités que nous en avons tirées.

Nous avons travaillé sur un échantillon de métadonnées issus de l'ensemble documentaire de TKM dans sa totalité, sans distinction entre les différentes études. L'échantillon dont sont issues ces métadonnées est constitué de 6 880 documents. Il représente un moyen d'aborder tous les types de données, toutes études confondues.

Le nombre d'entités par type peut varier en fonction de leur nature. Par exemple, étant donné qu'un document peut être la propriété de plusieurs organisations, en cas de co-publication d'un article, le nombre d'organisations propriétaires est généralement plus important que le nombre de documents. Nous présentons dans le tableau 7.1 page ci-contre la taille des échantillons par type de données.

Type de données	Taille de l'échantillon (en nb d'items)
organisations	13 187
auteurs	22 317
lieux	13 187
dates	6 880
documents	6 880

Table 7.1 – Taille des échantillons par type de données

Le nombre de dates est égal au nombre de documents, puisqu'un document est publié à une date donnée. Nous verrons qu'en fonction du type de document, plusieurs dates peuvent être présentes pour un même document, chacune ayant une sémantique particulière. Cependant, nous avons considéré que les dates ne sont pas entre elles en relation paradigmatique, comme le sont les organisations co-publiantes par exemple, mais se placent sur une linéarité temporelle. Or, il

existe une linéarité temporelle par document. Le nombre de 6 880 correspond donc précisément au nombre de linéarités temporelles.

De la même manière que les organisations, plusieurs auteurs peuvent être impliqués dans la production d'un même document. C'est pourquoi le nombre d'auteurs de l'échantillon est largement supérieur au nombre de documents.

Enfin, le nombre de lieux est égal à celui des organisations, pour la simple raison que lorsqu'ils sont exprimés, les lieux le sont au sein même des noms d'organisations bruts. Ce nombre correspond pour ces deux types à celui qui est constaté avant tout traitement de normalisation sur les unités correspondantes. Nous soulignons cependant qu'en l'occurrence, ce nombre correspond au nombre maximum de lieux potentiels, qui ne correspond pas obligatoirement au nombre réel : puisque les organisations et les lieux sont restitués au sein des mêmes noms, et que la présence des lieux de ces organisations, bien que très courante, n'est pas absolument systématique, il est impossible de prévoir à l'avance le nombre de lieux en présence. Chacun des lieux potentiels comptabilisés inclut virtuellement un pays, une ville et une adresse.

La forme brute et par conséquent la normalisation des deux derniers types d'entités nommées (types 3 et 4 de la liste), les dates et les personnes, ont posé relativement peu de problèmes : la variation des dates est extrêmement limitée, et celle des auteurs est réduite grâce aux règles simples conçues par les ingénieurs R&D de TKM. Les entités nommées d'organisation sont les plus complexes. Enfin, les lieux ont fait l'objet de fait du même échantillonnage que les organisations. Le fait qu'ils soient mêlés à ces entités de type différent les rend plus complexes à appréhender.

Nous décrivons dans ce qui suit les régularités typiques, ou irrégularités typiques, qui ont pu être tirées des corpus pour chaque type d'entités nommées étudié.

7.1.1.1 Les organisations

Nous avons vu en 4.2 page 114 que les organisations étaient les entités nommées les plus élaborées de nos données. Les noms d'organisations tels qu'ils sont importés sont en effet, d'un point de vue quantitatif, les chaînes de caractères les plus longues, et du point de vue référentiel, les plus riches, dans la mesure où souvent, ainsi que nous l'avons signalé dans la section 6.1 page 159, elles désignent des structures complexes du monde réel.

Des régularités se dégagent de notre échantillon de 13 187 noms d'organisations bruts. Tout d'abord, du point de vue de la forme, des particularités d'ordre syntaxique ressortent de nos observations. Les noms d'organisations bruts sont issus de champs dédiés d'une base de données, eux-mêmes remplis à partir des en-têtes des documents. S'il s'agit bien de données langagières, elles ne relèvent cependant pas du texte au sens habituellement entendu. En effet, les séquences de mots ainsi extraites sont courtes, et ne sont pas à proprement parler des phrases. Ce sont toutes, du point de vue syntaxique, des formes nominales, qu'il s'agisse de noms propres ou de

descriptions définies (4.2.1.2 page 115) sous forme de syntagmes réduits. De fait, la structure syntaxique est généralement moins riche que dans du texte. Une syntaxe réduite n'en est pas moins présente. Elle est stéréotypée, dans le sens où les mêmes structures reviennent de manière régulière.

En nous fondant entre autres sur cette syntaxe, nous pouvons dégager trois plans sur lesquels les organisations se distinguent les unes des autres. Ces trois plans sont des tendances, et ne s'appliquent pas systématiquement à un type donné d'organisations.

1. La complexité syntaxique : tous les noms d'organisations obéissent à une certaine syntaxe, ainsi que nous venons de la voir. Cependant, certaines structures syntaxiques, même limitées, sont plus complexes que d'autres.
2. La longueur des chaînes : le degré de complexité de la structure est corrélée avec la longueur des chaînes de caractères : plus la structure syntaxique d'un nom d'organisation est complexe, plus la chaîne de caractères est longue [53]. Inversement, les chaînes de caractères courtes reflètent des structures syntaxiques simples.
3. Les identifiants : un grand nombre de noms d'organisations contiennent des mots génériques caractéristiques d'un type d'organisation, que nous nommons pour le moment des identifiants, dans la mesure où ils permettent d'identifier la présence, et souvent le type, d'une entité d'organisation. Le fait qu'un nom d'organisation n'en comporte pas peut lui aussi être significatif.

En 3.2.2.9 page 94, nous avions recensé trois types d'organisations à inclure dans la facette correspondante de la ressource termino-ontologique. Nous rappelons ici ces trois types :

1. les entreprises ;
2. les organismes académiques et/ou publics, que nous nommerons à partir de maintenant, pour faciliter la lecture, les organismes ;
3. les particuliers.

Or, en corrélant ces trois types d'organisations avec les critères formels que nous venons d'énoncer, nous obtenons la répartition représentée dans le tableau 7.2 page précédente.

Les noms d'organismes sont donc, de manière générale, longs, syntaxiquement complexes et contiennent des identifiants d'organismes, tandis que les noms d'entreprises sont souvent courts, syntaxiquement simples et comportent des identifiants d'entreprise. Enfin, les noms de particuliers sont courts, syntaxiquement simples et ne contiennent pas d'identifiants spécifiques.

Un nom d'organisme possède donc souvent une structure syntaxique relativement élaborée, qui révèle la complexité de l'entité du monde réel qu'elle désigne. Il arrive que le nom soit court

53. Nous considérons pour la taille des noms la longueur des noms d'organisations, à l'exclusion de la présence éventuelle d'une adresse. Ces dernières observations valent donc pour la partie du nom brut qui correspond effectivement au nom de l'organisation, et non à sa localisation géographique.

	Organismes	Entreprises	Particuliers
Structure syntaxique	complexe	simple	simple
Taille des chaînes	longue	courte	courte
Identifiants	organismes	entreprises	aucun

TABLE 7.2 – Croisement des critères formels et des types d'organisations

et la structure simple, comme dans le nom :

1. *Université de Toulouse*

Néanmoins, il est courant de se trouver face à des noms tels que :

2. *School of Electrical Engineering and Telecommunications, The University of New South Wales, Sydney, NSW 2052, Australia*
3. *Department of Pediatrics, Steele Memorial Children's Research Center, University of Arizona, 1501 N. Campbell Ave., Tucson, Arizona 85724, USA*
4. *Department of Botany, University of Coimbra Center for Functional Ecology, Coimbra, Portugal*

Pour ces trois derniers exemples, c'est toute une structure hiérarchique qui est restituée. Cette structure est partiellement explicitée, par la présence des virgules fonctionnant comme des délimiteurs dans l'ensemble des noms d'organisations observés. Cependant, si la grande majorité des virgules jouent le rôle de délimiteurs, toutes les structures hiérarchiques ne sont pas forcément délimitées par ce moyen. Par exemple, dans 5, *University of Coimbra* et *Center for Functional Ecology* sont deux unités hiérarchiques différentes où l'une, le centre, « appartient » à l'autre, l'université, et ne sont pas explicitement et « typographiquement » séparées par une virgule. A côté de cette structuration paratactique, des mots grammaticaux sont utilisés pour structurer syntaxiquement ce nom brut, tels que, en l'occurrence, *of* ou *for*.

La structure syntaxique complexe de ces noms justifie leur longueur : au lieu de deux à trois mots graphiques pour un nom d'entreprise par exemple, il n'est pas rare qu'un nom d'organisme ait une longueur d'une dizaine de mots. Le fait que les noms d'organismes soient généralement des descriptions définies, et non des noms propres au sens habituellement entendu en linguistique (voir sous-section 4.2.1.2 page 115), rentre également en jeu ici.

Enfin, des identifiants sont systématiquement présents dans les noms d'organismes. Pour la plupart d'entre eux, ils sont caractéristiques des organismes, par opposition aux entreprises. Nous avons cité plus haut *University*, typique d'un organisme académique, ainsi que *Center* et *Department*, qui sont moins univoques mais restent majoritairement des représentants d'organismes

publics, tout au moins dans notre contexte et dans nos données.

Un nom d'entreprise a quant à lui, la plupart du temps, une structure beaucoup plus simple et courte qu'un nom d'organisme. Nous présentons ici quelques exemples :

5. *RENAULT S.A.S., 13-15, quai Le Gallo, F-92100 Boulogne Billancourt (FR)*
6. *Xerox Corporation, Norwalk CT [US]*
7. *WYETH CORP*
8. *MERCK*

Les chaînes de caractères sont beaucoup plus courtes que celles des noms d'organismes. Rappelons que les adresses et localisations présentes dans les noms d'organisations ne sont pas prises en compte dans la taille des chaînes de caractères. Dans ces quatre exemples, le nom de l'entreprise est constitué *a minima* du nom de l'entreprise, et pour trois d'entre eux, d'un identifiant d'entreprise. En l'occurrence, les identifiants sont *S.A.S.*, *Corporation* et *CORP*. Ils représentent en fait le statut légal de l'entreprise. Il arrive cependant que des noms d'entreprises apparaissent sans identifiant spécifique, comme c'est le cas en 7 avec l'entreprise Merck.

Quoi qu'il en soit, la structure syntaxique est ici inexistante, puisque seul apparaît le nom propre, suivi le cas échéant de l'identifiant. Ce dernier est presque toujours propre aux noms d'entreprises. Ici, *CORP* et *Corporation* sont deux variantes d'un même identifiant.

D'autres noms d'entreprises peuvent être plus longs, sans toutefois atteindre la taille des noms d'organismes :

9. *Tuoyin Digital Technology Co., LTD.*

Cependant, là encore, le nom d'entreprise se divise en nom de l'organisation d'une part, et identifiants d'entreprises d'autre part. Certaines entreprises sont détaillées, mais ce cas de figure est minoritaire. D'autre part, même dans ce cas, le nom ne comporte généralement que deux niveaux de hiérarchie :

10. *HAPTION S.A., Service de Robotique Interactive, Route de Laval, 53210 Soulgé sur Ouette, France*

Enfin, certaines filiales de maisons-mères donnent le nom de leur filiale, qui permet de les distinguer des autres :

11. *Saint-Gobain Glass France*
12. *Saint-Gobain Technical Fabrics Europe*

Les noms de particuliers, enfin, sont des noms propres de personnes. Ils sont placés de fait dans les noms d'organisations, et ce uniquement quand le document est un brevet, lorsque celui-ci a été déposé par un individu en son nom propre, et non par une organisation. Ils fonctionnent

sur un modèle classique de *Prénom(s) + Nom* ou *Nom + Prénom(s)*, en fonction de la manière dont les données ont été saisies lors de la publication du document, où le(s) prénom(s) peuvent être représentés par son (leurs) initiale(s). Nous présentons ici quelques exemples :

13. Cachin, Franck, 84 Rue Bayard, 75015 Paris (FR) [54]

14. Kiselyov Alexander

15. Xing Weifan

Comme pour les organismes et les entreprises, nous ne considérons pas les adresses dans le nom d'organisation. La virgule peut là encore apparaître comme un délimiteur entre nom de famille et prénom (9), sans que cette dernière soit présente de manière suffisamment récurrente pour en tirer une tendance fiable. Par ailleurs, l'ordre des nom et prénom(s) est souvent aléatoire. Le problème des noms de personnes est complexe du point de vue du traitement des entités nommées. L'ordre des composants d'un nom, les variantes que représentent un prénom développé et son intiale, etc., sont des écueils à leur identification et à leur normalisation, d'autant plus si aucun identifiant n'est présent. Cependant, dans le cadre de cette thèse, ce problème n'a pas été abordé, en particulier en raison de la faible fréquence de ces noms en tant qu'organisations et de leur faible intérêt pour les analystes de TKM.

7.1.1.2 Les lieux

Ainsi que nous l'avons signalé plus haut, les lieux, qu'il s'agisse du pays, de la ville ou de l'adresse complète, apparaissent dans le champ dédié au nom d'organisation. La plupart du temps, le pays est mentionné : de manière générale, sur notre corpus, 75,5% des noms d'organisations contiennent leur pays d'origine. Il existe en revanche une grande disparité entre les types de publications. Dans notre échantillon de 13187 organisations, seuls 52,91% des noms d'organisations propriétaires d'un brevet spécifient explicitement leur pays d'origine, contre 98,1% des noms d'organisations issus des articles scientifiques.

L'information la plus cruciale pour TKM est celle du pays de l'organisation. La ville et l'adresse précise sont utiles, mais moins importantes que le pays qui permettent de positionner les pays entre eux sur un domaine donné.

Lorsqu'ils sont présents, les pays apparaissent presque systématiquement à la fin du nom de l'organisation. La seule exception vaut pour les brevets, où lorsque le pays n'est pas spécifié en tant que tel, il peut toutefois être cité dans le nom même de l'organisation, comme *Japan Polypropylene Corp* où *Japan* fait partie intégrante du nom d'organisation, et non de l'adresse elle-même. De manière générale, le pays peut être indiqué par son nom complet, comme *France*,

54. Cet exemple a été construit à partir de noms effectivement trouvés dans nos données. Nous n'avons pas souhaité faire apparaître les nom et adresse précise de personnes existantes, dans un souci de respect de la vie privée. Cependant, nous avons obéi strictement à la structure rencontrée, ne changeant que les noms de personne et de rue. Ce nom est le seul à avoir été construit : tous les autres sont des cas attestés.

ou par son isocode, comme *FR*. En pratique, cette différence est là encore régulièrement imputable à la source des documents d'origine, source qui est directement liée au type des publications : les isocodes sont régulièrement utilisés pour les brevets, tandis que les noms complets se trouvent la plupart du temps dans les articles scientifiques.

D'autre part, la présence du pays (sous forme de nom complet ou d'isocode) n'implique pas systématiquement que l'adresse complète sera présente. En revanche, en cas de présence d'un numéro et d'un nom de rue, cette adresse est systématiquement complétée par la ville, mais pas toujours par le pays. la réciproque n'est pas toujours vraie : la présence d'une ville n'implique pas forcément celle du numéro et nom de rue. Par ailleurs, les informations de lieux en général, et les adresses complètes en particulier, sont plus susceptibles d'être trouvées dans les en-têtes d'articles scientifiques que dans ceux des brevets.

Enfin, si l'adresse complète est présente, elle l'est systématiquement sous la forme :

numéro et nom de rue, ville, pays.

Les virgules ont été placées ici pour le confort de la lecture, mais ne sont pas systématiquement présentes dans nos données. D'autre part, le cas échéant, l'état ou la province de l'organisation peut également être présente, là encore sous la forme de son nom complet ou de son isocode. Le code postal peut lui aussi être indiqué, avant ou après la ville.

7.1.1.3 Les dates et les personnes : entités nommées ne nécessitant pas le recours à des règles de déduction

Les dates La forme des dates est très régulière, du moins dans la version qui en est intégrée à la base documentaire interne au moment de l'import de documents. Elles font en effet l'objet d'un nettoyage rapide. Pour les brevets comme pour les articles, l'année est fournie, en quatre chiffres. Ce sont donc, sans grande surprise, des dates au format *2011*, *1979*, etc.

Pour les brevets uniquement, des dates détaillées sont également fournies. Cela s'explique par l'importance de ces dernières pour établir l'antériorité pour un brevet, et donc pour attribuer la propriété d'une invention à un inventeur donné du point de vue légal. Toutes les dates détaillées ont la même forme : *aaaammjj*.

Il existe des dates de demande, de publication et de priorité. Nous prenons en compte pour l'instant seulement les dates de publication. Il peut y avoir plusieurs dates de publication, si le document a été modifié après la précédente publication. Dans ce cas, les dates ont la forme suivante : *aaaammjj ;aaaammjj ;aaaammjj*. La présence du *;* comme délimiteur est systématique.

Les personnes Les entités nommées de personnes listent les auteurs, en tant qu'individus, qui ont écrit un article scientifique ou un brevet. En réalité, les entités nommées « auteurs » rassemblent donc les auteurs et les inventeurs.

Selon les sources de données, la forme de ces entités varie dans les documents originaux. Pour les brevets, les noms d'inventeurs ont généralement la forme : *WASCHEUL MICHAEL ; LELASSEUX XAVIER*. Les noms des différents auteurs sont donc séparés par des points-virgules, qui représentent des délimiteurs fiables. Ces noms sont déjà standardisés grâce à une règle générale définie par les ingénieurs R&D de TKM, que nous énonçons comme suit :

La chaîne de caractères est découpée sur les points-virgules ; dans chacune des suites obtenues, la première chaîne de caractères alphabétiques, jusqu'au premier espace, est identifiée comme le nom de famille de l'auteur ; la chaîne de caractères restante est identifiée comme le prénom de l'auteur.

Pour les articles scientifiques, les noms d'inventeurs originaux ont la forme suivante :

16. Viana, M.a b , Ribet, J.a , Rodriguez, F.a , Chulia, D.a

Là encore, un délimiteur fiable permet de distinguer plusieurs noms d'auteurs : il ne s'agit plus du point virgule, mais de la séquence [_ ,_], où « _ » représente une espace. Le nom de famille est séparé du (des) prénom(s) ou de son (leur) initiale(s) par une virgule, non entourée d'espaces. Enfin, la lettre minuscule après la dernière initiale du prénom est la note renvoyant à l'affiliation de l'auteur, et doit être éliminée. Au moment de l'import, la règle permettant de les normaliser est donc :

La chaîne de caractères est découpée en différents individus sur les suites _ ,_ ; dans chacune des suites obtenues, la chaîne de caractères jusqu'à la virgule est identifiée comme le nom de famille de l'auteur ; la chaîne de caractères suivant cette virgule est identifiée comme le(s) prénom(s) de l'auteur ou son (ses) initiale(s), à l'exclusion de la dernière lettre, en minuscule, qui est éliminée.

Dans les données dont nous disposons, les noms des individus sont donc déjà identifiés, et pour un nom d'individu donné, le nom de famille d'une part et le(s) prénom(s) d'autre part sont distingués et placés dans des champs séparés de la base interne.

A l'issue de la standardisation, les données comportent encore du bruit. Cependant, ainsi que nous l'avons énoncé page 246, nous ne traitons pas spécifiquement, dans le cadre de cette thèse, le problème complexe des entités nommées de personnes. Par ailleurs, les noms de personnes pour les auteurs sont de manière générale moins soumis à variation, notamment en ce qui concerne l'ordre des composants d'un nom (prénom(s) et nom de famille), que les noms de particuliers en tant qu'organisations. Enfin, la standardisation telle qu'elle est effectuée aujourd'hui est la plupart du temps jugée satisfaisante par les analystes.

Même si ces quatre types d'entités nommées sont utilisés comme autant de points d'entrée dans le système d'immersion que nous cherchons à construire, il apparaît, à la lumière des descriptions que nous venons de faire des types d'entités, que l'élément central reste l'organisation. C'est pourquoi, dans ce qui suit, nous nous intéressons particulièrement à ce type

d'entités nommées, ainsi que, dans une moindre mesure, aux lieux, puisqu'ils apparaissent dans les mêmes séquences textuelles que les organisations.

7.1.1.4 De la régularité dans la variation, de la variation dans la régularité

De manière générale, la forme des noms est soumise à variation, au sein d'un même type d'organisations, voire pour une même organisation publiant plusieurs documents. L'une des causes de cette variation est le caractère multilingue des données. Les cinq langues les plus courantes que nous avons relevées sont l'anglais, le français, l'allemand, l'espagnol et l'italien. Les organisations japonaises publiant des articles ou déposant des brevets sont nombreuses, mais la plupart du temps, leur nom est traduit en anglais. Les noms d'organisations saisis en russe ou en chinois - transcrits en alphabet latin - sont de plus en plus nombreux, mais restent cependant minoritaires. La langue la plus courante reste toutefois l'anglais, puisqu'à l'image des organisations japonaises, un certain nombre de noms d'organisations non anglo-saxonnes sont traduits, parfois approximativement, dans cette langue.

Cet aspect multilingue des données doit être pris en compte, dans la mesure où la variation en est augmentée, et ce de plusieurs manières. D'une part, les identifiants changent de forme d'une langue à l'autre par traduction, voire par utilisation d'équivalents de référents. Par exemple, alors qu'en France le sigle *SARL* sera utilisé pour désigner une entité juridique à responsabilité limitée, les pays anglo-saxons emploieront le sigle *Ltd.* ou *LLC*. Si les deux référents ne sont pas strictement égaux, ils sont des équivalents d'un système juridique à un autre. Enfin, une même maison-mère peut posséder des filiales portant le même nom d'un pays à l'autre, mais avec un statut légal différent. Ainsi, elle peut avoir un statut de SARL en France et un statut de compagnie *Incorporated* aux Etats-Unis. Nous nous trouvons face à des formes telles que :

17. *Laboratoire GREYC, Université de Caen, 14032 Caen Cedex, France*
18. *GREYC Lab, University of Caen, Caen, France*
19. *Pfizer Ltd.*
20. *Eatops SARL*
21. *Coventor Inc., 625 Mt Auburn Street, Cambridge, MA 01778, United States*
22. *Coventor Sarl, 3 Ave. du Quebec, Villebon Sur Yvette 91140, France*

La variété des langues entraîne donc une variété d'identifiants pourtant équivalents, du côté des organismes comme pour les entreprises. Or, pour envisager une normalisation correcte, ces identifiants équivalents mais formellement différents doivent être identifiés, de manière à établir des correspondances entre variantes d'organisations.

Il est d'autant plus nécessaire de maîtriser cette variation entre identifiants que ces derniers sont nombreux, pour les organismes comme pour les entreprises. Nous avons recensé 37

identifiants d'entreprises différents exprimant des statuts légaux, à l'image de *Inc.* ou *SARL*, et 34 identifiants d'organismes, comme *Université*, *Laboratoire* ou *Département*. Ces chiffres ne prennent pas en compte les variantes possibles pour chacun des identifiants. Les identifiants d'organismes rassemblent environ 140 variantes, qu'elles soient dues à leur traduction, leur abréviation, ou à des conventions de notation concernant la présence ou l'absence du point (« . »). Les identifiants d'entreprises rassemblent quant à eux approximativement 110 variantes, pour les mêmes raisons [55].

Nous avons expliqué que la structure syntaxique et paratactique des noms d'organisations reflétait la plupart du temps la structure hiérarchique des organisations désignées. Cependant, l'ordre dans lequel sont agencés les éléments de cette hiérarchie au sein des noms varie d'un nom d'organisation à un autre :

23. Lab. Spectrometrie Masse Bio-O., Université Louis Pasteur, 25 rue Becquerel, 67087 Strasbourg Cedex 2, France

24. Université Pierre et Marie Curie, Laboratoire de Physiologie Cellulaire et Moléculaire des Plantes, Le Raphaël, 3 R. Galilée, 94200, Ivry-sur-Seine, France

Dans les deux cas, une université et un laboratoire sont présents, et dans les deux cas, l'université englobe le laboratoire. Pourtant, d'un nom à un autre, l'ordre linéaire de présentation est inversé.

La hiérarchie exprimée dans les noms d'organisations est soumise à un autre type de variation. Cette hiérarchie est restituée dans des noms aux structures qui varient peu, nous l'avons vu : intuitivement, nous pourrions supposer qu'un identifiant d'organisme se place toujours au même niveau par rapport aux autres identifiants dans la hiérarchie, et donc qu'à une rigidité syntaxique correspondrait une rigidité de référent, de désignation. Cela se vérifie effectivement pour les laboratoires et les universités par exemple : lorsqu'un identifiant d'université et un identifiant de laboratoire sont tous deux présents dans un même nom d'organisation, alors le laboratoire est toujours systématiquement inférieur hiérarchiquement à l'université : celle-ci inclut celui-là. Mais cet ordre hiérarchique n'est pas toujours aussi tranché, et peut varier selon les noms d'organisations. Le sens du lien d'inclusion entre un institut et un centre, par exemple, est beaucoup plus aléatoire. Ainsi des deux organisations suivantes [56] :

25. National Cancer Center Research Institute

26. Brain Tumor Research Centre, Montreal Neurological Institute

Dans la première d'entre elles, le *National Cancer Center* inclut le *Research Institute*. Dans la deuxième, c'est le *Montreal Neurological Institute* qui inclut le *Brain Tumor Research Centre*.

55. La liste des identifiants recensés ainsi que l'ensemble de leurs variantes est présentée en annexe

56. Nous avons raccourci les noms d'organisations bruts dans les deux exemples qui suivent, pour plus de lisibilité.

Pour établir ce fait de source sûre, nous avons dû aller vérifier la structure de ces organisations sur leur site internet respectif. Ainsi, sans connaissances du monde, et uniquement sur la base d'une liste hiérarchisée des amorces, il est difficile de décider l'ordre au cas par cas. Or, nous avons précisé plus haut que notre objectif était de normaliser ces données sans ressources de type dictionnaire importantes, pour des raisons de coût et d'exhaustivité.

L'ordre hiérarchique des différentes organisations liées les unes aux autres dans un même nom n'est donc pas parfaitement prédictible : la diversité de la hiérarchie des entités du monde réel, et en particulier des organismes, se répercute sur les expressions linguistiques qui les désignent.

Malgré ces divergences sur certains des identifiants d'organismes, il est possible d'en tirer une hiérarchie « idéale », canonique, à partir de nos observations sur l'échantillon, avec des invariants, comme *Université*, et des éléments variables, comme *Centre* et *Institut*[57].

Enfin, la variation vient aussi, nous l'avons vu dans la section 6.1 page 159, des erreurs de typographie et des fautes d'orthographe présentes dans les noms d'organisations. Il est difficile de tirer une régularité quelconque de ces dernières, puisqu'elles relèvent d'accidents et non d'un raisonnement.

Nous venons de lister les régularités et les variations présentes dans les données de notre échantillon, et que nous étendons à la totalité des données à traiter[58]. Si les dates et les noms d'auteurs ont une forme très régulière, les noms de lieux et surtout les noms d'organisations sont en revanche soumis à d'importantes variations. Cependant, au sein même de la variation se dégagent des régularités qu'il est possible d'exploiter pour leur normalisation. Dans la sous-section suivante, nous présentons brièvement le travail qui a été effectué sur les dates et les noms d'auteurs, puis nous attachons à développer les traitements qui ont été appliqués aux noms de lieux et d'organisations afin de les standardiser, et de normaliser les entités nommées correspondantes.

7.1.2 Système à base de règles symboliques

Les types d'entités nommées que nous venons de décrire sont l'expression d'une grande partie des types d'informations nécessaires à la constitution de la ressource termino-ontologique, et par là à la mise en place du système d'immersion. Or, l'importance de la variation apparaissant dans ces entités nommées, et en particulier dans les noms d'organisations et de lieux, en fait des données dont il est difficile de tirer parti en l'état. Par conséquent, leur intégration à un quelconque système s'appuyant sur ces dernières nécessite, au préalable, leur normalisation, de

57. Cette hiérarchie canonique est disponible en annexe.

58. A condition que les sources de données ne varient pas, ou, le cas échéant, que les conventions de notation soient sensiblement les mêmes. Durant les trois années qu'ont duré ces travaux, les sources de données ont peu varié, et les conventions de notation étaient comparables d'une source à l'autre.

manière à lisser ces variations. Si nous prenons l'exemple simple des entités nommées d'organisations *Université de Toulouse* et *Univeristy of Toulouse*, leur normalisation doit permettre leur identification comme une seule et même organisation.

7.1.2.1 Hypothèses

Nous venons de voir que le corpus portait des régularités assez nettes, et ce même dans la variation. Celles-ci sont exploitables par des méthodes endogènes. De manière générale, les processus fondés sur ces dernières font appel à des données externes au corpus à traiter, qu'il s'agisse de dictionnaires, de lexiques, et/ou de règles. Concernant nos travaux, la plupart des régularités repérées manuellement peuvent être utilisées pour l'écriture de règles intégrées à un système de normalisation, permettant de les repérer automatiquement et à grande échelle, et ainsi de corriger la variation.

Nous formulons donc l'hypothèse suivante :

Hypothèse II.10. *Une méthode exogène fondée sur un système à base de règles peut être utilisée pour exploiter les régularités du corpus, avec pour objectif de normaliser les entités nommées impliquées dans les informations nécessaires à la constitution du système d'immersion.*

Les systèmes fondés sur des règles pour le traitement des entités nommées Les systèmes fondés sur des règles sont couramment utilisés pour la reconnaissance d'entités nommées. Bien que notre objectif ne soit pas à proprement parler la reconnaissance des entités nommées, mais bien leur normalisation, notre système doit toutefois entre autres identifier les composants et les limites de ces entités avant de les normaliser. Certains principes de la reconnaissance d'entités nommées peuvent donc être utilisés également pour leur normalisation.

Les « systèmes fondés sur une base de règles » [Poibeau, 2001] pour la reconnaissance d'entités nommées sont fondés, comme leur nom l'indique, sur un ensemble de règles permettant d'identifier les entités nommées dans du texte. La plupart du temps, les règles sont écrites à la main, et décrivent la forme et la structure des entités de manière à les repérer [59]. Ces règles peuvent se fonder sur des indices lexicaux, syntaxiques, etc. Les textes analysés peuvent avoir été préalablement étiquetés ou être encore dans leur forme brute. Le cas échéant, l'étiquetage peut porter sur les propriétés grammaticales et/ou morphologiques des mots. D'autre part, la majeure partie des systèmes emploient des dictionnaires de noms propres dont les entrées sont

59. Ces systèmes, bien qu'ils soient fondés sur un traitement humain préalable, ne sont pas à considérer comme des systèmes anthropogènes : ici, le concepteur ou le constructeur des règles doit être distingué de l'utilisateur final du système (si le concepteur des règles peut en pratique être également l'utilisateur final du système, ces rôles doivent être dissociés). Le concepteur intervient et apporte une ressource, mais le caractère anthropogène serait avéré seulement dans le cas où l'utilisateur final, en tant qu'expert, pouvait enrichir lui-même les règles du système et apporter ses propres connaissances. Voir à ce sujet le chapitre 8.

utilisées dans les règles. Pour présenter brièvement certains de ces systèmes, nous nous fondons sur l'état de l'art établi par [Zaghouani, 2009], au cours de ses travaux sur la reconnaissance des entités nommées en langue arabe.

Le système FUNES [Coates-Stephens, 1993] utilise des règles syntaxiques pour décrire la structure du co-texte des noms propres anglais, répartis par types. Les règles sont écrites sous la forme d'expressions régulières. Il emploie également un lexique de 500 formes verbales et de 2 000 formes nominales.

[Karkaletsis et al., 1999] ont développé sur la plateforme GATE le système GIE (*Greek Information Extraction*), dédié aux entités nommées en langue grecque [Cunningham et al., 1996]. Les textes sont d'abord segmentés et étiquetés, et la reconnaissance des entités nommées en elle-même est fondée sur deux modules. Les entités sont d'abord recherchées à l'aide de dictionnaires de noms propres, puis des grammaires locales, développées à la main, sont appliquées pour repérer les entités non présentes dans les dictionnaires. Ce système peut faire appel à un module d'apprentissage automatique, puisqu'un dernier module permet de créer et d'entraîner des données pour créer automatiquement de nouvelles règles de repérage.

Le système SPRACH-R [Renals et al., 1999] est fondé sur des automates à états finis et sur des dictionnaires de noms propres. Le module de reconnaissance compare grâce à des règles sous forme d'automates les mots des phrases aux entrées des dictionnaires. Des grammaires locales permettent d'attribuer un type - organisation, lieu ou personne - à chaque entité ainsi identifiée.

Enfin, le système Nominator de [Wacholder et al., 1997] utilise un ensemble d'indices textuels pour la définition de ses règles. Ces indices sont les majuscules, des mots-clés et la ponctuation. La particularité de ce système est qu'il n'utilise pas d'informations syntaxiques, et des ressources lexicales limitées à une courte liste de noms propres. Selon cet auteur, le fait de limiter l'emploi de ressources lexicales externes permet d'augmenter la robustesse et la rapidité de son système.

L'un des avantages reconnus pour ce type de système réside dans le fait que les connaissances linguistiques mobilisées sont facilement accessibles et lisibles par l'humain, puisqu'elles sont présentes dans des listes. De fait, elles sont modifiables facilement, et peuvent être enrichies régulièrement sans difficulté.

En revanche, un certain nombre d'inconvénients sont avancés. Tout d'abord, les règles sont écrites initialement pour un certain type de textes et pour une ou des langues données, et peuvent donc ne pas être adaptées à d'autres types de productions textuelles ou à d'autres langues. Le cas échéant, il est donc nécessaire de reprendre et de modifier les règles, et il en va de même pour les lexiques utilisés en entrée. Par ailleurs, seuls les types d'entités prévus par les règles sont repérés. Dans le cas particulier du système Nominator, qui utilise peu de ressources lexicales au profit d'indices tirés du texte, les entités ne sont pas reconnues si lesdits indices ne sont pas présents dans les textes.

De plus, certaines règles peuvent être extrêmement complexes à écrire, afin de parvenir à

extraire le plus grand nombre possible d'entités nommées. L'approche par règles peut donc être coûteuse, tout en ayant une couverture et une transportabilité non optimale.

L'approche par apprentissage pour les systèmes de reconnaissance d'entités nommées est souvent présentée dans la littérature comme l'alternative aux systèmes fondés sur des règles [Poibeau, 2001]. Dans cette approche, le système apprend un modèle d'annotation à partir d'un corpus étiqueté. Selon ce dernier auteur, « [l]e résultat de la phase d'apprentissage peut être un ensemble de règles, un arbre de décision ou un modèle numérique (proche d'un modèle de langage) ». Les méthodes d'apprentissage peuvent être supervisées, semi-supervisées ou non supervisées, selon le degré d'intervention de l'humain dans le processus. Ces systèmes ont l'avantage majeur de s'adapter au corpus qu'ils doivent traiter, surpassant en cela les méthodes à base de règles construites à la main. Les problèmes d'adaptation à d'autres langues ou d'autres types de textes sont de la sorte éliminés, et la complexité des règles rentre beaucoup moins en jeu pour l'humain. L'inconvénient de ces systèmes est que les ressources sont souvent opaques, non accessibles aux utilisateurs ou concepteurs.

Nous récapitulons, dans le tableau 7.3 page 256, l'ensemble des systèmes que nous avons cités, ainsi que les avantages et inconvénients qu'ils présentent.

Les systèmes fondés sur des règles manuelles présentent donc deux inconvénients majeurs : d'une part, la non transposabilité à d'autres types de textes et à d'autres langues, et d'autre part, la complexité que doivent atteindre certaines règles pour pouvoir repérer le plus possible d'entités nommées. Cependant, les auteurs de ces systèmes s'attachent à détecter les entités nommées dans du texte, et non à les normaliser dans des éléments d'en-têtes de documents. Ainsi, il convient d'examiner ces inconvénients du point de vue de la normalisation sur des éléments textuels courts, et de voir s'ils valent également dans ce contexte.

Le problème de la non transposabilité d'un système fondé sur des règles pour la normalisation se pose peu dans notre contexte. Les données à traiter et normaliser sont en effet très formatées, d'abord du fait de la nature de ces données : elles sont des extraits d'en-têtes de documents, qui obéissent à des conventions de saisie relativement uniformes et comportent souvent les mêmes types d'informations, et suivent donc à une structure peu soumise à la variation **dans sa globalité**. D'autre part, elles sont issues de brevets et articles scientifiques, documents au genre très contraint. De fait, nous pouvons supposer que des règles écrites pour ce type de tâche et pour ce type de données pourront s'adapter à un grand nombre de systèmes visant à normaliser des données issues de documents de ce type. L'aspect multilingue des données à traiter pourrait enfin constituer un obstacle. Néanmoins, puisque les en-têtes des documents comportent souvent le même type de lexique, encore une fois en raison d'une certaine rigidité, le recensement d'indices nécessaires à l'écriture des règles est envisageable.

Auteur(s)	Méthode/ Système	Ressources	Avantages	Inconvénients
[Coates-Stephen, 1993]	FUNES	Règles syntaxiques Lexique verbal et lexique nominal	Enrichissement aisé des règles et du lexique par l'humain	Règles adaptées à un seul corpus Complexité potentielle des règles à écrire à la main Couverture limitée des lexiques
[Karkaletsis *et al.*, 1999]	GIE	Grammaires locales manuelles Dictionnaires de noms propres Module d'apprentissage optionnel	Enrichissement aisé des règles et du lexique par l'humain	Grammaires adaptées à un seul corpus Complexité potentielle des grammaires locales Couverture limitée des dictionnaires Ne reconnaît que les types d'entités prévus
[Renals *et al.*, 1999]	SPRACH-R	Automates à états finis Dictionnaires de noms propres Grammaires locales	Enrichissement aisé des règles et du lexique par l'humain	Grammaires adaptées à un seul corpus Complexité potentielle des grammaires locales Couverture limitée des dictionnaires Ne reconnaît que les types d'entités prévus
[Wacholder *et al.*, 1997]	Nominator	Règles fondées sur indices textuels (majuscules, ponctuation, mots-clés) Ressources lexicales limitées (une liste courte de noms propres)	Enrichissement aisé des règles et du lexique par l'humain Robustesse Rapidité	Entités non reconnues si les indices prévus ne sont pas présents dans les textes. Ne reconnaît que les types d'entités prévus
multiples	Méthodes par apprentissage	Corpus étiqueté	Adaptation au corpus des règles générées Complexité des règles sans impact sur l'humain	Règles générées opaques pour l'humain

TABLE 7.3 – Récapitulatif des systèmes de reconnaissance d'entités nommées, fondé sur l'état de l'art de [Zaghouani, 2009]

La complexité potentielle des règles à construire, pour les mêmes raisons que celles que nous venons d'énoncer, est largement réduite si les données à traiter sont des en-têtes de documents plutôt que du texte : nous avons expliqué plus haut que la syntaxe des éléments à traiter était limitée, et il en est de même pour le lexique. Non seulement les séquences à traiter sont courtes, mais la nature des informations contenues dans ces séquences étant récurrente, il paraît envisageable de constituer un ensemble de règles sans que certaines atteignent une complexité ingérable.

Les inconvénients souvent attribués aux systèmes fondés sur des règles non acquises automatiquement se posent donc peu dans notre objectif de normalisation. De plus, le fait de pouvoir intervenir par la suite sur les ressources associées au système est quant à lui un avantage dans notre contexte, puisque cela laisse la possibilité aux utilisateurs d'enrichir ces ressources ou de les modifier au besoin.

Au vu de ces éléments, nous partons donc de l'hypothèse suivante :

Hypothèse II.11. *La particularité de notre corpus, constitué de structures syntaxiques non verbales courtes, et de nos objectifs, c'est-à-dire la normalisation des entités nommées, nous permet de constituer un système de règles cohérent et efficace pour cette normalisation.*

En somme, un système fondé sur des règles est donc une solution adaptée à nos données et à nos besoins. Dans ce qui suit, nous présentons notre propre système de règles, puis nous attachons à décrire les règles que nous avons développées, en nous appuyant sur des exemples.

7.1.2.2 Application : un système fondé sur des règles pour la normalisation des entités nommées

La nature et la forme de nos données nous ont permis d'envisager la mise en place d'un système de normalisation des entités nommées fondé sur des règles, sans que cela soit trop coûteux en temps de développement ni en temps de traitement.

Le développement de ressources importantes comme des dictionnaires d'entités nommées aurait pu constituer un dernier obstacle au développement de ce système. En effet, de tels dictionnaires, pour avoir une couverture suffisante, devraient recenser un nombre très important d'entités, et ce pour l'ensemble des types d'entités nommées en présence. Or, si cela reste possible pour les lieux comme les pays par exemple, les noms d'organisations ou de personnes auraient été plus problématiques : ces deux catégories sont des catégories ouvertes, et il n'est pas envisageable de lister la totalité des noms de personnes et d'organisations, d'autant plus que de nouveaux noms de ces deux catégories apparaissent très régulièrement. D'un autre coté, se limiter à une liste des plus grandes organisations et des personnes apparaissant le plus souvent ne serait pas suffisant, et offrirait une couverture trop limitée. En effet, les auteurs d'articles

scientifiques et les inventeurs de brevets sont si nombreux que la tâche semble impossible. De même, en rester aux plus grandes organisations ne serait pas pertinent, puisqu'un grand nombre de petites entreprises déposent des brevets par exemple, dans des domaines très spécialisés.

La solution peut consister à limiter au minimum les ressources de type dictionnaire, et à nous fonder le moins possible sur des lexiques contenant des entités nommées, mais majoritairement sur les indices lexicaux, syntaxiques et de ponctuation que nous avons relevés dans les données. En cela, nous nous rapprochons de la méthode de [Wacholder et al., 1997] pour son système Nominator, que nous avons évoquée ci-dessus. Il évite les dictionnaires de noms propres par l'exploitation des indices que sont les majuscules, les mots clés et la ponctuation.

La particularité de ce système est qu'il n'utilise pas d'informations syntaxiques, et que les ressources lexicales sont limitées à la liste des mots-clés, nommés par les auteurs « special "name words" »(*ibid.*). Ces derniers sont des mots permettant de détecter la présence d'une entité nommée, ainsi que de lui attribuer un type. Il s'agit de ce que nous avons appelé plus haut des identifiants.

Nous n'exploitons pas strictement les mêmes indices que [Wacholder et al., 1997]. En particulier, nous ne pouvons tirer systématiquement parti de la différence de casse, et ce pour deux raisons : d'abord, un certain nombre de noms non normalisés sont saisis entièrement en majuscules ; de plus, même lorsque seules les premières lettres des mots sont en capitalisées, elles le sont en général pour chaque mot plein d'un nom, qu'il s'agisse d'un identifiant ou du nom de l'organisation à proprement parler. La capitalisation est donc, dans notre cas, souvent non discriminante, hormis quelques cas de figure particuliers. En revanche, nous utilisons bien la ponctuation, et particulièrement la virgule (cf. *supra*), ainsi que les identifiants.

Nous utilisons également la structure syntaxique des noms d'organisations, contrairement au système Nominator, mais sans faire appel à un étiquetage systématique. Etant donné le caractère multilingue du corpus, ainsi que la structure nominale des données, étiqueter celles-ci en parties du discours aurait posé des difficultés sans apporter une grande valeur ajoutée. Puisque nous nous trouvons dans le cadre d'une application commerciale, trouver des étiqueteurs en accès libre tels que Tree Tagger [Schmid, 1994] pour cinq langues différentes, et exploitables dans le cadre d'une activité commerciale, aurait été problématique. De plus, les résultats en seraient potentiellement décevants, puisque les données ne suivent pas le schéma classique d'une phrase verbale, à partir duquel Tree Tagger fonctionne.

L'ensemble de nos règles sont donc fondées sur des patrons lexico-syntaxiques, qui exploitent la structure syntaxique limitée mais présente des noms d'organisations, les identifiants d'entités nommées, ainsi que la ponctuation.

Pour atteindre ces objectifs, les ressources exogènes exploitées sont donc les suivantes :

1. le système à base de règles fondées sur des patrons lexico-syntaxiques ;

2. un lexique des identifiants recensés ;
3. un lexique multilingue de mots grammaticaux ;
4. un lexique de pays et villes, pour les entités nommées de lieux, où chaque ville est associée à son pays d'appartenance, et réciproquement.

Alors que la première de ces ressources est une ressource dynamique, les trois suivantes sont des ressources statiques.

Les identifiants sont répartis par le type d'entités nommées qu'ils désignent, puis par sous-types. Ils concernent exclusivement les noms de lieux et les noms d'organisations, puisque les autres types de noms ne contiennent pas d'identifiants spécifiques. Les sous-types sont, pour les organisations, les organismes et les entreprises ; les identifiants de lieux sont en fait des désignateurs d'adresses, comme *avenue*, *Cedex*, *street* ou *park*. Les identifiants servent, nous l'avons dit, à repérer dans les données la présence d'une organisation ainsi que son type. Ainsi, la présence de l'identifiant *université* par exemple permet de savoir qu'une entité nommée d'organisme est présente. Il est à noter que certains identifiants n'ont cependant pas de type précis, et peuvent relever de noms d'entreprises ou d'organismes. Par exemple, *Laboratoire* ou *Département* peuvent indiquer la présence d'organisations des deux types, bien que leur présence dans les organisations de type *organisme* soit plus courante que dans les entreprises. C'est pourquoi ces identifiants sont exploités, mais considérés comme non discriminants quant au type. Ces cas sont néanmoins peu nombreux : il s'agit des identifiants *Laboratoire, Département, Fondation, Society, Service*. Par défaut, ils sont considérés comme des identifiants d'organismes, puisqu'ils relèvent plus souvent de cette catégorie que de de celle des entreprises. Cependant, ils ont un degré de fiabilité moindre par rapport aux identifiants discriminants, et ce caractère est pris en compte lors de certains calculs les utilisant.

Les règles exploitent l'ensemble de ces identifiants, puisque certaines des règles de normalisation sont déclenchées en fonction de la présence d'un identifiant d'un type donné. A partir de maintenant, nous nommons ces identifiants des *amorces*, puisqu'elles amorcent l'application des règles. Notons que pour chaque amorce, les formes développées comme les formes abrégées sont prises en compte. Nous considérons que cette liste est relativement fermée, puisque ces amorces sont en quelque sorte des marqueurs morphologiques qui « signalent » l'instanciation d'un type donné d'entité. A ce titre, si le nombre d'entités nommées est potentiellement infini, celui des marqueurs de type d'entité est lui relativement restreint.

Un classement hiérarchique a été établi pour ces amorces. Il a été constitué grâce à des observations sur des échantillons constitués spécifiquement dans cet objectif. Des ensembles de 100 noms d'organisations par amorce ont été sélectionnés, avec pour contrainte de contenir au moins deux amorces d'organismes différentes en plus de l'amorce testée. De cette façon, nous avons pu observer le comportement en contexte de ces amorces.

Les conclusions que nous avons pu en tirer sont de l'ordre de la tendance, et non du fait systématique. Certaines tendances sont très lourdes, comme l'université qui, lorsqu'elle est présente dans un nom, se place toujours en haut de la hiérarchie ; d'autres sont plus nuancées : l'amorce *Division*, par exemple, se place dans 65% des cas de cooccurrence plus haut que l'amorce *Laboratoire*. Nous avons donc pris le parti de les représenter de la sorte dans notre hiérarchie. Cependant, les 35% de cas où la laboratoire englobera la division ne sont pas pris en compte dans ce classement.

D'autre part, d'autres amorces se positionnent encore moins nettement. C'est ce que nous avons vu plus haut, avec les amorces *Institut* et *Centre*, qui viennent se placer vers le haut de la hiérarchie, mais dont il est difficile de déterminer le sens d'inclusion : en cas d'apparition de ces deux amorces dans un même nom, il est difficile de savoir sans connaissances du monde laquelle est hiérarchiquement supérieure à l'autre (voir les exemples de noms 25 et 26). Ainsi, dans la hiérarchie canonique qui a été constituée à partir de ces observations, et qui permet d'effectuer les classements entre entités liées entre elles hiérarchiquement lors de l'exécution des règles, il existe des zones « floues » dans lesquelles plusieurs amorces peuvent être placées au même niveau [60].

Ce lexique hiérarchisé des amorces prend la forme d'un réseau : il ne s'agit pas strictement d'un arbre, puisqu'une amorce de niveau hiérarchique 3 par exemple peut être fils direct d'une amorce de niveau 1, sans la présence obligatoire d'une amorce de niveau 2. Nous fournissons en figure 7.2 un extrait de ce réseau. La structure en réseau orienté permet une plus grande souplesse qu'une structure en arbre, puisque celui-ci obligerait la présence d'une amorce de niveau 2 entre un niveau 1 et un niveau 3 pour que le lien hiérarchique puisse s'établir.

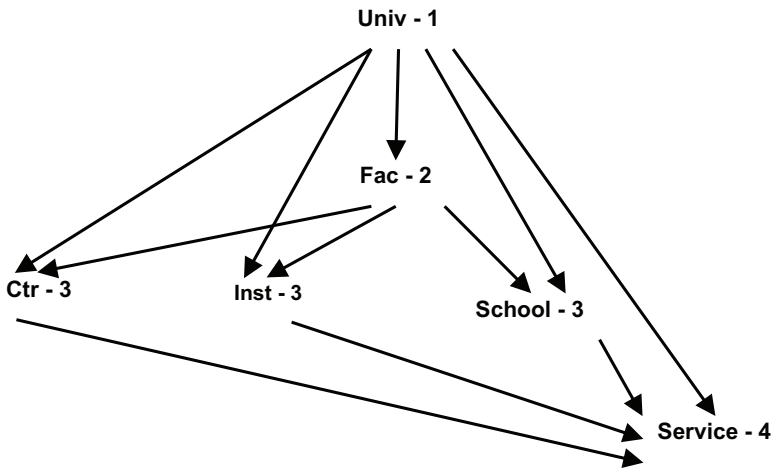

Figure 7.2 – Extrait du réseau hiérarchisé d'amorces

Ce lexique, dans sa forme énumérative ou hiérarchisée, est également utilisé dans les processus

60. La liste hiérarchisée complète de ces amorces est disponible en annexe.

endogènes de normalisation, puisque les amorces sont des mots creux pour le calcul de distance d'édition (voir la sous-section 6.1.1 page 161), et des identifiants hiérarchisés dans les calculs de segmentation (sous-section 6.1.2 page 177). Il représente le seul apport externe de ces traitements.

Le deuxième lexique contient des mots vides au sens de [Tesnière, 1959], ou mots grammaticaux, qui servent à déterminer certaines constructions syntaxiques utiles à la normalisation. Ils permettent en effet, en combinaison avec le lexique d'amorces, de construire des patrons d'extraction et de normalisation utilisés dans les règles. Cette liste regroupe les mots vides utilisés dans les données en cinq langues différentes, de manière à couvrir le plus d'occurrences possible dans le corpus. Les langues les plus présentes sont l'anglais, l'allemand, l'espagnol, le français et l'italien. Notre lexique est constitué d'une quarantaine d'entrées, qui présentent la catégorie grammaticale du mot-outil, le mot-outil lui-même, et enfin sa langue.

Enfin, le lexique des pays et celui des villes sont utilisés pour repérer, dans les noms d'organisations, les entités permettant de localiser plus ou moins précisément les organisations propriétaires des documents.

L'ensemble de ces ressources est majoritairement utilisé pour la normalisation des entités nommées d'organisations et de celles de lieux. Les deux autres catégories d'entités nommées sont moins concernées par ce processus : les auteurs sont déjà pré-traités lors de l'import des données, et ne seront pas retouchées ; quant aux dates, quelques règles simples permettent de les délimiter.

Ces lexiques sont les seules ressources de type dictionnaire que nous utilisons. Nous avons expliqué que nous avons pris le parti de faire appel le moins possible à des ressources de type dictionnaire ou thésaurus, dans la mesure où nous considérons comme impossible, ou pour le moins trop coûteux, d'avoir un dictionnaire de noms propres suffisamment important pour avoir une couverture satisfaisante sur nos données, qui plus est dans des domaines spécialisés (Poibeau, 2001).

Les règles que nous avons mises en place servent plusieurs objectifs, tous concourant à normaliser les données :

1. la détection des entités nommées et de leurs frontières ;
2. l'élimination du bruit, c'est-à-dire des éléments ne correspondant pas à des entités nommées dans notre corpus et pour notre application ;
3. la délimitation des entités nommées multiples pour les noms d'organisations en contenant plus d'une ;
4. la hiérarchisation des entités nommées multiples ;
5. la réécriture des entités nommées identifiées et délimitées pour leur harmonisation.

Détail des règles et exemplification

Les attendus Notre objectif global est de normaliser les entités nommées pour qu'elles correspondent aux attendus définis en amont avec les utilisateurs experts, c'est-à-dire les analystes de TKM. Nous prenons donc en compte leurs attentes, et non une norme de représentation existante telle que les normes AFNOR. Dans la mesure où ces attendus ont été définis en collaboration avec les utilisateurs, et où les attentes individuelles pouvaient varier, une certaine part de subjectivité est inhérente aux attendus eux-mêmes. Malgré tout, nous avons objectivé ceux-ci au maximum, en fixant pour chaque type d'entité nommée une forme standard à atteindre qui réponde au mieux aux besoins des utilisateurs. Nous présentons ici les standards par type d'entités nommées :

Les dates :

- Une date doit avoir la forme : *année-mois-jour*, comme dans *2005-02-27* ;
- En cas de dates multipes, chacune d'entre elles doit être distinguée, et prendre la forme présentée ci-dessus.

Les lieux :

- Les pays doivent être identifiés sous la forme de leur nom ou de leur isocode ;
- Les villes doivent, autant que possible, être repérées, et restituées grâce à leur nom ;
- Les adresses précises, lorsqu'elles sont présentes, doivent être détectées et isolées ; aucune forme standard n'a été établie, puisqu'aucun recoupement n'est opéré sur les adresses précises : elles viennent compléter l'information.
- Lors de leur identification, ces entités de lieux de différents niveaux détectées au sein d'un même nom doivent être liées et hiérarchisées.

Cette hiérarchisations des noms de lieux permet de respecter une partie de notre engagement sémantique (voir la sous-section 3.2.1 page 79). En effet, c'est cette hiérarchisation qui permet parfois de lever une ambiguïté intratypique (sous-section 4.2.2.2 page 132) d'une entité de lieu. Il peut par exemple arriver que les villes *Paris* en France et *Paris* aux Etats-Unis (plus précisément au Texas) soient présentes dans les données. Si aucun lien n'est établi entre ville et pays, alors le nom de cette ville reste ambigu, et il n'y a aucun moyen de le raccrocher au pays correspondant de manière certaine. En revanche, si la relation hiérarchique est restituée, il devient possible de savoir de quel pays un document associé à Paris est issu.

Les organisations : la forme attendue d'une organisation est fonction de son type.

- Les entreprises sont normalisées par la suppression, lorsqu'elles sont présentes, des amorces d'entreprises ; par exemple, *Mitsubishi Corp.* devient *Mitsubishi* ;
- Les noms d'organismes doivent, après leur découpage hiérarchique, avoir une forme différente :
  - Pour les universités, l'amorce doit être réduite à Univ, ce qui permet d'éliminer les variations dues aux différentes variantes utilisées pour Université. De plus, cette amorce est basculée en tête du nom. Ainsi, l'organisation *Southampton University* devient *Univ*

Southampton ;

- Les autres noms d'organismes doivent avoir une forme similaire, à l'exception de l'amorce qui n'a pas à être déplacée : dans certains cas, placer l'amorce en tête de nom gêne la compréhension, et il est donc préférable, pour les utilisateurs, de respecter la structure globale du nom. En revanche, les amorces sont là encore réduites à une forme canonique, afin d'éliminer la variation. Ainsi, pour les amorces *center* et *institute* par exemple, *Riverside Research Institute* devient *Riverside Research Inst*, et *Texas Medical Center* devient *Texas Medical Ctr*.

- Dans tous les noms d'organisations, les mots outils sont supprimés, afin de réduire encore la variation. Un nom comme *Université de Caen* devient *Univ Caen*, et *Medical Center of Central Georgia* est transformé en *Medical Ctr Central Georgia*.

Aucune forme particulière n'a été définie pour les particuliers : d'un point de vue quantitatif, ils restent minoritaires et sont donc peu exploités par les analystes ; par ailleurs, étant donnée leur forme et l'absence d'amorces ou de délimiteurs, il est difficile d'envisager un traitement permettant de normaliser de manière sûre ces noms de personnes, tout en restant peu coûteux. Le rapport entre coût et bénéfice serait trop déséquilibré.

A côté de ces formes standards, le type de chaque organisation - entreprise, organisme ou particulier - doit également être détecté.

les dates Le jeu de règles de normalisation pour les dates se résume à deux règles, puisque les dates prennent toujours la même forme :

1. Lorsque des point-virgules sont présents dans une date « brute », un découpage de la chaîne est réalisé sur ces caractères. Cela permet de distinguer plusieurs dates ;
2. Sur une date donnée, les année, mois et jour sont isolés de manière à obtenir une date à trois niveaux. Dans *20050227* par exemple, l'année *2005*, le mois *02* et le jour *27* sont distingués.

les lieux Les noms de lieux ont des règles de normalisation plus riches que les dates. Cela est en particulier dû au fait que les noms de lieux, qu'il s'agisse des pays, des villes ou des adresses, sont placés dans les noms d'organisations. En premier lieu, il convient donc de repérer ces lieux au sein du nom brut. Pour cela, des règles différentes sont appliquées en fonction du type de lieu à repérer - pays, ville ou adresse précise. Cependant, le point commun entre les règles s'appliquant aux pays et celles s'appliquant aux villes est qu'elles se concentrent en priorité sur la fin du nom d'organisation brut. En effet, les localisations géographiques des organisations, lorsqu'elles sont présentes, se situent très souvent à la fin du nom. Cela peut certes engendrer un silence, mais celui-ci reste faible, tandis que nous limitons le bruit potentiellement engendré par des noms d'organisations pouvant contenir des éléments habituellement identifiés comme lieux.

Pour le repérage des pays, le lexique correspondant, contenant leur nom et isocode, sont projetés sur le nom d'organisation brut. Si le pays est trouvé sous l'une ou l'autre forme, il est sélectionné.

Pour le repérage des villes, le même principe est appliqué.

Enfin, pour le repérage des adresses, des patrons lexico-syntaxiques simples sont construits, à partir d'un lexique d'amorces d'adresses [61], de la ponctuation - très souvent les virgules, et de la présence de nombres. En effet, quand un nombre suivi d'une virgule précède un segment contenant une amorce d'adresse, le nombre en question est très souvent le numéro de voie de l'adresse correspondante.

Nous présentons ici, à titre d'exemple, l'une des règles permettant de détecter une adresse :

$$(ADD) \Rightarrow [ADD] \tag{7.1}$$

Les éléments entre parenthèses renvoient à des amorces, tandis que les éléments entre crochets symbolisent des sections, soit des séquences placées entre deux virgules.

La règle signifie donc que si une amorce d'adresse est trouvée dans une section, alors cette section est une adresse.

Lorsque plusieurs entités de lieux de niveaux différents sont repérées dans un même nom, les listes sont utilisées pour établir les liens qui les unissent, et donc les hiérarchiser :

- lorsque les deux entités sont présentes, la ville et le pays sont associés grâce au lexique à double entrée correspondant. Ils sont de fait hiérarchisés, puisque leur niveau est déduit directement de leur type dans la liste ;
- lorsqu'une adresse a été repérée grâce à une amorce, et qu'un autre niveau d'entité de lieu est présent, elle y est associée en tant que hiérarchiquement inférieure, qu'il s'agisse d'une ville ou d'un pays.

Dans les cas où une adresse précise est détectée, et si aucun pays ni ville n'a été trouvé, alors tout ce qui suit potentiellement l'adresse précise est considéré également comme faisant partie de l'adresse, même en l'absence d'amorce.

Les règles de normalisation des organisations Les règles concernant les organisations sont, nous l'avons exposé, les plus complexes, en raison de leur structure plus élaborée que les autres types d'entités et la richesse des informations véhiculées. Les règles de normalisation se répartissent en quatre groupes, en fonction de leur objectif :

1. les règles de nettoyage et de standardisation préliminaires ;
2. les règles de découpage des noms d'organisations bruts ;
3. les règles de découpage et de classement au sein des segments ;

61. La liste des amorces d'adresses est disponible en annexe.

4. les règles finales de découpage et de réécriture des noms d'organisations.

Type de règles 1 - Nettoyage et standardisation préliminaires

Ces règles ont pour objectif de procéder à une première standardisation, sommaire, des noms d'organisations bruts. Cette standardisation est motivée par le fait que les traitements de normalisation et de structuration sont plus efficaces sur des noms d'organisations débarrassés de certaines variations qui ont un impact sur l'application des règles.

Dans un premier temps, les parasites les plus évidents sont éliminés. Ces parasites sont des éléments présents dans les noms bruts d'organisations qui ne nous servent pas et/ou n'endossent pas, dans notre application, le rôle d'entités nommées. Un jeu de règles de nettoyage préliminaire est donc appliqué à chaque nom d'organisation.

Les règles permettent d'éliminer par exemple des adresses e-mail, certains signes de ponctuation que nous n'exploitons pas, comme les parenthèses, ou des indications externes aux noms d'organisations eux-mêmes qui ont été ajoutées dans les bases de données en ligne. Des indications comme *[machine translation]* par exemple ne sont pas présentes dans les documents d'origine, mais ont été placées dans les noms d'organisations pour signifier que le document associé a été traduit automatiquement.

A titre d'exemple, la règle permettant d'éliminer les adresses e-mail est la suivante :

$$+/@/+ \Rightarrow |Mail| \tag{7.2}$$

Dans la règle formalisée, les $+$ entourent une chaîne de caractères issue des données, et $/@/$ renvoie aux patrons de détection d'adresse mail qui ont été écrits ; les / encadrent une sous-section, c'est-à-dire des éléments extraits d'une section. Nous rappelons qu'une section est une séquence d'un nom d'organisation placée entre deux virgules. Les marques d'adresse mail sont entre autres la présence d'un @, et de formes correspondant à des patrons tels que :

suite_de_caracteres_et_de_points_ou_tirets + @ + suite_de_caracteres.

Elle permet d'éliminer ces adresses e-mail de noms tels que :

27. *Division of Gynecologic Oncology, Department of Obstetrics and Gynecology, University of Kentucky Medical Center, 800 Rose Street, Whitney-Hendrickson Building, Lexington 40536, USA.* ***jowell@gx.net***

Dans un deuxième temps, des règles permettent de donner aux amorces leur forme standardisée, et s'appuient sur la liste des amorces que nous avons constituée. A ce stade, le niveau hiérarchique des amorces n'est pas pris en compte, puisqu'il s'agit uniquement d'une étape de réécriture. De même, le type n'est pas attribué à ce stade en fonction des amorces rencontrées.

Ici, les variantes que sont [*University, Université, Univ.*], par exemple, sont toutes ramenées à leur forme standard UNIV, et les variantes[*Corporation, Corp., Corp*] au standard *CORP*.

La règle utilisée pour cela est la suivante :

$$(Amorce) \Rightarrow (AMORCE) \tag{7.3}$$

Ici, la chaîne *(Amorce)* renvoie à la forme non standardisée d'une amorce, tandis que la version en majuscules représente sa forme standardisée.

Type de règles 2 Découpage des noms d'organisations bruts

Nous avons signalé qu'il était possible de trouver, dans les noms d'organisations bruts, des noms à entités multiples faisant référence à plusieurs organisations distinctes. Dans les cas de co-dépôt de brevets, ou de co-publications d'articles, plusieurs organisations sont présentes dans les entêtes. Si la séparation n'est pas clairement exprimée entre les différentes organisations présentes, le découpage effectué au moment de l'import ne peut être effectué correctement. Auquel cas, nous retrouvons ces organisations au sein d'un même nom brut, et il convient alors de les distinguer.

Le point-virgule est fréquemment utilisé par les scripteurs pour séparer les deux organisations. Ce signe de ponctuation constitue un indice fiable de découpage, bien qu'il ne soit pas systématique. Par ailleurs, le point-virgule est parfois utilisé pour séparer un nom d'organisation de son adresse. Là encore, il est donc un indice qu'il est possible d'exploiter. En somme, dans la plupart des cas, l'utilisation du point virgule est le signe d'une rupture dans un nom d'organisation brut, qu'il indique la co-présence d'organisations distinctes, ou celle d'une organisation et de son adresse.

Ainsi les quatre exemples :

28. *BHA GROUP INC ; E.I. DU PONT DE NEMOURS AND COMPANY*
29. *Centre de Recherche en Acquisition et Traitement de l'Image pour la Santé, CNRS ; INSERM, Lyon, France*
30. *CNRS ; TOTAL SA ; Université de Pau et des Pays de l'Adour*
31. *COLLAGEN MATRIX, INC. ; 509 Commerce Street, Franklin Lakes, NewJersey 07869 (US)*

Utiliser le point-virgule permet donc de déstructurer un nom brut, afin d'en tirer les différents composants, que nous nommons *segments*, plus ou moins indépendants. La règle utilisée pour cela est :

$$+x;y+ \Rightarrow [x]||[y] \tag{7.4}$$

Si une chaîne de caractères contient un point-virgule, alors les deux éléments de la chaîne de part et d'autre du point-virgule sont des segments, symbolisés dans la règle par les *[]*.

Cependant, cette utilisation du point-virgule n'étant pas d'une fiabilité absolue, il convient, après cette segmentation sommaire, de vérifier que les segments ainsi obtenus sont bien destinés

à être séparés, et de quelle façon : s'il s'agit de deux organisations distinctes, alors cette indépendance devra être prise en compte en tant que telle ; si nous sommes face à un nom d'organisation et à son adresse, cette dernière doit être identifiée en tant que lieu et reliée à son organisation ; enfin, il peut arriver que le point-virgule sépare deux organisations liées hiérarchiquement : dans ce dernier cas, le sectionnement par le point-virgule ne doit pas être prise en compte, comme dans l'exemple suivant :

32. *The Research Foundation of State University of New York ; Office of Science and Technology Transfer & Economic, Outreach Intellectual Property Division, UB Technology Incubator, Suite 111, Baird Research Park, 1576 Sweet Home Road, Amherst, NY 14228 (US)*

En l'occurrence, l'organisation *Office of Science and Technology Transfer & Economic*, et les éléments suivants, dépendent bien de la *Research Foundation*, elle-même appartenant à la *State University of New York*. Le point-virgule n'indique donc pas, ici, une rupture entre noms distincts.

C'est pourquoi d'autres indices doivent être utilisés, et en particulier les amorces, afin de découper correctement organisations distinctes et/ou organisations et adresses.

Ici, la liste des amorces est utilisée avec le type associé à chaque amorce, ainsi qu'avec le niveau hiérarchique de chacune.

Le type associé aux amorces permet de distinguer des segments contenant chacun des entités que nous qualifions d'*incompatibles* : par exemple, un segment contenant l'amorce *Univ*, de type *organisme*, et un autre contenant *Corp*, de type *entreprise*, réfèrent forcément à deux entités distinctes, non liées.

C'est le cas de figure que nous trouvons dans le nom :

30. *CNRS ; TOTAL SA ; Université de Pau et des Pays de l'Adour*

Ici, le CNRS et Total d'une part, et Total et l'université d'autre part, sont incompatibles du point de vue du type.

En revanche, un segment contenant l'amorce *Street*, de type *adresse*, face à tout autre segment contenant une amorce d'entreprise ou d'organisme, constitue l'adresse du nom d'organisation.

C'est le cas dans notre exemple :

31. *COLLAGEN MATRIX, INC. ; 509 Commerce Street, Franklin Lakes, NewJersey 07869 (US)*

Dans les cas où les deux amorces sont de types incompatibles *organisme vs. entreprise*, mais où l'amorce d'organisme est recensée comme non discriminante, une règle que nous pouvons qualifier de *contextuelle* est appliquée : le contexte d'apparition de l'amorce non discriminante est pris en compte. Le type par défaut de cette dernière est mis à jour et considéré comme *entreprise*. Or, ces amorces non discriminantes sont placées dans le bas de la hiérarchie, et en tout cas toujours plus bas que les amorces d'entreprises discriminantes. De fait, les deux segments

sont dans ce cas rapprochés et considérés comme une seule organisation à entités liées. Il en est ainsi pour le nom :

33. Hot and Cold Metal-forming ***Laboratory*** *for Lubricants ; Condat* ***SA***

où le laboratoire est une partie de l'enteprise Condat.

Hormis ce cas précis, deux possibilités s'offrent lorsque les types d'amorces sont les mêmes d'un segment à l'autre. Si deux segments contiennent des amorces discriminantes de type *entreprise*, alors ces deux segments sont considérés comme indépendants :

28. BHA Group ***Inc*** *; E.I. Du Pont De Nemours* ***and Company***

En effet, les amorces d'entreprises discriminantes étant toutes de même niveau hiérarchique, la présence de deux d'entre elles dans deux segments différents est souvent le signe d'un co-dépôt, puisqu'une organisation ne peut être incluse dans une autre de même niveau, si ce niveau est déterminé de manière absolue. Cependant, le type seul ne permet pas de trancher quant à l'indépendance de deux organisations contenues dans des segments dans lesquelles les amorces sont toutes de type *organisme*.

Le niveau hiérarchique de ces amorces rentre alors en jeu : si les deux amorces d'organisme sont de niveau 1, c'est-à-dire au niveau le plus haut, alors les deux segments sont considérés comme indépendants, et référant à deux organisations distinctes. Ce sera le cas des amorces *CNRS* et *INSERM* dans (29), et des amorces *Univ* et *CNRS* dans (30).

29. Centre de Recherche en Acquisition et Traitement de l'Image pour la Santé, ***CNRS*** *;* ***INSERM****, Lyon, France*

30. ***CNRS*** *; TOTAL SA ;* ***Univ*** *de Pau et des Pays de l'Adour*

En revanche, si l'une des amorces est de niveau 1, et que l'autre lui est hiérarchiquement inférieure, alors les deux segments sont considérés comme liés hiérarchiquement, et comme référant à une seule organisation contenant plusieurs niveaux d'entités. Ainsi, *Univ* et *Office* sont considérées comme référant à des entités liées :

32. The Research Fdn of State ***Univ*** *of New York ;* ***Office*** *of Science and Technology Transfer & Economic, Outreach Intellectual Property Division, UB Technology Incubator, Suite 111, Baird Research Park, 1576 Sweet Home Road, Amherst, NY 14228 (US)*

Le même raisonnement est appliqué, si deux amorces d'organismes de niveau inférieur à 1 sont comparées. Une amorce *Inst* et une amorce *Ctr* seront donc elles aussi considérées comme liées.

Formellement, et à titre d'exemple, la règle permettant de déterminer que deux segments comportant respectivement *Univ* et *Ctr* sont liés est :

$$\begin{aligned}&[(ORGANISME)\&(NIV1)], [(ORGANISME)\&(NIV2+)]\\ \Rightarrow\ &[(ORGANISME)\&(NIV1)] > [(ORGANISME)\&(NIV2+)]\end{aligned} \tag{7.5}$$

Etant donnés deux segments, si l'un d'entre eux contient une amorce d'organisme de niveau 1 et le second une amorce d'organisme de niveau 2 ou plus, alors le premier inclut hiérarchiquement le second.

De même, il se peut que deux types d'organisations incompatibles, ou qu'une adresse et une organisation, apparaissent dans un nom sans être séparés par un point-virgule mais par une virgule. Dans ce cas, il faut pouvoir séparer ces éléments, de manière à obtenir un découpage fiable. Cela permet de découper des noms comme :

34. Institut Pluridisciplinaire de Recherche Appliquée dans le domaine du génie pétrolier (IPRA), CNRS, Université de Pau et des Pays de l'Adour (UPPA), TOTAL SA

35. RENAULT S.A.S., 13/15 Quai le Gallo, F-92100 BoulogneBillancourt (FR)

Dans le premier nom, l'Institut Pluridisciplinaire de Recherche Appliquée dans le domaine du génie pétrolier, le CNRS, l'Université de Pau et Total SA doivent être distingués comme des organisations différentes. En particulier, le type de l'organisation *Total SA* n'est pas compatible avec le type des autres organisations, qui relèvent de la catégorie des organismes publics et assimilés. Dans le second exemple, la société Renault SAS doit être séparée de son adresse.

Là encore, les amorces, ainsi que leur type et leur niveau hiérarchique, sont exploités, ainsi que la présence des virgules. Si deux parties d'un même nom sont des sections, et ne sont séparées que par une virgule, qu'elles contiennent des amorces de type et/ou de niveaux incompatibles sur le même modèle que ci-dessus, alors elles sont distinguées comme des segments, c'est-à-dire comme si elles avaient été séparées, dans le nom d'organisation brut, par un point virgule.

C'est ce que permet de faire la règle :

$$\begin{aligned}&\{(ORGANISME)\}, \{(ENTREPRISE)\}\\ \Rightarrow\ &[(ORGANISME)]||[(ENTREPRISE)]\end{aligned} \tag{7.6}$$

Etant donnés deux segments, si l'un d'entre eux contient une amorce d'organisme et le second une amorce d'entreprise, alors les deux sections sont séparés et deviennent des segments distincts.

A l'issue de ces traitements, les organisations distinctes sont donc identifiées comme telles, ce qui permet de les traiter indépendamment les unes des autres lors des traitements suivants.

Les adresses repérées sont quant à elles isolées du reste du nom dont elles dépendent, tout en leur restant liées.

Enfin, les organisations liées hiérarchiquement sont rassemblées dans une seule et même séquence, qui sera traitée comme un nom d'organisation à entités hiérarchiques multiples.

Type de règles 3 Découpage et classement des segments

A ce stade, les organisations distinctes ont été séparées, les adresses isolées, et les organisations liées rapprochées. Ainsi, chaque segment arrivant en entrée de ces traitements représente un et un seul nom d'organisation, même s'il est un nom à entités hiérarchiques multiples. Les segments sont conservés dans les données, car ils représentent une étape intermédiaire exploitable entre les noms d'organisations bruts et les noms structurés et normalisés attendus en sortie du système.

Les traitements appliqués ici ont pour objectif de définir un classement hiérarchique de ces entités multiples, le cas échéant. Ils s'effectuent en deux étapes : la première phase consiste à découper à nouveau ces segments, mais cette fois sur les virgules. Nous nommons le résultat de ce découpage des *sections*. La seconde permet d'effectuer un classement hiérarchique de ces sections.

Nous avons expliqué que dans la grande majorité des cas, chaque section entre virgules renvoie à une « sous-entité » d'une même organisation. Dans les noms d'organisations :

27. Division of Gynecologic Oncology, Department of Obstetrics and Gynecology, University of Kentucky Medical Center, 800 Rose Street, Whitney-Hendrickson Building, Lexington 40536, USA.

36. Department of Radiology, Hospital of the University of Pennsylvania, Philadelphia19104.

chaque section entre virgules renvoie à un objet du monde intégré à celui dénoté par la séquence suivante : en (36) par exemple, le département de radiologie fait partie de l'hôpital universitaire de Pennsylvanie, qui se trouve dans la ville de Philadelphie.

Pour découper ces noms, la règle suivante est appliquée :

$$[x, y] \Rightarrow \{x\} - \{y\} \tag{7.7}$$

Si une chaîne de caractères contient une virgule, alors les deux éléments de la chaîne de part et d'autre de la virgule sont des sections, symbolisées dans la règle par les *{}*, potentiellement liées hiérarchiquement ; ce lien potentiel est symbolisé par -.

De cette manière, nous pouvons nous assurer que chaque section est relativement cohérente et autonome, bien qu'elle soit liée hiérarchiquement aux autres. Cette segmentation n'est pas toujours parfaitement adéquate. Toutefois, elle permet un premier découpage grossier, qui s'avère souvent suffisant.

Pour nos exemples (27) et (36), les sections retenues comme candidats seront donc les suivants :

27. Division of Gynecologic Oncology, Department of Obstetrics and Gynecology, University of Kentucky Medical Center

(a) Division of Gynecologic Oncology

(b) Department of Obstetrics and Gynecology

(c) University of Kentucky Medical Center

36. Department of Radiology, Hospital of the University of Pennsylvania, Philadelphia19104.

(a) Department of Radiology

(b) Hospital of the University of Pennsylvania

Chaque section ainsi isolée est considérée comme contenant potentiellement une entité d'organisation. Si aucune virgule n'est présente dans le segment, c'est la totalité du segment qui est considérée comme « section » et envoyée directement en entrée des règles suivantes. A ce titre, cette section est conçue comme un nom d'organisation indépendant. Ce sera le cas pour le nom d'organisation :

37. Mitsubishi KK

Il obéit à la règle :

$$[x\neg, y] \Rightarrow \{xy\} \tag{7.8}$$

Si un nom d'organisation ne contient pas de virgule, alors ce nom est traité comme une section.

Dans les cas où plusieurs sections ont été identifiées, la deuxième phase consiste à sélectionner celles qui contiennent une amorce de la liste. Chaque section ainsi sélectionnée est considérée comme une entité d'organisation, même s'il s'agit d'une organisation liée à d'autres.

La sélection des sections est fondée sur la règle suivante :

$$\{(\neg ADD)\} \Rightarrow \{(\neg ADD)\} = //ORGANISATION// \tag{7.9}$$

Si une section contient une amorce qui n'est pas une amorce d'adresse, c'est-à-dire une amorce d'entreprise ou d'organisme, alors cette section est une entité d'organisation.

A partir du résultat de cette règle, si une et une seule des sections contient une amorce, aucun classement n'est effectué, et la section en question est considérée comme l'entité d'organisation.

Dans le cas contraire, l'ensemble des sections sont comparés entre elles, et un classement est établi sur la base du niveau hiérarchique des amorces qu'elles contiennent. Une fois encore, ces règles s'appuient sur la liste structurée des amorces. Par exemple, une section candidate contenant *Univ*, amorce de niveau 1 selon la liste, sera considérée comme contenant une entité nommée de niveau supérieur à celle contenue dans une section où *Dept*, amorce de niveau 5, est présente.

Notons que, dans des cas où plusieurs amorces de même niveau sont présentes dans deux sections différentes, comme *Ctr* et *Inst* par exemple, la première section rencontrée est considérée arbitrairement comme la plus haute hiérarchiquement, englobant donc la seconde.

Revenant à nos exemples, pour (27) et (36), les sections sont classées hiérarchiquement comme indiqué dans le tableau 7.4.

Nom d'organisation brut	Rang relatif	Sections hiérarchisées correspondantes
Department of Radiology, Hospital of the University of Pennsylvania, Philadelphia19104	1	Hospital of the University of Pennsylvania
	2	Department of Radiology
Division of Gynecologic Oncology, Department of Obstetrics and Gynecology, University of Kentucky Medical Center	1	University of Kentucky Medical Center
	2	Department of Obstetrics and Gynecology
	3	Division of Gynecologic Oncology

TABLE 7.4 – Classement hiérarchique des sections issues de deux segments à entités multiples

Le rang hiérarchique qui est attribué à chaque section est relatif : il représente le classement des sections les unes par rapport aux autres, et non le classement des amorces qu'elles contiennent dans la liste hiérarchisée des amorces disponible en annexe.

Formellement, cette hiérarchie est obtenue grâce aux règles :

$$\{(NIVSUP)\} \Rightarrow \{NIVSUP\} \tag{7.10}$$

Si l'amorce contenue dans une section est identifiée comme étant de niveau supérieur aux amorces contenues dans les autres sections, alors cette section est hiérarchiquement supérieure aux autres.

$$\{(NIVINF)\} \Rightarrow \{NIVINF\} \tag{7.11}$$

Si l'amorce contenue dans une section est identifiée comme étant de niveau inférieur à certaines amorces contenues dans d'autres sections, alors cette section est hiérarchiquement inférieure à ces autres sections.

Il est à noter que si aucune amorce n'est trouvée, un autre ensemble de règles tente de détecter une suite de majuscules. En effet, ces suites de majuscules au sein de noms saisis majoritairement en minuscules sont dans la plupart des cas des acronymes d'organisations. C'est le cas particulier que nous avons mentionné en 7.1.2.2 page 257, concernant l'utilisation des majuscules comme indices. En cas de réussite, c'est la section contenant cette suite de majuscules qui est conservée. En cas de nouvel échec, le système envoie l'ensemble du segment en entrée du groupe de règles suivant.

A l'issue des traitements effectués par les règles de sélection et de structuration, les entités nommées d'organisations ont donc été identifiées, en partie délimitées, et hiérarchisées.

Type de règles 4 Dernier découpage et réécriture des noms d'organisations

Il arrive que le premier découpage sur les virgules ne soit pas suffisamment fin, et une section placée en entrée du deuxième ensemble de règles peut contenir plusieurs amorces de niveau hiérarchique différent. En (27) et (36), nous trouvons en effet deux amorces différentes dans la même section :

27. Division of Gynecologic Oncology, Department of Obstetrics and Gynecology, University of Kentucky Medical Center, 800 Rose Street, Whitney-Hendrickson Building, Lexington 40536, USA

- ***University** of Kentucky Medical **Center***

36. Department of Radiology, Hospital of the University of Pennsylvania, Philadelphia19104

- ***Hospital** of the **University** of Pennsylvania*

Dans un tel cas, une règle est appliquée pour tenter un second découpage, à l'intérieur même de la section. Elle s'appuie sur des patrons lexico-syntaxiques permettant de repérer des structures qui autorisent une segmentation claire en syntagmes nominaux, et donc une identification sans équivoque des différents noms d'organisations présents. Nous nous appuyons notamment sur la structure :

$$/préposition(+article) + amorce\ de\ niveau\ supérieur/ \tag{7.12}$$

Cette structure est utilisée dans la règle :

$$\{/ + xy + PrepArt(NIVSUP) + za + /\} \tag{7.13}$$

$$\Rightarrow |PrepArt(NIVSUP) + za + | = //ORGANISATION// \tag{7.14}$$

$$\Rightarrow | + xy + | = //ORGANISATION// \tag{7.15}$$

Si une section à deux amorces est composée d'une chaîne de caractères suivie du patron *Préposition + article + amorce de niveau supérieur* suivie d'une nouvelle chaîne de caractères, alors la sous-section composée de l'amorce de niveau supérieur et de la chaîne de caractères suivante est une organisation ; par conséquent, la chaîne de caractères apparaissant avant le patron est une autre organisation.

En isolant le syntagme contenant l'amorce de plus haut niveau, cette règle permet de déterminer de façon certaine que tout ce qui intervient avant l'amorce n'appartient pas à l'entité hiérarchiquement supérieure, et peut donc être considérée comme l'autre entité nommée présente. Ainsi en (36), nous sommes en mesure d'affirmer que la sous-section *Hospital of the* contient l'une des deux organisations, et que la seconde est : *University of Pennsylvania*

De même, pour le nom brut suivant :

38. Department of Botany, University of Coimbra Center for Functional Ecology, Coimbra, Portugal

University of Coimbra et *Center for Functional Ecology* sont deux unités hiérarchiques différentes où l'une, le centre, « appartient » à l'autre, l'université, et ne sont pas explicitement et « typographiquement » séparées par une virgule. Cependant, dans le cas qui nous occupe, la structure syntaxique couplée à la présence des amorces d'organismes permet de repérer la rupture entre les deux éléments. En effet, puisque *University* et *Center* sont considérés comme deux amorces d'organisme, et puisque *University* est placé linéairement avant *Center*, nous savons qu'à partir de *Center*, la suite de mots relève du centre et non de l'université, et ce jusqu'à la virgule suivante. Les seuls mots restant encore à rattacher à l'une ou à l'autre des amorces sont alors *of* et *Coimbra*. Or, une séquence de mots débutant par *of* est un syntagme prépositionnel, qui peut jouer le rôle de modifieur de nom. Nous pouvons dès lors utiliser le patron suivant :

$$amorce\ 1 + préposition(+article) + chaîne\ de\ caractères\ sans\ espaces + amorce\ 2 \quad (7.16)$$

Ce patron est utilisé dans la règle :

$$\begin{aligned} &\{+ab+/(AMORCE)PrepArt+x+(AMORCE)/+yz+\} \\ \Rightarrow\ &|+ab+(AMORCE)PrepArt+x+|=//ORGANISATION// \\ &\Rightarrow |(AMORCE)+yz+|=//ORGANISATION// \end{aligned} \quad (7.17)$$

Si une section à deux amorces est composée d'une chaîne de caractères suivie du patron *Amorce + Préposition + article + chaîne de caractères sans espaces (soit un mot graphique) + Amorce*, lui-même suivi d'une autre chaîne de caractères, alors la sous-section composée de tout ce qui est placé avant la seconde amorce est une organisation ; par conséquent, tout le reste, soit la seconde amorce et la chaîne de caractères suivante est une deuxième organisation.

Puisque l'identifiant d'organisme est forcément un nom, et que le rattachement prépositionnel se fait de la droite vers la gauche, nous considérons que *of Coimbra* est très probablement le modifieur de *University*, et donc que ce syntagme fait partie du nom de l'université et non du

centre qui suit. Cela est possible uniquement si la chaîne de caractères placée entre la préposition et la seconde amorce est un mot graphique, donc si elle ne contient pas d'espaces. En effet, à partir de deux mots graphiques, le second peut être une partie du modifieur de la première amorce, ou bien un modifieur de la seconde. Puisque nous n'utilisons pas d'analyseur syntaxique, cela n'est pas décidable dans notre système de règles.

Cependant, de telles structures ne sont pas systématiques, et un découpage univoque entre des groupes ou syntagmes n'est pas toujours possible à l'aide de règles. Ainsi, certains cas restent problématiques même après cette phase du traitement, comme en (27), pour lequel il est impossible de déterminer à l'aide de patrons et sans ressource externe exhaustive où effectuer la segmentation, de manière à isoler *University of Kentucky* et *Medical Center*, pour la raison évoquée ci-dessus. C'est ici que les traitements endogènes détaillés dans la section 6.1 page 159 rentrent en jeu, et permettent de déterminer la frontière entre les deux entités.

Les traitements effectués ensuite sur la section diffèrent en fonction du type de l'amorce identifiée. S'il s'agit d'une amorce de type entreprise, elle sera supprimée, de manière à pouvoir identifier comme une seule et même entreprise toutes les filiales d'une même maison mère, sous réserve que les filiales aient toutes le même nom. Le nom d'organisation (37) subira donc la transformation suivante :

37. Mitsubishi KK :
    - ***Mitsubishi***

Cette suppression du sigle obéit à la règle :

$$\{(ENTREPRISE)\} \Rightarrow\gg (ENTREPRISE) \tag{7.18}$$

Si l'amorce de la section est une amorce d'entreprise, alors cette amorce est supprimée.

En revanche, si l'amorce relève du type organisme, elle ne sera pas supprimée, puisqu'elle permet de distinguer plusieurs sous-catégories d'institutions, comme un hôpital avec *Hosp*, un institut avec *Inst*, ou une université avec *Univ*.

Elle sera parfois déplacée en tête de section, notamment pour les universités, de manière à ce que tous les noms d'organisations relevant de la même sous-catégorie *Université* soient identifiables immédiatement.

Les règles permettant de traiter les noms d'université sont :

$$\{(UNIV)\} \Rightarrow \{^{\frown}(UNIV) \tag{7.19}$$

Si l'amorce de la section est l'amorce *UNIV*, alors elle est placée en tête de la section.

Les autres amorces sont maintenues à leur position d'origine dans le nom. En effet, dans le cas des instituts ou des laboratoires par exemple, les noms obtenus après déplacement des amorces

sont parfois trop peu compréhensibles, ce qui gêne le travail des analystes.

Ces règles de découpage et de réécriture permettent d'obtenir, à partir des sections identifiées, les noms d'organisations suivants :

27. Univ of Kentucky Medical Ctr
- *Univ of Kentucky Medical Ctr*

36. Hosp of the Univ of Pennsylvania
- *Hosp of the*
- *Univ of Pennsylvania*

39. Harvard Univ
- *Univ Harvard*

La dernière étape consiste à nettoyer une dernière fois les noms ainsi transformés en éliminant les mots grammaticaux listés dans nos lexiques. Cette transformation peut ne pas paraître opportune dans certains cas ; cependant, elle présente l'avantage de lisser un peu plus les noms d'organisations, et d'éliminer un certain nombre de variantes sur cette base. Par exemple, elle permet de rassembler des noms comme *Univ de Bretagne Sud* et *Univ Bretagne Sud*.

Pour nos trois exemples, les sections récupérées à l'issue du système de règles, et donc les entités nommées identifiées et normalisées à ce stade, sont présentées dans le tableau 7.5, (dérivé du tableau 7.4 page 272) en regard de leur segment d'origine.

Nom d'organisation brut	Rang relatif	Sections hiérarchisées correspondantes
Department of Radiology, Hospital of the University of Pennsylvania, Philadelphia19104	1	Univ Pennsylvania
	2	Hospital
	3	Dept Radiology
Division of Gynecologic Oncology, Department of Obstetrics and Gynecology, University of Kentucky Medical Center	1	Univ Kentucky Medical Center
	2	Dept Obstetrics Gynecology
	3	Div Gynecologic Oncology
Mitsubishi KK	1	Mitsubishi

TABLE 7.5 – Classement hiérarchique des noms normalisés à partir de deux segments à entités multiples

Ces entités, ainsi que les segments dont elles sont issues, sont restituées dans la base de données internes de TKM. Un nom brut contenant plusieurs organisations distinctes est dupliqué en autant d'exemplaires qu'il existe d'organisations distinctes dans le nom de départ.

Nous présentons dans le tableau 7.6 un récapitulatif des types de règles employés pour la normalisation.

Type	Sous-catégories	Exemple
1. Nettoyage et standardisation	Elimination des parasites	*jowell@gx.net*
	Standardisation des amorces	*University / Université → UNIV*
2. Découpage des noms bruts	Séparation des organisations distinctes	*CNRS || Total SA*
	Séparation des organisations et de leur adresse	*COLLAGEN MATRIX, INC. || 509 Commerce Street*
	Rapprochement des organisations hiérarchiquement liées	*The Research Fdn of State Univ of New York > Office of Science and Technology Transfer*
3. Découpage et classement des segments	Découpage interne des organisations à entités multiples	*Dept of Radiology || Hospital of the Univ of Pennsylvania*
	Classement hiérarchique en fonction des amorces	*Hospital of the Univ of Pennsylvania > Dept of Radiology*
4. Dernier découpage et réécriture	Derniers découpage et classement internes aux sous-séquences	*Univ of Pennsylvania > Hospital*
	Réécriture	*Mitsubishi KK → Mitsubishi* *Harvard Univ → Univ Harvard* *Univ of Pennsylvania → Univ Pennsylvania*

TABLE 7.6 – Récapitulatif des types de règles employés dans la normalisation

7.1.2.3 Validation : évaluation du système de règles

Protocole d'évaluation Afin de mesurer la qualité de nos résultats de manière fiable, nous avons établi un protocole d'évaluation. L'évaluation a consisté à comparer les noms obtenus avec les attendus que nous avons présentés (cf. *supra*), à l'aide des mesures qui sont traditionnellement utilisées pour les systèmes de traitement automatique des langues comme par exemple dans les conférences Text REtrieval Conference (TREC) ou Message Understanding Conference (MUC). Notre évaluation porte uniquement sur les entités d'organisations et celles de lieux, puisque ce sont ces dernières qui posent véritablement problème. Les auteurs ne sont pas retouchés après l'import, puisque leur forme est satisfaisante pour les traitements à effectuer par la suite, et la normalisation des dates est triviale. Puisque les entités de lieux se trouvent dans les même unités que les entités d'organisations, ce sont des noms d'organisations bruts qui ont été sélectionnés pour le calcul d'évaluation. Nous avons choisi de calculer le bruit et le silence obtenus pour un échantillon de 1000 noms d'organisations bruts, ainsi qu'un certain nombre de mesures que nous pouvons qualifier d'intermédiaires, et que nous détaillons ci-dessous.

Corpus d'évaluation Nous avons équilibré notre corpus d'évaluation de la manière suivante : La moitié des 1000 noms d'organisations sont des noms de déposants de brevets, tandis que les

autres sont des auteurs d'articles. Les publications sont issues de divers domaines que nous avons présentés en 1.2.2 page 21.

Détail des mesures Les éléments évalués sont la conformité des entités d'organisations normalisées aux formes standards, l'extraction des adresses, et le type - entreprise, organisme ou particulier - attribué aux entités d'organisations. Nous présentons les différents critères qui nous permettent de réaliser une évaluation précise de nos résultats sur les noms, et de savoir où sont les difficultés majeures. Nous ne nous sommes pas limitée à la quantification des bruits et des silences, mais avons opté pour une typologie plus fine. Nous présentons cette typologie dans le tableau 7.7.

Intitulé de l'évaluation	Critères de sélection
Correct	Forme attendue, complète et sans parasites
Correct mais mots ou entités en désordre	Les mots extraits sont les bons, mais pas dans le bon ordre ; les entités multiples sont bien délimitées, mais ne sont pas dans le bon ordre
Partiellement correct avec bruit	Forme attendue + éléments non pertinents
Partiellement correct avec silence	Forme attendue – un ou plusieurs éléments pertinents
Partiellement correct avec bruit et silence	Forme attendue + éléments non pertinents – un ou plusieurs éléments pertinents
Incorrect avec bruit	Le nom ramené n'est pas celui attendu et ne contient que du bruit ; ex. : *203 Faculty Street* au lieu de *ELS Language Centers*
Incorrect avec silence	Le système n'a effectué aucune normalisation, et le nom brut est incorrect

Table 7.7 – Critères d'évaluation de la normalisation des entités d'organisations

En ce qui concerne le type d'organisation et l'adresse, les critères d'évaluation sont plus restreints. L'évaluation de l'extraction du pays a été réalisée en priorité, puisque c'est l'information cruciale pour les analystes, qui peuvent alors établir des classements des pays les plus performants sur un domaine donné. Pour le pays, trois possibilités existent :

- le pays n'a pas été ramené, et c'est un problème de rappel ;
- le pays ramené n'est pas le bon, et c'est un problème de précision ;
- le pays a été correctement extrait.

Ce sont les mêmes critères qui sont appliqués aux types.

Pour systématiser au maximum la tâche d'évaluation, nous avons créé une interface permettant de procéder à la vérification de chaque nom de l'échantillon, et de traiter les données (voir figure 7.3 page précédente). De cette manière, les calculs sont automatisés et moins soumis au

risque d'erreur.

L'évaluateur vérifie chaque nom normalisé à partir de cette interface, en se fondant à la fois sur le nom original et le nom normalisé en colonnes 2 et 3, et sur les modèles standards élaborés (cf. supra). Il doit cocher les cases correspondant au cas de figure qui lui est présenté pour déterminer s'il est ou non correct, et ce pour l'ensemble des noms d'organisations sur lesquels le système de normalisation a été appliqué.

Qualité de la normalisation obtenue Nous présentons les taux de normalisation correcte pour les 1000 noms d'organisations de l'échantillon, en fonction du type de documents. Nous cumulons ensuite les résultats de la normalisation du nom avec l'extraction du type, puis avec l'extraction du type et du pays (tableau 7.8).

organisations correctement normalisées	Nom	Type	Pays	Nom + Type	Nom + Type + Pays
Brevets	94,4% (472/500)	73,4% (367/500)	45,2% (226/500)	70,4% (352/500)	43% (215/500)
Articles scientifiques	73,6% (368/500)	94,2% (471/500)	97,2% (486/500)	71,6% (358/500)	70,8% (354/500)
Total	84% (840/1000)	83,8% (838/1000)	71,9% (719/1000)	71% (710/1000)	56,9% (569/1000)

TABLE 7.8 – Organisations correctement normalisées et typées, pays correctement identifiés

Le tableau 7.9 page ci-contre montre la répartition des erreurs par type de document, ce qui apporte de l'information supplémentaire sur les raisons des erreurs de normalisation.

Répartition des erreurs	Correct mais mots en désordre	Partiellement correct avec bruit	Partiellement correct avec silence	Incorrect avec bruit	Incorrect avec silence	Total d'erreurs par type de document
Brevets	0%	3,8%	1,8%	0	0	5,6%
Articles	0,3%	19,5%	1,4%	3,5%	1,7%	26,4%

TABLE 7.9 – Répartition des erreurs du système sur les noms, par type de document

Sur les 1000 noms d'organisations, 136 avaient déjà la forme attendue avant que le processus de normalisation ne soit appliqué. Cela confirme le fait que la normalisation est une étape indispensable, puisque 86,4% des noms bruts ont une forme différente de celle que nous attendons.

Globalement, nous obtenons des résultats relativement satisfaisants sur la normalisation des noms, puisque sur les 1 000 noms d'organisations, 840 ont la forme standard attendue (tableau 7.8). Si nous retirons les 136 formes déjà correctes, qui ne sont pas modifiées par le système,

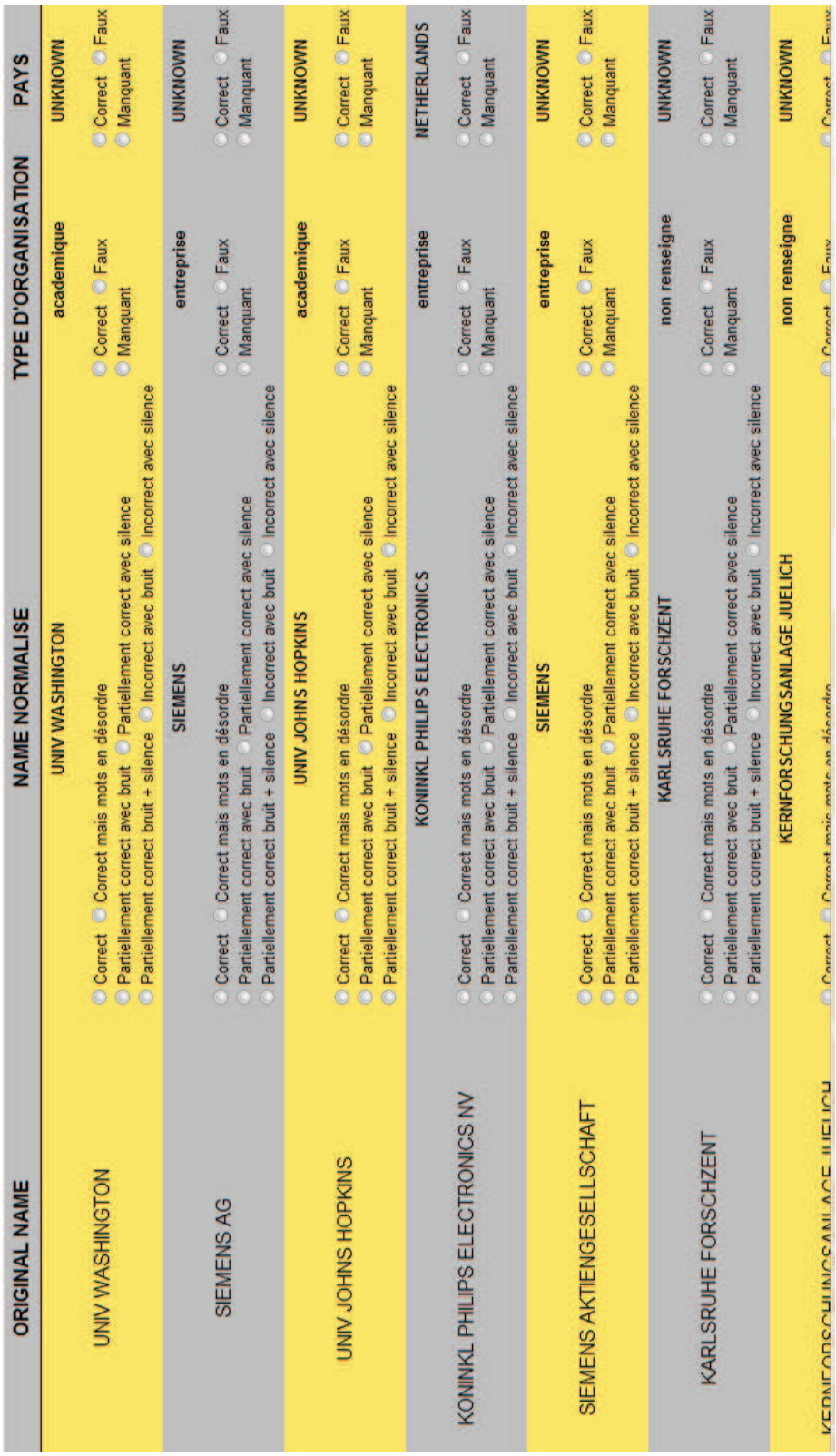

FIGURE 7.3 – Interface d'évaluation de normalisation

704 noms ont été correctement normalisés d'après nos standards.

Il existe une différence importante entre les résultats pour les brevets et ceux pour les articles scientifiques en ce qui concerne la normalisation des noms d'organisations (donc hors extraction du type et du pays) : alors que le nom des déposants de brevets est correctement normalisé dans 94,4% des cas, seuls 73,6% des noms d'auteurs d'articles correspondent aux standards établis. Le tableau de répartition des erreurs par type de document nous permet d'interpréter cet écart de manière plus précise : 19,8% du nombre total de noms d'auteurs d'articles sont mal normalisés en raison d'un manque partiel de précision, à savoir qu'il contient à la fois toute l'information pertinente, et un certain nombre d'éléments non pertinents. C'est le cas pour seulement 3,8% des noms de déposants de brevets. La plupart des déposants de brevets sont des entreprises, tandis que la majeure partie des auteurs d'articles sont des institutions publiques. Or, nous l'avons signalé, les noms d'institutions publiques sont généralement beaucoup plus longs que ceux des entreprises, et contiennent beaucoup d'informations qui doivent être éliminées. De fait, ils sont beaucoup plus difficiles à normaliser.

Le faible taux d'erreurs dues au silence montre que notre traitement permet, sinon une parfaite exhaustivité, du moins une couverture assez large.

Lorsque les types sont considérés comme non pertinents, dans tous les cas, il s'agit d'un problème de silence, et jamais d'un manque de précision : soit l'information est présente et correctement identifiée, soit elle n'apparaît pas dans le nom d'organisation brut. Notons qu'il aurait été possible d'équilibrer le corpus par type d'organisations. Cependant, nous ne l'avons pas fait car nous ne l'avons pas jugé souhaitable, parce qu'une répartition égale entre entreprises, organismes et particuliers n'aurait pas été représentative des données. Nous voulions restituer la réalité d'une étude, qui se partage en majorité entre entreprises et organisations. De plus, puisque ces noms sont minoritaires, ils représentent une part faible des déposants de brevets, et même nulle pour les publiants d'articles scientifiques. Par conséquent, leur poids dans les analyses des ingénieurs d'étude de TKM est très faible.

Quant à la répartition générale des types d'organisations sur un corpus donné, il est difficile d'en tirer une moyenne. En effet, la proportion des entreprises et des organismes est directement corrélée à la proportion de brevets et articles scientifiques. En général, les déposants de brevets sont souvent des entreprises, tandis que les organisations publiantes sont très fréquemment des organismes. Nous présentons à titre d'exemple, dans la figure 7.10 page suivante, la répartition de ces types d'organisations pour une étude réalisée au sein de TKM sur la télé-médecine, que nous mettons en regard des genres des documents rassemblés.

Là encore, le nombre de noms d'organisations est supérieur à celui des documents, puisque plusieurs organisations peuvent être propriétaires d'un même document. Le type des 2 624 organisations non déterminées n'a pas été identifié automatiquement : il s'agit majoritairement d'entreprises ne contenant pas d'amorce, et plus rarement de noms de particuliers. Les proportions

Etude TKM				
Nombre de documents		**Nombre de noms d'organisations**		
8692		13891		
Brevets	Articles sc.	Entreprises	Organismes	Non déterminés
3105	5587	3241	8026	2624
%	%	%	%	%
35,7	64,3	23,3	57,8	18,9

TABLE 7.10 – Répartition des genres de documents dans une étude TKM effectuée sur la télémédecine, et répartition des types d'organisations

entre entreprises et académiques n'en sont pas modifiées, même avec un examen des organisations au type non renseigné. Les types dépendent donc du genre des documents sélectionnés par les analystes, en fonction de leurs besoins.

Concernant l'identification des pays, nous observons la même tendance. Les résultats montrent une grande disparité entre les brevets et les articles scientifiques. La raison en est que les noms d'organisations associés aux brevets, souvent des noms d'entreprises, véhiculent beaucoup moins d'informations que les noms associés aux articles, la plupart du temps publiés par des organismes. Sur notre corpus de travail de 13 187 noms bruts d'organisations, nous avons constaté que 1,99% seulement des noms associés aux articles scientifiques ne contenaient aucun nom ni isocode de pays, contre 47,09% des noms associés aux brevets (voir la sous-section 7.1.1.2 page 247). Ce dernier chiffre explique la faible performance de notre système de normalisation pour les pays sur notre corpus d'évaluation.

Nous soulignons pour finir le fait que l'évaluation de cette normalisation a été réalisée à partir des attendus objectivés que nous avions définis à l'aide des critères fournis par les utilisateurs, et en collaboration avec ces derniers. Par conséquent, nous avons appliqué ces attendus pour l'ensemble des noms normalisés. Or, dans certains cas, les utilisateurs auraient probablement identifié des noms mal normalisés, là où nous avons considéré qu'ils rentraient dans la catégorie des noms « corrects ».

Typiquement, pour les noms considérés comme « corrects mais avec des mots ou des entités dans le désordre », nous obtenons un taux d'erreur très faible, avec un taux nul pour les brevets, et de seulement 0,3% pour les articles (voir tableau 7.9 page 280). En effet, la plupart des noms

normalisés à entités multiples ont été ordonnés automatiquement dans l'ordre prévu, grâce à la liste hiérarchisée des amorces. Cependant, ainsi que nous l'avons souligné plus haut, l'ordre des amorces décrivant la structure d'un ensemble d'organisations liées hiérarchiquement a été déterminé d'après des tendances générales, certes lourdes, mais pas systématiques. Par conséquent, l'ordre hiérarchique déterminé automatiquement pour ces entités peut ne pas correspondre à la réalité de la structure. Or, pour certaines études, les niveaux les plus bas hiérarchiquement peuvent avoir leur importance, et il est alors nécessaire de les réagencer correctement. Ainsi, sur des noms bien précis, et dans un contexte et pour une tâche donnés, il est possible que l'objectivation nécessaire au traitement par un programme soit trop rigide pour des utilisateurs finaux avec des besoins spécifiques. Cela ne remet pas en cause la qualité globale du système ; néanmoins, cela montre que les résultats d'un traitement automatique peuvent ne pas toujours correspondre à des besoins précis à un moment *t*.

7.1.3 Conclusion

La normalisation des entités nommées, et plus particulièrement celle des entités d'organisations et de lieux, est un processus complexe fondé entre autres sur un système de règles exogène.

Celui-ci passe par l'exploitation des régularités présentes de manière récurrente dans tous les ensembles documentaires constitués. Il permet de réduire la variation, qu'elle soit due à la divergence de conventions de notation, à la traduction appliquée ou non, ou inhérente aux noms à entités multiples par exemple.

Les régularités exploitées par les règles relèvent de la syntaxe et/ou d'unités lexicales, et les noms ainsi normalisés comportent des structures syntaxiques simplifiées, et une variation lexicale relativement maîtrisée. De plus, ils sont structurés de manière hiérarchique le cas échéant.

Notre système de règles est appliqué aux noms bruts lors de leur entrée dans l'ensemble documentaire. Les noms normalisés sont distingués de leur nom brut d'origine, mais le lien reste maintenu entre les uns et les autres, en particulier à travers la création de l'unité intermédiaire qu'est le segment.

Ce maintien de la relation entre entité normalisée et noms bruts permet leur capitalisation en vue d'une exploitation ultérieure, sur d'autres ensembles documentaires.

Par ailleurs, les entités ainsi normalisées sont utilisées à leur tour pour la constitution de la ressource termino-ontologique multi-plans, en combinaison avec des ressources exogènes.

7.2 Des ressources exogènes pour la constitution de la ressource termino-ontologique

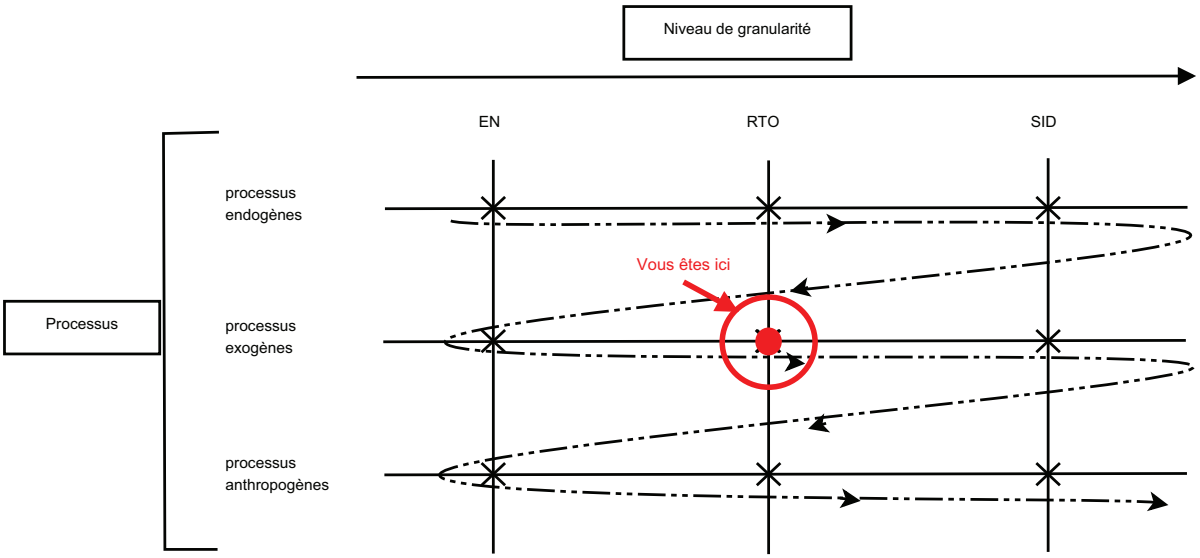

FIGURE 7.4 – Positionnement des traitements exogènes pour la ressource termino-ontologique multi-plans dans l'ensemble des processus

Dans le chapitre 6 (section 6.2 page 199), nous avons présenté les méthodes et ressources endogènes permettant de constituer la ressource termino-ontologique multi-plans. Les ressources en question sont issues du corpus, et sont intégrées dans les différentes facettes de la ressource, éventuellement après transformation par des traitements de normalisation. Ces méthodes ne sont pas les seules qui ont été utilisées. De plus, certaines ont été enrichies de ressources exogènes. Ainsi, nous employons pour la constitution de la RTO des méthodes pouvant être qualifiées d'hybrides puisque faisant appel aux données du corpus pour en tirer de la connaissance, mais également à des ressources exogènes pour compléter ou améliorer les traitements endogènes (voir figure 7.4). Dans les sous-sections suivantes, nous exposons d'abord l'hypothèse qui a motivé l'exploitation de ressources exogènes pour plusieurs facettes de la RTO ; puis nous présentons les applications réalisées, qu'elles soient à l'état de tests ou de réalisations effectives et intégrées aux outils de la société TKM. Enfin, nous montrons en quoi cet aspect de nos travaux a été évalué.

7.2.1 Hypothèse

Dans le chapitre 2 (section 2.3 page 48), nous avons exposé les apports épistémologiques dont nos travaux pouvaient bénéficier. En particulier, nous nous sommes positionnée du côté des épistémologies constructivistes, qui postulent que la connaissance n'est pas une simple copie du réel, mais une construction cognitive entre un sujet humain et un objet observé. Dans ce cadre,

nous considérons qu'il n'existe pas d'objectivité, de manière absolue, mais une intersubjectivité stable permettant d'acquérir et de partager des connaissances. Par voie de conséquence, la réalité ontologique postulée par les épistémologies positivistes est exploitée par le courant constructiviste comme une « fiction utile ».

Ainsi, si la réalité ontologique n'existe pas en tant que telle pour le constructivisme, elle n'en représente pas moins une fondation sur laquelle il est possible d'appuyer des structures de représentation des connaissances. Par conséquent, il est possible de tendre vers un idéal d'objectivité, même si celui-ci reste « fictionnel ».

Or, les méthodes endogènes sont avant tout fondées, comme nous l'avons décrit dans le chapitre 6, sur des données issues du corpus à traiter. Leur avantage majeur est, nous l'avons vu, l'adéquation des ressources ainsi extraites au corpus à traiter. Cependant, elles posent aussi une série d'inconvénients : d'abord, dans le cadre de la ressource termino-ontologique que nous mettons en place, les biais éventuellement présents dans le corpus lui-même se répercutent de fait sur la ressource construite à partir de ce dernier ; d'autre part, l'exhaustivité n'est pas envisageable par ces méthodes, puisqu'un corpus ne saurait être représentatif de la totalité d'un phénomène, quel qu'il soit.

Si, dans certains cas, ces biais pèsent peu sur l'efficacité d'un processus, et bien que l'exhaustivité ne soit pas toujours cruciale, atteignable, ni même souhaitable, d'autres facettes, et en particulier celle des lieux, peuvent s'avérer plus cohérentes et utiles si certains aspects de leur contenu couvrent l'ensemble des entités possibles. Par ailleurs, utiliser des lexiques fermés, lorsqu'ils existent et dans le cas où les connaissances qu'ils représentent sont reconnues comme stables et viables, peut être un moyen d'éviter des calculs lourds et complexes, ou de corriger ou compléter des traitements endogènes. C'est le cas de la facette des thèmes fondée sur les collocations. Enfin, faire appel à une base de règles, elle aussi fondée sur des lexiques fermés, et lorsque les régularités se dégageant du corpus sont formalisables et exploitables, permet de se rapprocher de la réalité en se détachant des variantes textuelles ; c'est ce qui a été utilisé pour la facette des organisations.

Pour la constitution des facettes des lieux, des thèmes et des organisations, nous formulons donc l'hypothèse suivante :

Hypothèse II.12. *Pour compléter des traitements fondés sur des ressources endogènes pour la constitution d'une structure de représentation des connaissances, il est pertinent d'utiliser des ressources exogènes, et en particulier des lexiques et des bases de règles, de manière à rendre ces structures et leur contenu plus complets et/ou plus pertinents.*

7.2.2 Application et méthodologie : des lexiques et systèmes de règles à plusieurs étapes de constitution

A partir de cette hypothèse, nous avons mis en place des traitements fondés sur des ressources exogènes de façon à optimiser la constitution de certaines facettes de la ressource termino-ontologique (RTO) multi-plans. La facette des lieux, et celle des thèmes, ont été enrichies à l'aide de lexiques ; celle des organisations à partir d'une base de règles, *via* la normalisation des entités nommées.

7.2.2.1 L'enrichissement exogène par lexiques : les lieux et les thèmes

Des lexiques pour la facette des lieux de la RTO

Dans le chapitre 6 (section 6.2 page 199), nous avons présenté les différentes facettes de la ressource termino-ontologique multi-plans que nous avons modélisée. Dans ce modèle, la facette des lieux contient trois types d'entités de lieux : les pays, les villes, et enfin les adresses.

Pour constituer une partie de cette facette, des ressources exogènes statiques, sous forme de lexiques, peuvent être exploitées. Pour cela, nous avons utilisé des lexiques pré-constitués de pays et de villes. Le lexique des pays est issu du site web de l'Organisation Internationale de Normalisation [62]. Il a été enrichi par quelques zones régionales propres aux zones de couverture des brevets, telles que *World* ou *Europe*, qui ont également un isocode dédié. La liste ainsi constituée contient 251 pays ou zones. Chaque nom de pays est associé à son isocode.

Le lexique des villes a été constitué à partir d'une liste de 2 700 000 villes [63], chacune étant associée à son pays correspondant et à ses coordonnées géographiques en latitude et longitude.

Ces lexiques peuvent être directement intégrés dans la facette des lieux. Dans le cas des villes, l'entité mère de chacune d'elles, c'est-à-dire le pays, est également présent. Cela permet de restituer la hiérarchie entre pays et ville, et de retrouver le premier à partir de la deuxième. Les noms de villes homonymes, telles que *Vienne* en Autriche et *Vienne* en France, sont distinguées. Cela respecte l'engagement sémantique rendant possible la désambiguïsation des villes (voir la sous-section 3.2.1 page 79). Enfin, pour chaque pays, l'isocode est également présent.

Tous les items de ces lexiques sont en anglais, puisque la plupart du temps, les pays et villes restitués dans les données brutes le sont dans cette langue.

Ainsi, la majeure partie de la facette des lieux (la seule exception étant les adresses) est constituée *a priori*, indépendamment du corpus. Cela ne signifie pas pour autant que son contenu ne soit pas relié au reste des informations, ni aux documents. En effet, des liens sont créés entre les données intégrées *a priori* et les informations, pays ou villes, détectées dans les documents à partir de ces données.

62. Liste disponible sur : `http://www.iso.org/iso/fr/english_country_names_and_code_elements`
63. Liste des villes disponible sur le site web de Maxmind : `http://www.maxmind.com/`. Accès à la liste des villes : `http://www.maxmind.com/app/worldcities`

Cette détection a lieu lors du processus de normalisation, décrit dans la section 7.1 page 241, à l'aide du système à base de règles. La facette des lieux sert donc également de lexique placé en entrée du système pour la détection des pays et villes à partir des noms d'organisations bruts, pour enrichir cette même facette.

Des lexiques pour l'extraction des collocations Nous avons exposé dans la sous-section 6.2.3 page 209 la manière dont nous exploitons, pour notre modèle de ressource termino-ontologique et la structuration de l'ensemble documentaire, les résumés et titres de documents en tant que ressource endogène. Nous en extrayons les collocations brutes, afin de dégager les thèmes des documents d'un ensemble donné. Celles-ci sont des segments répétés au moins deux fois dans une unité textuelle déterminée, composés de un à cinq mots graphiques, non lemmatisés, et commençant et finissant par un mot plein. Nous postulons que ces collocations expriment le thème des documents.

Ce postulat a été le point de départ d'un stage réalisé à TKM par une étudiante de Master 1 en traitement automatique des langues de l'université Stendhal-Grenoble 3. [Farabet, 2010] a travaillé à l'implantation d'une telle méthode au sein des outils TKM, sous notre supervision.

Nous nous fondons sur son rapport *ibid.* pour présenter l'implantation et les résultats obtenus.

La méthode des segments répétés utilisée par [Lafon & Salem, 1983] n'utilise aucune ressource externe pour fonctionner : elle est entièrement fondée sur des statistiques, par le calcul des co-occurrences répétées dans le corpus à traiter. Il s'agit donc d'une méthode fondamentalement endogène. De plus, dans leur inventaire des segments répétés sont inclus des segments pouvant débuter et/ou finir par des mots grammaticaux, ce qui correspond à leur objectif d'analyse des spécificités de certains types de discours.

Cependant, dans notre propre cadre de travail, les mots grammaticaux, ou mots vides au sens de [Tesnière, 1959], lorsqu'ils occupent certaines positions dans les segments, peuvent parasiter les résultats. Par exemple, nous avons établi que les séquences commençant et/ou finissant par des mots outils sont dans notre cas peu utiles à la détection d'un thème ou sous-thème d'un ensemble documentaire. Par conséquent, nous avons cherché à éliminer les séquences dont les extrémités gauche ou droite sont des mots grammaticaux.

L'algorithme que nous avons mis en place a été appliqué à un corpus de 2 000 résumés et titres de publications, mêlant 1 000 articles et 1 000 brevets. Ces publications sont issues d'une étude réalisée dans le domaine de l'optique et portant plus précisément sur les caméras haute sensibilité et haute rapidité.

Pour ne prendre en compte que les segments que nous considérons comme potentiellement pertinents, nous fondons notre algorithme sur l'utilisation de lexiques. Ainsi, si l'utilisation du segment répété en tant qu'élément structurant de l'ensemble documentaire est une approche fondamentalement endogène, l'algorithme lui-même que nous avons mis en place, en revanche, s'appuie sur des ressources exogènes.

Nous avons pris le parti de mettre en place des lexiques de mots grammaticaux, et ce pour les cinq langues majoritaires dans les données : l'anglais, le français, l'italien, l'espagnol et l'allemand.

Puisque nous nous plaçons dans des données multilingues, la première étape du traitement est la détection de la langue : cette détection sert à savoir, par la suite, quel lexique de mots grammaticaux doit être utilisé. En effet, puisque nous cherchons à éliminer les séquences commençant ou finissant par des mots grammaticaux, il convient d'abord d'établir la langue dans laquelle ces mots doivent être recherchés.

Pour cela, nous nous fondons sur ces mêmes lexiques de mots grammaticaux. Nous utilisons ces unités comme déclencheurs, puisqu'à partir du moment où nous travaillons sur des données textuelles, nous considérons que nous sommes face à un discours. Or, aucun discours ne peut être rédigé sans l'aide de mots grammaticaux. Il s'agit donc d'éléments systématiquement présents, dès lors que le discours est constitué de phrases bien formées.

Le premier lexique utilisé pour ce repérage est le lexique anglais, puisque c'est la langue majoritaire dans les documents. Ses items des lexiques sont projetés sur chacun des résumés ; s'ils sont présents dans un résumé, l'algorithme considère qu'il est rédigé en anglais. Sinon, les autres lexiques sont projetés les uns après les autres, jusqu'à trouver des items correspondant à l'une des langues. Si aucune langue n'a été identifiée à l'issue de ce traitement, l'anglais est la langue attribuée par défaut, toujours en raison de la prédominance de cette langue dans le corpus.

Une fois la langue identifiée, la détection des segments répétés potentiels est effectuée, à l'aide d'une fenêtre glissante contenant un à cinq mots, ce qui correspond à la taille maximale de la fenêtre utilisée par [Lafon & Salem, 1983]. En effet, au-delà de cinq mots, l'algorithme devient plus complexe, ce qui représente un obstacle à sa mise en application. De plus, le nombre de segments porteurs de thème comportant plus de cinq mots est très réduit.

Pour chaque document, le titre, puis le résumé de la publication sont parcourus grâce à la fenêtre. A l'aide du lexique correspondant à la langue détectée, les séquences commençant et/ou finissant par des mots grammaticaux sont éliminées. De même, les séquences contenant des « délimiteurs de séquences » [Lafon & Salem, 1983] ne sont pas prises en compte. Les délimiteurs de séquences sont pour nous, tout comme pour ces auteurs, les « signes de ponctuation usuels », tels que le point, la virgule, le point-virgule, etc.

Ainsi, ne sont conservés que les segments ne contenant aucun délimiteur de séquence, et dont le premier et le dernier des mots de la suite sont des mots lexicaux et ne contenant aucun délimiteur de séquence. En revanche, des mots grammaticaux peuvent être présents à l'intérieur d'une séquence. Notre méthode récupère également les mots graphiques de manière isolée, afin d'obtenir, en plus des séquences, des mots simples pouvant représenter le thème d'un document.

De ces segments ainsi récupérés, ne sont conservés que ceux qui sont répétés au moins deux fois sur la totalité de l'ensemble documentaire. L'unité textuelle de la répétition, au sein d'un même

document ou dans des documents différents, n'est à ce stade pas prise en compte. Ces séquences sont alors stockés dans la facette de la ressource termino-ontologique dédiée aux thèmes des documents. Chaque segment extrait est relié au document dont il est issu.

Nous rappelons qu'à partir de notre corpus de travail, 314610 collocations brutes répétées au moins deux fois dans l'ensemble documentaire ont été extraites, avec une moyenne de 157 collocations brutes par document (voir la section 6.2.3 page 209 pour plus de détails).

Nous présentons dans le tableau 7.11 page 289 les séquences obtenues pour les mots *impact*, *of*, *tool*, *path* et *types* en tête de segment, extrait de la phrase :

Hence, this paper first discusses the impact of tool path types and other programming parameters on process implementation through an experimental campaign performed on a parallel kinematics machine tool.

Cette phrase est elle-même issue d'un résumé d'article scientifique.

Mot débutant la fenêtre	Suites détectées
impact	impact
	impact of tool
	impact of tool path
	impact of tool path types
of	Sans objet
tool	tool
	tool path
	tool path types
	tool path types and other
path	path
	path types
	path types and other
	path types and other programming
types	types
	types and other
	types and other programming
	types and other programming parameters

TABLE 7.11 – Séquences extraites pour les mots *impact*, *of*, *tool*, *path* et *types* en tête de segment, issus d'une phrase attestée en corpus

Les segments extraits sont donc de longueur 1 à 5 en nombre de mots, ou items, et la fenêtre

se déplace d'un item à chaque fois que sa taille maximale de cinq a été atteinte. Les seuls cas dérogeant à cette règle sont les délimiteurs de séquence, ainsi que les mots grammaticaux lorsqu'ils passent en tête de segment, et pour lesquels les suites ne sont pas calculées du tout, comme c'est ici le cas pour *of*.

De même, les séquences commençant par un autre mot mais finissant par un mot grammatical ne sont pas prises en compte.

Pour ces cinq mots, nous recensons 16 segments à extraire. L'élimination des séquences que nous venons de décrire permet donc de réduire le nombre de segments extraits, puisque sans ce tri, le nombre aurait pu atteindre 25 suites.

Etant donnée la phrase d'origine, il est d'ores et déjà possible de déterminer en partie les segments pouvant éventuellement être vecteurs de thème. Les séquences *impact of tool path types* ou *tool path types*, par exemple, peuvent éventuellement, en fonction du contexte, relever d'unités thématiques. En revanche, des suites comme *types and other* ne semblent que peu pertinentes, notamment en raison de structures syntaxiques tronquées et de mots peu porteurs de sens en eux-mêmes. Cependant, récupérer l'ensemble de ces suites nous permet de laisser à l'expert le soin de décider quelles suites sont pertinentes pour lui et pour une étude donnée.

Afin de limiter le poids des traitements, nous avons pris le parti de ne pas lemmatiser les formes extraites. De plus, cette lemmatisation pourrait entraîner une perte d'information, par exemple si le nombre d'un groupe nominal est significatif.

A ce stade du traitement, chacun des segments, à partir du moment où sa fréquence est d'au moins 2 dans l'ensemble documentaire, est considéré comme une collocation brute potentielle, et, par conséquent, comme un thème potentiel pour un document, un ensemble documentaire ou un sous-ensemble de ce dernier.

Dans le cadre de ce processus, le rappel est donc privilégié par rapport à la précision : l'objectif est en effet de rassembler le plus possible de suites potentiellement représentatives d'un thème, de manière à restituer un maximum d'informations. Cela se fait au prix d'un bruit relativement important, puisque toutes les séquences ainsi extraites ne peuvent relever de ce que les analystes pourraient considérer comme thématique. Néanmoins, nous limitons ce bruit grâce à l'élimination des suites ayant moins de chances d'être porteuses d'un thème, grâce à la non prise en compte des séquences commençant et/ou finissant par des mots grammaticaux.

La facette thématique ainsi constituée est destinée à être utilisée dans le système d'immersion final. Pour un ensemble documentaire précis et propre à une étude donnée, l'utilisateur pourra consulter les documents par une interrogation sur la base de mots simples ou de suites de mots représentant un thème ou sous-thème recherché.

C'est à la suite de cette recherche, sous forme de requête utilisateur, que sont calculées de manière plus fine et dédiée les collocations brutes. Cette méthode en deux temps bien distincts présente deux avantages : d'une part, c'est la seule manière de garantir le caractère récurrent d'une

collocation pour un ensemble documentaire spécifique, et ainsi de respecter le critère statistique de la collocation.

7.2.2.2 Enrichissement exogène par base de règles pour structurer le plan des organisations

Nous avons présenté, dans la section précédente, le système à base de règles permettant la normalisation des entités nommées. Cette normalisation est nécessaire à la qualité des données insérées dans les facettes, et en particulier dans celle des organisations. En effet, sans elle, la variation présente dans les noms d'organisations bruts empêcherait une exploitation efficace de ces derniers. Ainsi, ce traitement des unités d'informations par le biais d'une ressource exogène dynamique nous permet de respecter notre engagement sémantique [Bachimont, 2000], voir 3.2.1.

La maîtrise de cette variation passe entre autres par le caractère structurant de la normalisation, qui permet une segmentation hiérarchisante des noms d'organisations bruts, et une intégration en conséquence dans la facette des organisations.

La normalisation, fondée en partie sur un système à base de règles et donc sur des ressources exogènes, est une étape intermédiaire entre la ressource endogène dont sont tirées les informations intégrées à la facette des organisations, c'est-à-dire le corpus, et la ressource termino-ontologique.

Le système à base de règles, avec le lexique d'amorces sur lequel il s'appuie, permet donc simultanément de traiter la standardisation des noms d'organisations, et leur structuration en hiérarchie le cas échéant. La liste des amorces prend en effet la forme d'un réseau agencé hiérarchiquement, en fonction du niveau que nous avons attribué à chacun des items. Cette structuration hiérarchique permet d'établir des règles de découpage, mais aussi d'ordonnancement des segments obtenus en fonction des amorces qu'ils contiennent.

Par là, le système à base de règles permet de se rapprocher du réel auquel les entités nommées font référence.Utiliser les noms bruts, directement tirés de la ressource endogène qu'est le corpus, ne permet pas de restituer la richesse des entités d'organisations auxquelles il est fait référence. En revanche, la structuration résultant entre autres du système à base de règles dans la facette des organisations permet de représenter, de manière plus fidèle, les objets du réel désignés par les noms : ces objets ne sont pas systématiquement unitaires et simples, mais sont souvent complexes, et hiérarchisés.

7.2.3 Validation et conclusion

L'ensemble de ces traitements est fondé, pour tout ou partie, sur des ressources exogènes, qu'elles soient statiques ou dynamiques. Toutes permettent de compléter les traitements endogènes là où ils présentent des faiblesses et des manques, pour la constitution de la ressource termino-ontologique et par conséquent pour la structuration du corpus.

Chacune de ces ressources exogènes est utilisée pour un ou des type(s) de données spécifique(s). En cela, elles répondent à la structuration en facettes de la ressource : les traitements dédiés permettent la séparation claire des données résultantes, alors intégrées aux facettes distinctes.

Par ailleurs, ces ressources exogènes rendent les processus endogènes décrits dans le chapitre 6 globalement plus efficaces.

D'abord, l'apport de ces ressources permet d'aboutir à des traitements plus complets : le plan des lieux, en particulier, est constitué en grande partie *a priori* grâce à des lexiques fermés, pour les pays, ou du moins limités dans le cas des villes, qui rendent possible une quasi-exhaustivité pour ces unités d'informations. De même, l'enrichissement de ce plan par la mise en correspondance de chaînes de caractères issues des documents avec les items des listes permet de récupérer un maximum d'informations.

D'autre part, l'utilisation du système à base de règles, pour le plan des organisations, et de lexiques, pour le plan des thèmes, permet de tendre vers une plus grande précision. Les règles de normalisation et de hiérarchisation éliminent un grand nombre d'éléments parasites et structurent les entités de manière à ce qu'elles correspondent le plus possible aux objets réels auxquels elles réfèrent. Quant aux lexiques utilisés pour la détection de collocations brutes potentielles, ils éliminent une partie des segments non pertinents grâce à la détection de ceux qui débutent et/ou finissent par un mot grammatical.

En somme, les traitements exogènes améliorent les traitements endogènes portant sur la structuration de la ressource termino-ontologique multi-plans, et de fait, la structuration de l'ensemble documentaire représenté par cette RTO.

Les ressources exogènes que nous avons citées et présentées sont fondées sur des observations des données attestées, mais restent constituées en-dehors de tout apport direct d'un corpus. Pourtant, il existe d'autres ressources exogènes constituées selon un autre mode. La capitalisation des données est un moyen de tirer profit de traitements de différents types effectués sur un corpus donné et d'en utiliser les résultats sur les ensembles documentaires ultérieurs. Puisque ces résultats ont été obtenus à partir d'un autre corpus que celui qui est alors traité, ils constituent bien une ressource exogène, et non une ressource dont le type est le même que celui qui a permis de l'obtenir en premier lieu. Ainsi, un traitement endogène peut produire des résultats sur une ensemble documentaire particulier qui, eux, deviennent une ressource exogène pour les corpus suivants.

Il est à noter que seules sont capitalisées les données qui ont été validées de manière anthropogène à la suite de processus endogènes ou exogènes. Nous présentons le déroulement de cette capitalisation sur la figure 7.5. Par exemple, la segmentation d'un nom d'organisation à entités multiples (voir la sous-section 6.1.2), processus endogène, doit être validée par l'utilisateur pour être intégrée comme donnée capitalisée dans la facette correspondante (voir le chapitre 8, sec-

tion 8.1.2). C'est donc la combinaison entre traitements endogènes ou exogènes d'une part, et traitements anthropogènes d'autre part, qui permet la capitalisation, et transforme le résultat de ces processus en ressource exogène pour les ensembles documentaires suivants.

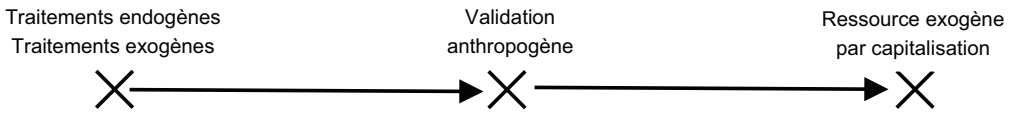

FIGURE 7.5 – Déroulement du processus de capitalisation permettant de générer de nouvelles ressources exogènes

Ces données sont alors utilisées comme toute autre ressource exogène, à ceci près qu'elles sont considérées comme les plus fiables, puisque émanant, à leur dernière étape de constitution, d'une décision humaine.

7.3 Des ressources exogènes pour la projection d'information dans l'immersion documentaire

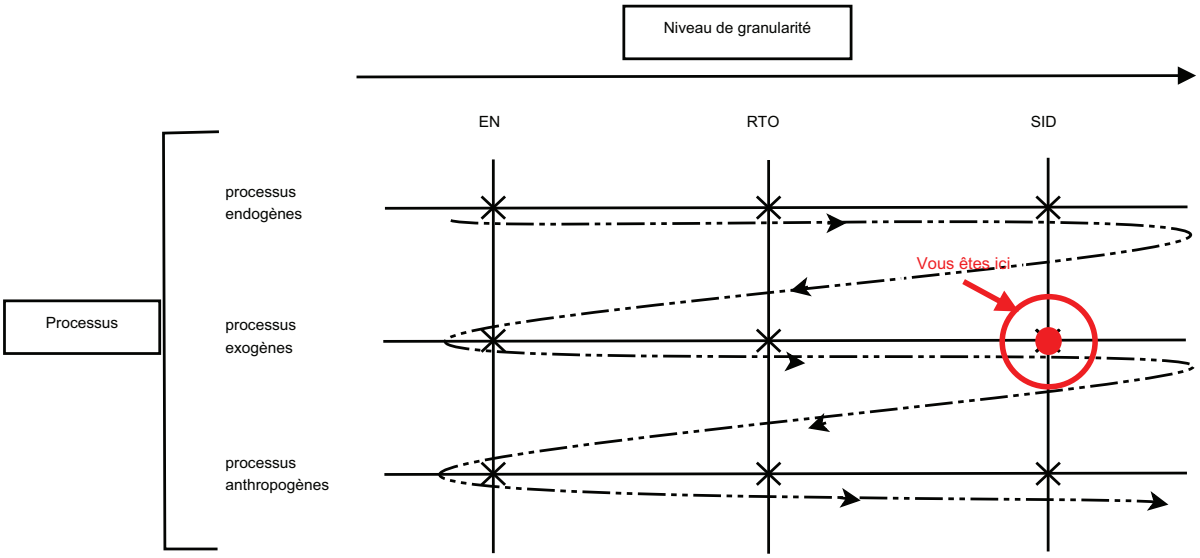

FIGURE 7.6 – Positionnement des traitements exogènes pour le système d'immersion documentaire dans l'ensemble des processus

Il est possible de considérer le système d'immersion, c'est-à-dire le niveau de granularité global dans nos travaux, comme l'interaction de trois types d'informations : celles qui concernent le traitement des documents inclus dans l'ensemble documentaire à étudier du point de vue de son sens, celles qui permettent de rendre ce contenu accessible par visualisation, et enfin, celles qui sont apportées par l'utilisateur pour interpréter le contenu et construire les connaissances. Le premier type d'informations est en grande partie constitué, nous l'avons vu, par la ressource endogène qu'est la ressource termino-ontologique (RTO). Cependant, il est aussi complété par des ressources exogènes. Les informations nécessaires à la visualisation des données sont quant à elles souvent fournies, dans notre cas, par des ressources externes. Dans le présent chapitre, nous abordons l'ensemble de ces ressources et processus exogènes employés au niveau du système d'immersion (voir figure 7.6 page 294) [64].

Du point de vue calculatoire, et donc pour enrichir les données endogènes structurées au sein de la RTO endogène, des ressources externes peuvent être exploitées pour l'immersion. Elles sont un moyen de corriger des biais inhérents à un ensemble documentaire limité, dans lequel la somme des attestations ne peut prétendre à la complétude. De plus, si elles sont correctement structurées pour l'usage visé, le coût engendré par leur mise en place reste faible, et économise des traite-

64. Le troisième type d'informations relève de ressources et processus anthropogènes ; nous les traitons dans le chapitre suivant.

ments relevant d'autres types, pouvant être coûteux. Par ailleurs, dans le cas de listes fermées, l'exhaustivité devient envisageable. Enfin, elles peuvent offrir des possibilités supplémentaires d'exploration d'un ensemble documentaire, toujours en ayant pour base la RTO endogène et les dimensions qui en sont tirées. Dans ce cas, cet enrichissement qu'elles permettent fonctionnent à la manière d'une augmentation de la réalité virtuelle décrite par les dimensions informationnelles. En effet, elles permettent de superposer aux informations endogènes de nouvelles informations qui les complètent, et qui sont visualisables. C'est pourquoi, dans ce qui suit, nous parlerons d'*augmentation exogène des dimensions*.

D'autre part, relativement à la visualisation, des ressources exogènes permettent l'interfaçage final entre les documents et l'utilisateur, et offrent à celui-ci la possibilité d'agir sur ceux-là. Ces ressources externes sont bien entendu incontournables dès lors qu'il s'agit de visualiser graphiquement des informations. Mais il s'agit de savoir quels types de visualisations peuvent être utilisés dans le cadre du système d'immersion tel que nous l'avons conçu, et reposant sur une structure en facettes et sur des dimensions informationnelles. En réalité, il existe une infinité de possibilités pour visualiser une même information. Cependant, certaines sont mieux adaptées que d'autres, et mènent à des représentations pertinentes pour l'utilisateur. La pertinence de ces visualisations est optimisée par le fait que les types d'informations ont été clairement séparés dans les différentes facettes de la RTO, sur laquelle s'appuie l'immersion.

Dans ce qui suit, nous commençons par décrire les ressources exogènes qui peuvent être exploitées pour compléter l'information endogène fournie par la ressource termino-ontologique, de façon à enrichir les possibilités calculatoires du système d'immersion et à augmenter les potentialités des dimensions. Puis nous présentons un court tour d'horizon des types de ressources cartographiques externes existantes, et déterminons les plus adaptées en tant que support de visualisation du système d'immersion.

7.3.1 Ressources exogènes pour augmenter l'information endogène

Chaque dimension informationnelle est exploitée dans le système d'immersion à partir de la facette correspondante dans la RTO multi-plans.

A chaque facette peut être associée une ressource exogène, pour l'enrichir, la compléter. Celle-ci peut être disponible de manière générale, comme dans le cas d'informations géographiques sous forme de lexiques par exemple. Elle peut aussi avoir été conçue en interne au sein de TKM, notre terrain applicatif, spécifiquement pour le système d'immersion ou pour d'autres outils. Enfin, un client peut également fournir une telle ressource, comme une ontologie de domaine constituée par une entreprise cliente par exemple.

L'utilisation de ressources exogènes offre une multitude de possibilités d'immersion. Dès l'instant où une représentation d'informations est disponible et adaptable à la structure et au contenu

d'au moins une des dimensions, elle devient exploitable. Les points d'entrée, correspondant à la partie « Je veux voir » de la requête d'un utilisateur (voir section 6.3 page 222), peuvent alors être positionnés en fonction des informations de la ressource exogène utilisée et de sa structure.

Puisqu'une partie des systèmes d'immersion seront fondés sur cette ressource, quelle qu'elle soit, et quel que soit son type, la qualité des résultats obtenus est directement corrélée à sa propre qualité, et en particulier à celle de sa structure de représentation.

Quoi qu'il en soit, ce processus permet de désigner les éléments des dimensions endogènes sous d'autres termes que ceux présents dans les documents, et d'organiser différemment l'information. Cette organisation est fonction des éléments présents dans les ressources exogènes.

Dans la section 6.3 page 222, nous avons présenté les dimensions de l'immersion, selon le type de leur structure et les possibilités de calculs et/ou de visualisation qu'elles offraient. Les dimensions énumératives non ordonnées (ENO), qui correspondent aux facettes des auteurs et des thèmes dans la RTO, permettent des regroupements ou la prise en compte unitaire de leurs éléments. La dimension énumérative ordonnée (EO), celle des dates, est exploitable par les mêmes traitements que les dimensions ENO, auxquels s'ajoutent des possibilités de traitement en segments, c'est-à-dire par intervalles temporels. Enfin, les dimensions hiérarchisées, celle des organisations et celle des lieux, cumulent l'ensemble des traitements envisageables pour les autres types de listes, ainsi que la possibilité de les traiter comme des hiérarchies.

Des ressources exogènes, appliquées à ces dimensions endogènes de base, permettent de leur apporter un ensemble de données supplémentaires et d'augmenter les calculs et visualisations potentielles au cours de l'immersion documentaire. Elles permettent principalement, selon leur nature et le type d'informations qu'elles contiennent, d'ordonner les listes ENO, de hiérarchiser des listes ENO ou EO, et de mettre en réseau des listes ENO, EO ou hiérarchisées.

7.3.1.1 Augmentation des listes énumératives non ordonnées par ajout de segments

Les listes ENO, c'est-à-dire les dimensions des auteurs et des thèmes, ne sont pas ordonnées, ainsi que nous l'avons expliqué en 6.3 page 222 : elles sont constituées d'éléments sur lesquels sont possibles des sélections unitaires, ou bien des regroupements sur la base d'équivalents d'expressions régulières. Or, une ressource exogène peut permettre de les traiter comme des listes EO, si la structure qu'elle apporte fonctionne par segments : par exemple, une liste de certains des auteurs présents dans l'ensemble documentaire, ordonnée par genre masculin ou féminin, ou en fonction de leur année de naissance, pourrait dans l'absolu être utilisée pour ordonner les éléments de la dimension des auteurs en fonction de ces critères [65]. Dans ce cas, la dimension est

65. Les exemples que nous citons pour cette dimension sont des cas d'école, présentés pour illustrer notre propos et montrer les possibilités du modèle. Cependant, les informations relatives aux personnes en tant qu'individus sont contrôlées par la loi Informatique et Liberté. Les cas que nous présentons ne sont donc ni légaux, ni souhaitables. En conséquence de quoi, la facette des auteurs ne sera jamais associée, dans les faits, à d'autres informations que celles qui sont autorisées par la loi.

augmentée par segmentation de ses unités : les ressources exogènes jouent le rôle de structures englobantes, qui intègrent au sein de segments ou intervalles les unités d'informations des listes ENO.

7.3.1.2 Augmentation des listes énumératives par ajout d'arborescences

Certaines ressources exogènes, si elles sont organisées en arborescences, permettent de structurer hiérarchiquement des listes énumératives, qu'elles soient ou non ordonnées. Les dimensions concernées sont donc en l'occurrence les auteurs et les thèmes, ainsi que la dimension temporelle. Toutes présentent, par défaut, une structure plate - ou presque en ce qui concerne les dates.

Les informations qu'elles apportent peuvent être augmentées par l'apport d'arborescences externes. En ce qui concerne la dimension ordonnée des dates, une arborescence structurée par périodes historiques, entre autres possibilités, peut être envisagée pour une répartition des éléments de dates au sein de hiérarchies plus ou moins profondes. Par exemple, une période très large telle que l'après-guerre peut être scindée en périodes plus précises, comme les Trente Glorieuses, suivies de la période s'étendant du premier choc pétrolier à la chute du mur de Berlin, puis de celle partant de ce même point et s'étendant jusqu'aux attentats de 2001. Les dates issues de la dimension temporelle endogènes peuvent alors venir se placer dans cette hiérarchie, répartissant alors les documents eux-mêmes en fonction de leur contexte historique de production, si cette structuration s'avère pertinente pour l'utilisateur.

De même, une ontologie de domaine par exemple peut être exploitée pour structurer les thèmes, représentés par les collocations brutes, au sein de concepts de niveau plus ou moins élevé.

Dans tous les cas, de telles ressources importées dans le système d'immersion modifient la représentation des dimensions, en leur attribuant une profondeur là où il n'en existait pas auparavant.

7.3.1.3 Augmentation des trois types de listes par ajout de réseaux

Des ressources exogènes structurées en réseaux ou en hiérarchie peuvent également permettre la création de réseaux pour les trois types de listes, qu'elles soient énumératives ou hiérarchisées.

Nous précisons que, bien qu'une hiérarchie prenne la forme d'une arborescence, et qu'à ce titre elle soit une forme particulière de réseau, nous distinguons celle-là de celui-ci, car les réseaux permettent des représentations plus riches du point de vue de l'information que les arborescences.

De plus, cette augmentation par création de réseaux peut avoir lieu à l'intérieur d'un plan donné, ou bien dans une combinaison de plans. Par exemple, la projection d'une arborescence externe sur l'arborescence de l'une des dimensions hiérarchiques permet d'obtenir un réseau.

Ainsi, un réseau peut par exemple être établi au sein de la dimension hiérarchisée des orga-

nisations par l'apport d'une ressource externe structurée en réseau, contenant des informations sur des sociétés possédant des parts sociales d'autres sociétés. La projection de ce réseau externe sur la hiérarchie dimensionnelle permet de dégager les organisations présentes dans l'ensemble documentaire qui entretiennent des relations de propriété réciproque de parts. Le même raisonnement peut être appliqué sur toute dimension, comme sur celle des auteurs, et permet alors d'établir des réseaux de personnes sur des critères donnés si une ressource externe contient de telles informations.

De même, un réseau externe peut être appliqué pour augmenter plusieurs plans à la fois, par exemple pour déterminer les auteurs de l'ensemble documentaire travaillant pour des organisations données, et pendant une période donnée, et pas seulement à un moment t indiqué par une date de publication.

7.3.1.4 Conclusion

Jusqu'ici, en croisant entre elles les différentes dimensions endogènes tirées de la RTO, des réseaux pouvaient être obtenus.

Mais en apportant de l'information supplémentaire dans une dimension donnée, grâce à des ressources exogènes, il devient possible de créer des réseaux en son sein, à partir du plan correspondant. La structure de la dimension s'en trouve augmentée. De fait, en augmentant une dimension, c'est l'ensemble de l'immersion qui est améliorée. Il convient de souligner que cet ajout représente une potentialité, et est donc facultatif : l'utilisateur peut souhaiter s'en tenir aux dimensions endogènes, s'il n'a pas de besoins supplémentaires et/ou si aucune ressource externe n'est disponible. De plus, tirer parti d'une ressource externe quelle qu'elle soit implique que le lien entre ressource endogène et ressource exogène soit effectivement possible : l'intersection des nœuds de l'une et de l'autre doit être non nulle.

Ainsi, étant donnée la structure en facettes de la RTO dont naissent les dimensions, il est envisageable d'intégrer des données externes pour les compléter et les augmenter. D'un point de vue pratique, la ressource externe fournie doit être implantable dans la dimension correspondante. Il est donc nécessaire de prévoir son intégration à part entière dans le système, par le biais de formalismes adaptés. En effet, sa structuration et son contenu doivent permettre de formaliser les liens existant entre les unités informationnelles des dimensions, à travers les facettes de la RTO, et les éléments de la ressource externe.

7.3.2 Les ressources exogènes comme appui à la représentation

Notre système d'immersion passe par la visualisation des données issues des dimensions, en fonction d'arguments de requêtes précis déterminés par l'utilisateur. Ces visualisations sont possibles uniquement si des dispositifs de visualisation graphique sont disponibles, et interfacées avec

les données structurées. Des ressources et outils exogènes sont exploités pour cette représentation graphique, puisqu'elle passe par l'apport d'informations graphiques et cartographiques forcément externes à l'ensemble documentaire traité.

Toutefois, les choix cartographiques pour la visualisation d'informations ne vont pas de soi : ils nécessitent de prendre en compte un ensemble de paramètres influant tous sur l'efficacité de la visualisation dans un contexte donné.

Dans ce qui suit, nous présentons d'abord la notion de visualisation. puis nous abordons les différents types de cartographie existants, en fonction des paradigmes définis par [Tricot, 2006]. Enfin, nous faisons le point sur les types de cartographie adaptés à notre système d'immersion.

7.3.2.1 La notion de visualisation : représenter visuellement les données d'un ensemble documentaire

La visualisation d'information est un champ de recherche à part entière, intégré à l'étude de l'interaction homme-machine. Elle travaille à répondre à la question suivante [Tricot, 2006] :

Comment représenter un grand nombre d'informations sur un écran ?

Des éléments de réponse se trouvent dans la définition que donnent [Card et al., 1999] (cités par [Tricot, 2006]) de la visualisation, de manière générale :

> « La visualisation est l'utilisation de représentations visuelles interactives et informatisées de données pour amplifier la cognition. »

Les auteurs précisent cette définition du point de vue plus spécifique de la visualisation d'information *ibid.* :

> « La visualisation **d'information** est l'utilisation de représentations visuelles interactives et informatisées de données **abstraites** pour amplifier la cognition. »

La visualisation d'information est donc plus spécifiquement dédiée à des informations abstraites, qui ont la particularité de ne pas avoir de représentation graphique intrinsèque.

Dans tous les cas, les points ressortant de ces deux définitions sont d'abord le support informatisé des visualisations, et d'autre part leur caractère interactif. Par conséquent, elles appellent l'intervention des utilisateurs. Enfin, leur objectif est d'augmenter la cognition humaine, c'est-à-dire de faciliter l'appréhension d'informations par les utilisateurs, et ce particulièrement lorsque ces informations sont en grand nombre.

L'amplification de la cognition nécessite de mettre en place des représentations graphiques qui permette de mieux embrasser l'ensemble des informations à traiter. D'après [Fekete, 2010] :

> « La visualisation d'information permet d'utiliser des représentations graphiques faciles à saisir, immédiates à percevoir, et qui nous aident à comprendre des phénomènes complexes, à voir des tendances et à prendre rapidement des décisions appropriées. »

Pour cela, la visualisation peut exploiter la faculté humaine de perception préattentive : cette caractéristique du système visuel humain a été mise en avant par [Treisman, 1986] (citée par [Fekete, 2010]). Elle permet de percevoir certaines configurations visuelles de manière instantanée, sans effort, et avec précision. L'auteur donne l'exemple de la perception immédiate d'un point rouge placé au sein d'un ensemble de points bleus, et ce quel que soit le nombre de ces derniers.

Se fonder sur ces caractéristiques préattentives est un moyen de construire des visualisations d'informations adaptées au système perceptif humain, et lui permettant donc d'augmenter avec un minimum d'effort sa cognition. La perception visuelle est facilitée, et la mémoire de travail est moins sollicitée et peut se concentrer sur la résolution de problèmes complexes.

La question à se poser est alors de savoir quelles représentations sont adaptées pour véhiculer un type de message donné, ou en tout cas un type d'informations donné pouvant donner lieu à une interprétation.

[Card et al., 1999] (cités par [Tricot, 2006]) ont établi un modèle de la visualisation d'information, qui fait aujourd'hui référence dans la communauté. Nous le reproduisons dans la figure 7.7 page 300.

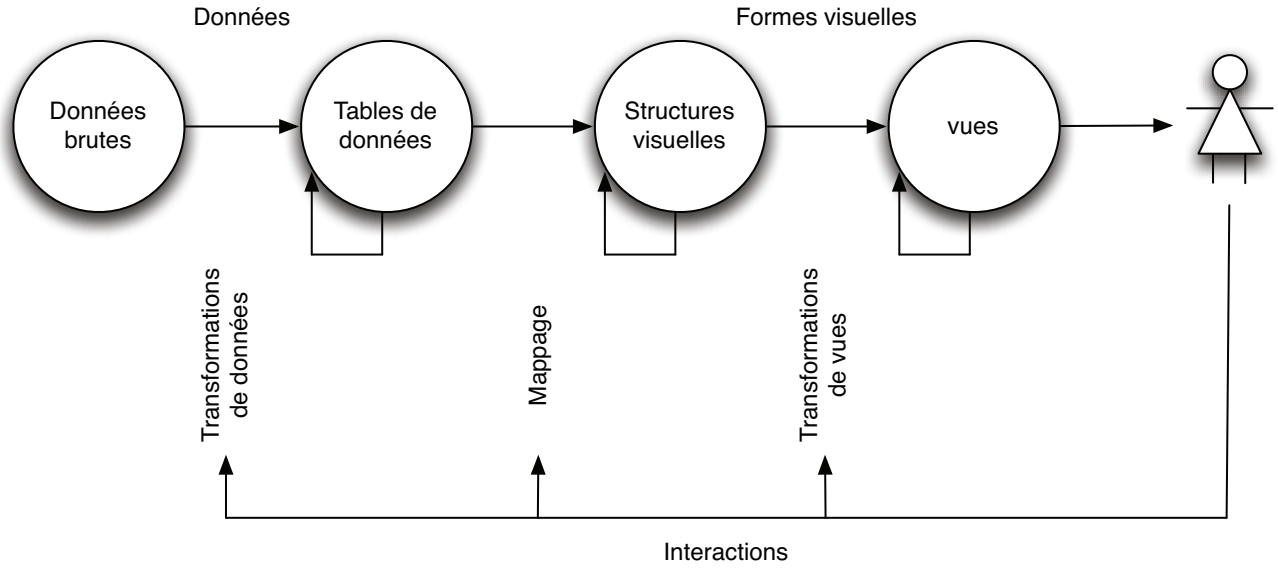

FIGURE 7.7 – Modèle de référence de la visualisation d'informations selon [Card et al., 1999]

Les opérations permettant de passer des données brutes, pour nous notre ensemble documentaire, aux vues, c'est-à-dire les éléments visuels effectivement présentés à l'utilisateur dans le système d'immersion, sont en fait une série d'abstractions et de transformations effectuées à chaque étape. Un agent humain peut intervenir à chacune de ces phases d'abstraction.

Les structures visuelles sont des objets graphiques, auxquels sont associés les éléments de la table de données. Cette association est réalisée par mappage, c'est-à-dire une « opération qui permet de passer d'un monde de données à un monde de formes visuelles » [Tricot, 2006]. Les vues effectivement présentées à un utilisateur sont formées à partir de ces structures.

[Tricot, 2006] s'inspire de ce modèle pour établir trois niveaux dans la visualisation de données abstraites, allant de l'information à sa visualisation, qui permet à des organisations de maîtriser leur espace informationnel pour leur activité. Nous présentons ces niveaux dans la figure 7.8.

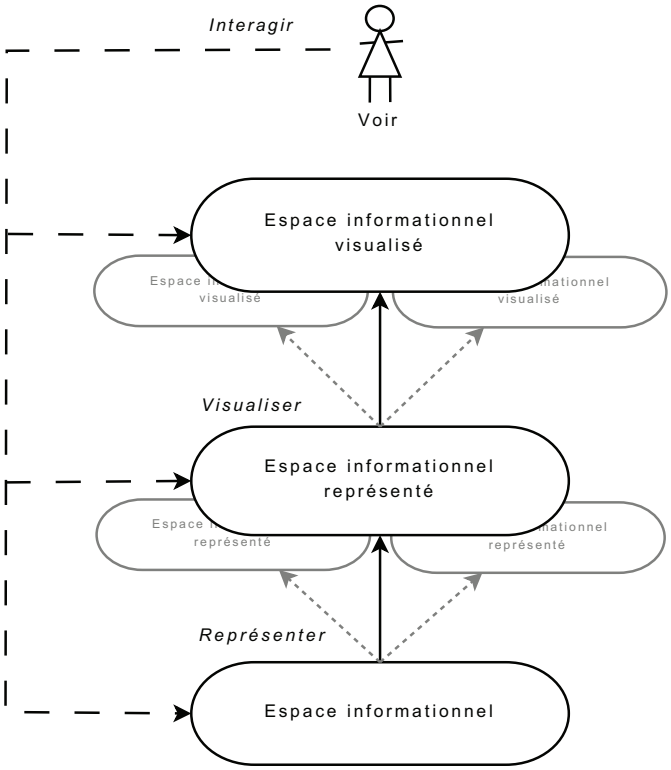

FIGURE 7.8 – Les niveaux de la cartographie de données abstraites selon [Tricot, 2006]

Sommairement résumé, un espace informationnel, composé de l'ensemble des documents et informations rassemblés au sein d'une organisation donnée, connaît un premier niveau d'abstraction afin d'en tirer une représentation ; puis cette représentation, composée de structures graphiques renvoyant aux informations de l'espace informationnel de départ, donne lieu à l'espace informationnel effectivement visualisé par un utilisateur.

L'auteur met en avant le fait qu'à chaque passage au niveau supérieur, un grand nombre de possibilités sont offertes quant aux modes de transformations.

Toutes ces possibilités résultent en des visualisations différentes, non obligatoirement pertinentes pour l'utilisateur.

7.3.2.2 De la cartographie géographique à la cartographie de données abstraites

Puisque l'espace informationnel qui nous occupe est un ensemble documentaire, les informations à exploiter sont d'ordre textuel. La navigation parmi ces documents a des particularités, notamment du point de vue de l'interaction avec l'utilisateur, qu'il convient de prendre en compte dans le choix des visualisations adéquates. [Shneiderman, 1996] (cité par [Roy, 2007]) recense sept critères à respecter pour permettre une visualisation efficace de tels ensembles.

- avoir un aperçu global de l'ensemble de documents ;
- pouvoir zoomer sur une collection d'éléments ;
- pouvoir supprimer certains éléments sur différents critères ;
- pouvoir obtenir des détails sur demande sur un groupe d'éléments ;
- voir les relations entre éléments ;
- pouvoir extraire des sous-collections d'éléments sur certains critères ;
- consulter un historique des actions réalisées.

Ces critères reposent sur un principe général de consultation cyclique d'un ensemble documentaire : l'utilisateur part de la vue d'ensemble, sur laquelle il remarque des éléments d'intérêt ; il peut se focaliser dessus à l'aide de zooms ou de filtres. Puis, une fois les détails consultés, il doit être possible de revenir à la vue d'ensemble, et d'itérer la même procédure [Fekete, 2010]. Ce principe cyclique influence donc forcément la sélection de moyens de visualisation, au niveau de la représentation de l'espace informationnel comme au niveau de la visualisation elle-même.

[Tricot, 2006] fait un état de l'art très riche de différentes méthodes de visualisations, et considère les cartes, au sens large, comme « un médium idéal entre un grand nombre d'informations et l'esprit » qu'il considère comme appartenant au paradigme de la cartographie. Nous nous fondons sur cet état de l'art pour développer ce qui suit.

Selon l'auteur, la cartographie ne s'applique donc pas seulement aux données géographiques, mais permet de visualiser ce qu'il nomme des « données abstraites » (*ibid.*) formant l'espace informationnel d'une organisation. Il est rejoint en cela par [Roy, 2007], qui établit trois catégories de cartes en fonction des entités à représenter, à partir des catégories de visualisation du site Places & Spaces[66]. Selon lui, les « cartes géographiques » mettent en évidence des territoires, les « cartes conceptuelles » représentent des entités immatérielles, et les « cartes de domaine » représentent des entités physiques de tout type, partageant des points communs relativement à un domaine donné, au sens large : les diverses communautés vivant à New York, les étoiles visibles depuis un point précis du globe à un moment t, ou encore des co-auteurs dans un domaine scientifique donné par exemple. La cartographie dépasse donc très largement les données géographiques, puisque des informations de diverses natures peuvent venir se placer sur une carte. A ces données de différentes sortes correspondent des représentations spécifiques : une

66. Adresse web de Places & Spaces : `http://www.scimaps.org/maps/browse/`

mappemonde par exemple n'est pas pertinente pour certains types d'informations.

A ce titre, [Tricot, 2006] considère des visualisations telles que des tables de données par exemple comme des cartes. Ce n'est pas le point de vue de [Roy, 2007], qui perçoit une distinction entre cartographie et autres visualisations. Pour lui, la différence se situe dans « la notion de distance entre éléments de l'ensemble analysé ainsi que [dans] les notions de vue globale et d'interactivité », qui seraient l'apanage des cartes. Nous préférons la définition de [Tricot, 2006], qui autorise plus de possibilités de visualisations, exploitables dans un grand nombre de situations. Cependant, les cartes doivent répondre aux conditions dans lesquelles se place la visualisation d'information : il n'est pas question de fournir n'importe quelle carte dans n'importe quelles circonstances, sans faire des choix raisonnés concernant les vues proposées.

En particulier, les cartes choisies doivent correspondre aux besoins des utilisateurs, que [Tricot, 2006] répartit en trois catégories : tout d'abord, les cartes choisies doivent autoriser la navigation selon la sémantique du domaine. Celle-ci renvoie « aux concepts permettant d'appréhender le domaine » (*ibid.*), et permet donc à l'utilisateur de comprendre et exploiter plus facilement une carte.

D'autre part, la cartographie doit proposer une vision à plusieurs échelles : un espace informationnel doit pouvoir être appréhendé dans sa globalité comme dans ses particularités. Ce besoin utilisateur fait écho aux critères de [Shneiderman, 1996].

Enfin, la carte proposée doit être adaptée à l'utilisateur, en fonction de son activité spécifique et de son niveau d'expertise.

Pour parvenir à la carte vue par l'utilisateur à partir de l'espace informationnel, plusieurs types de paradigmes rentrent en jeu et définissent le type final de visualisation. Nommément, il s'agit des paradigmes de représentation, de visualisation, et d'interaction [Tricot, 2006], qui correspondent aux trois niveaux de visualisation de l'espace informationnel présentés dans la sous-section précédente : les paradigmes de représentation permettent d'associer aux éléments de l'espace informationnel de départ des structures visuelles ; les paradigmes de visualisation déterminent la façon dont l'espace informationnel ainsi représenté sera visualisé dans une carte ; les paradigmes d'interaction sont utilisés pour définir les actions que l'utilisateur peut réaliser à partir de la carte.

Les paradigmes de représentation relèvent globalement de trois types : les représentations orientées valeurs, les représentations orientées relation, et les représentations arborescentes. Nous présentons ici un certain nombre d'exemples de paradigmes de *représentation*. Bien qu'ils ne permettent pas à eux tout seuls la *visualisation* des données, nous ne pouvons illustrer ces paradigmes sans passer par des visualisations qui peuvent en résulter. Les figures qui suivent sont donc des représentations *visualisées*.

Les premières sont utilisées lorsque les données à représenter sont, comme leur nom l'indique, des valeurs associées aux entités de l'espace informationnel. Ces représentations prennent souvent

la forme, au moment de la visualisation, de tables de données ou matrices. La figure 7.9 présente une matrice issue des outils TKM.

FIGURE 7.9 – Exemple de représentation orientée valeurs par matrice

Les nombres indiqués dans les cellules de la matrice correspondent aux documents se trouvant au croisement de deux valeurs. Par exemple, 127 documents remplissent à la fois le critère de la première ligne et celui de la première colonne.

Les représentations orientées relations sont très courantes, et sont exploitées pour représenter des données modélisées comme « un ensemble d'entités liées par des relations binaires » (*ibid.*). Dans ce cadre, les éléments, entités et relations, peuvent être représentés sous la forme de graphes, représentés ensuite soit par « nœuds-liens », orientés ou non, soit par structures matricielles. Nous présentons en figure 7.10 un graphe « nœuds-liens », établissant un réseau de collaboration entre IBM, Microsoft et d'autres sociétés qui leur sont reliées, réalisé par la R&D de la société TKM.

Enfin, les représentations arborescentes sont un type particulier de graphe, dont chacun des éléments possède un père au plus. De fait, toutes les techniques de représentation de graphes sont applicables aux arborescences. [Tricot, 2006] recense notamment les représentations en liste indentée (figure 7.11) [67], les représentations en « nœud-lien » d'arbres (figure 7.12), et enfin, les représentations en pavage, comme les Tree Maps. Le principe de ces dernières « consiste à découper en surface rectangulaire l'espace de la carte proportionnellement à chaque sous-arbre » [Tricot, 2006]. Les rectangles les plus petits renvoient aux feuilles de l'arbre, et peuvent à leur tour former des rectangles plus importants, représentant des nœuds plus hauts dans l'arborescence. Elles permettent de représenter de grands arbres, et de focaliser l'attention de l'utilisateur sur les feuilles. Sur la figure 7.13 page précédente, nous présentons une Tree Map permettant de visualiser l'occupation de l'espace du disque dur d'un ordinateur : chaque rectangle individuel représente un fichier, et leur accumulation renvoie à des répertoires, plus ou moins haut dans l'arborescence. Les couleurs renvoient quant à elles aux types des fichiers.

Les paradigmes de visualisation sont appliqués sur l'espace informationnel représenté, et permettent d'arriver, *in fine*, aux vues effectivement présentées aux utilisateurs. A partir d'un espace représenté, les possibilités de visualisation sont multiples. Elles se répartissent au sein de deux grands types de techniques : les visualisations uniformes ne déforment pas l'espace de la carte, mais peuvent faire intervenir des transformations affines. Ces dernières regroupent

67. Comme pour la sous-section précédente, les illustrations issues des outils TKM ont été nettoyées des données confidentielles qui pouvaient s'y trouver. Dans le cas de la liste indentée, des dossiers fictifs ont été créés, correspondant cependant à des structures effectivement observées.

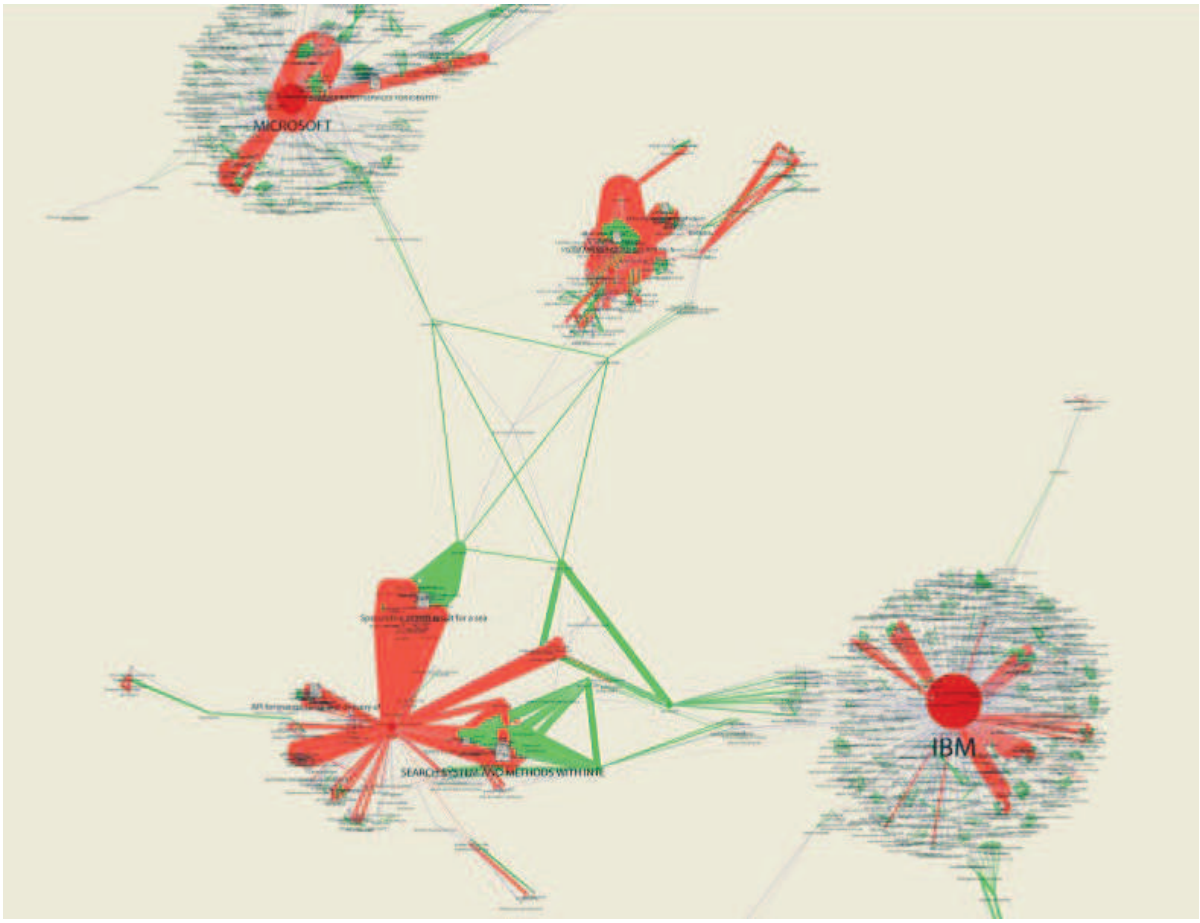

FIGURE 7.10 – Exemple de représentation orientée relations par graphe « nœuds-liens »

des opérations comme des translations, des rotations ou l'application d'un facteur de zoom par exemple. Les visualisations non uniformes déforment l'espace projeté pour visualiser un plus grand nombre de structures visuelles.

Enfin, les paradigmes d'interaction s'appliquent, comme leur nom l'indique, aux cartes interactives, où l'utilisateur passe du rôle de « spectateur » à celui d'acteur. L'interaction peut concerner le niveau de l'espace informationnel brut, et permettre de sélectionner uniquement certaines données et/ou de les modifier directement par exemple. Elle peut aussi influencer la représentation de cet espace, par l'application d'actions à certaines sélections de structures visuelles. Enfin, l'interaction au niveau de la visualisation, notamment par des mouvements de caméra et le contrôle de point de vue.

La cartographie de données offre donc une infinité de possibilités pour la visualisation d'un espace informationnel déterminé, selon les trois niveaux de paradigmes. Cependant, il convient de respecter la sémantique des informations à visualiser, afin de faciliter la compréhension de l'utilisateur.

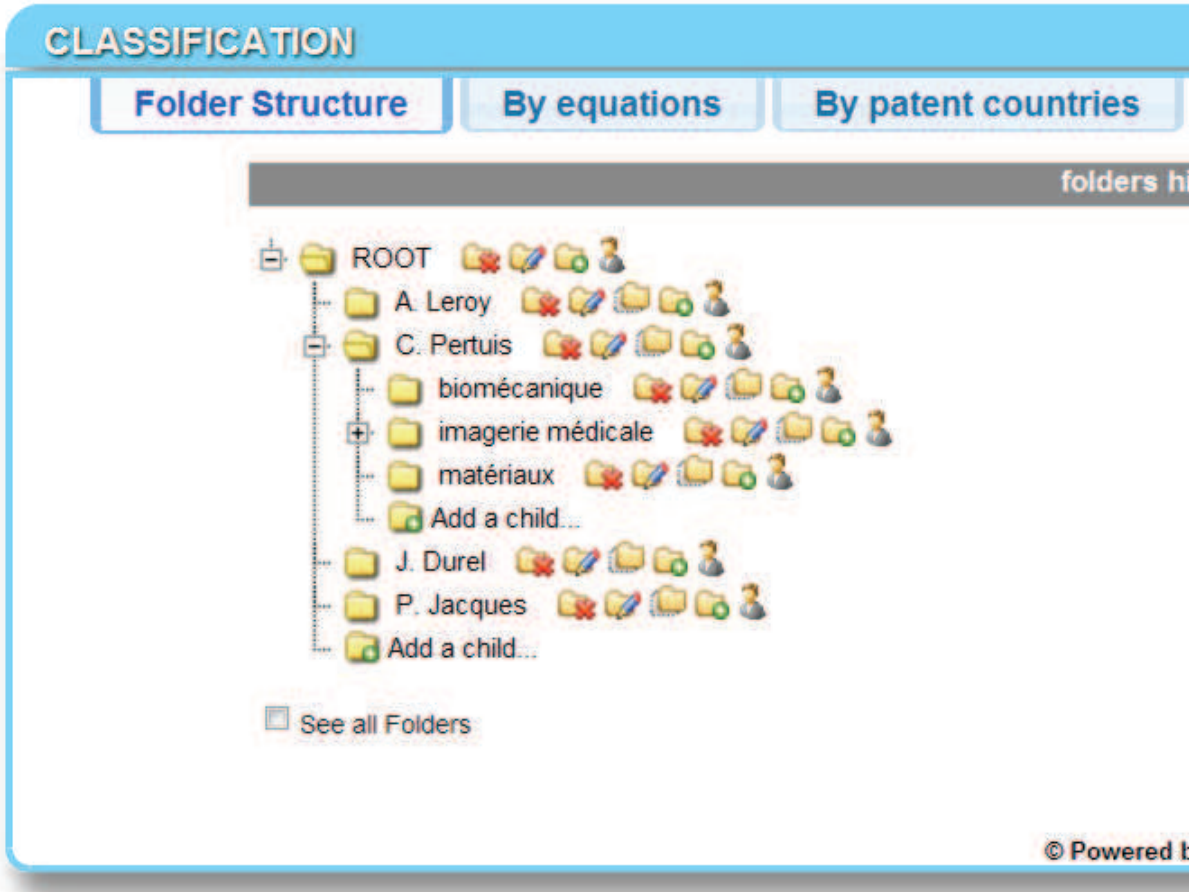

FIGURE 7.11 – Exemple d'arborescence sous forme de liste indentée, issue des outils de la société TKM.

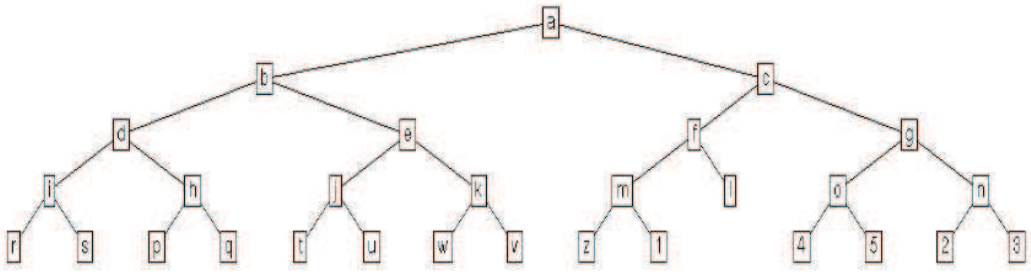

FIGURE 7.12 – Algorithme de dessin d'arbre de [Hascoët & Beaudouin-Lafon, 2001], présenté dans [Tricot, 2006]

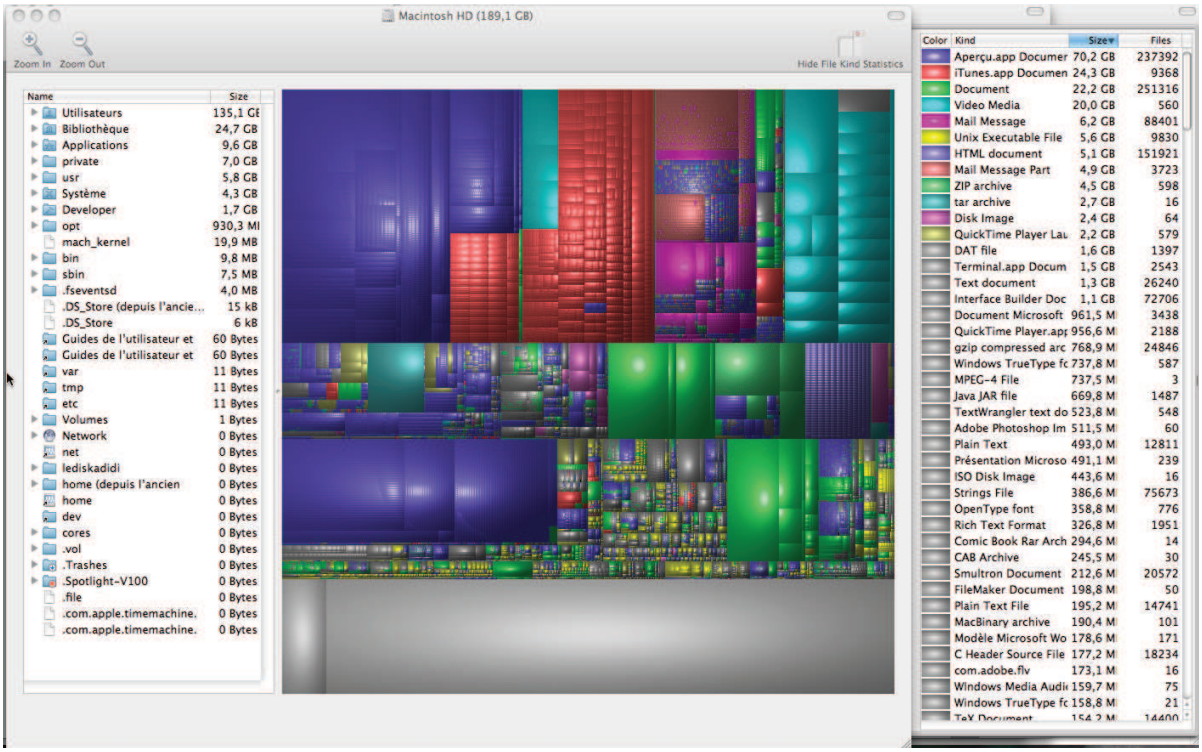

FIGURE 7.13 – Exemple de Tree Map, représentant l'occupation de l'espace d'un disque dur.

Dans notre cas, et celui d'un ensemble documentaire, les besoins sont variés, et plusieurs types de cartes peuvent donc être envisagées.

7.3.2.3 La cartographie pour le système d'immersion documentaire

L'espace informationnel que nous traitons est un ensemble documentaire, qui n'est jamais le même au fil des missions. Cependant, il est nécessaire de définir des types de cartographie suffisamment génériques et variés pour être applicable à chacun de ces ensembles. Quoi qu'il en soit, puisqu'il est composé de données textuelles, un tel ensemble doit être abordé grâce aux critères de fonctionnalités énoncés par [Shneiderman, 1996].

La définition des types de cartographie est aussi raisonnée en fonction des types d'informations que nous avons identifiés comme pertinents, et de la ressource termino-ontologique (RTO) endogène qui en a été tirée. En effet, c'est elle qui établit à la fois les unités et les relations entre unités qui doivent être visualisées.

Ainsi, ce que [Tricot, 2006] nomme la « sémantique du domaine » correspond dans notre cas aux informations contextuelles et thématiques de notre RTO.

D'autre part, le support que nous employons est également à prendre en compte. En l'occurrence, l'immersion est réalisée sur un écran d'ordinateur, et ce en deux dimensions.

En fonction de l'ensemble de ces caractéristiques, nous avons dégagé des types de cartographies adéquats pour notre objectif d'immersion parmi les différents paradigmes.

Au niveau de la représentation de l'ensemble documentaire, qui passe en fait plus précisément par la représentation de l'ensemble documentaire structuré par la RTO, la totalité des paradigmes sont utilisables. En effet, en fonction des dimensions sélectionnées par l'utilisateur dans sa requête, mais aussi de leur nombre et de leur combinaison, chacun des types de représentation s'avère pertinent.

Cependant, les représentations orientées relations sont les plus appropriées dans la majorité des cas envisagés. En effet, puisque notre RTO établit des relations entre les facettes et leurs éléments, ainsi qu'entre éléments d'une même facette, la projection des dimensions correspondantes implique une représentation en réseau des données sélectionnées. Par conséquent, les représentations « nœuds-liens » sont privilégiées. Des vues telles que celle de la figure 7.14 page suivante, issue des outils TKM actuels, sont pertinentes.

Sur cette carte, un réseau de collaboration est établi entre plusieurs organisations, toutes représentées par leur nom. Les entreprises sont symbolisées par des casques, tandis que les organismes sont figurés par des jeunes diplômés. Les points d'interrogation indiquent les organisations non typées. Le pays d'origine est lui aussi présent, sous la forme de drapeaux. Cette vue permet donc d'établir des liens de collaboration caractérisés par le type des organisations, mais aussi par leur pays d'origine, à travers les documents dont les organisations sont « copropriétaires ».

Les représentations arborescentes sont elles aussi utilisées, puisque des dimensions comme celle des organisations par exemple peuvent être l'objet de l'argument de visualisation de la requête. Dans ce cas, les informations sont réparties en fonction des organisations auxquelles elles se rattachent, et ces dernières sont structurées hiérarchiquement si cela est demandé par l'utilisateur. Les représentations « nœud-lien » sont préférables, puisqu'elles placent visuellement les entités hiérarchiquement inférieures en les plaçant sur la carte en-dessous de leur entité-mère, à la manière de la vue présentée en figure 7.15 [68]. En cela, elles restituent sur la carte la représentation mentale que peut avoir l'utilisateur des relations entre une organisation mère et ses organisations filles.

Cette carte représente la hiérarchie de la société Atos Origin. Intégré à notre système, un tel type de représentation hiérarchique permet à un utilisateur de moduler l'affichage des détails d'un élément d'information qui en englobe d'autres. Il peut par exemple décider de ne considérer que la maison-mère, en l'occurrence Atos Origin, ou de descendre dans la hiérarchie de ses filiales s'il a besoin de situer KPGM Consulting ou SEMA parmi elles par exemple.

A ces paradigmes s'ajoutent les représentations sur cartes géographiques, puisque notre ensemble documentaire est également structuré en fonction de données de localisation spatiale. Les

68. Cette vue est issue de Wikipedia pour l'entrée *Atos Origin* (`http://fr.wikipedia.org/wiki/Atos_Origin`), et non pas d'une étude TKM.

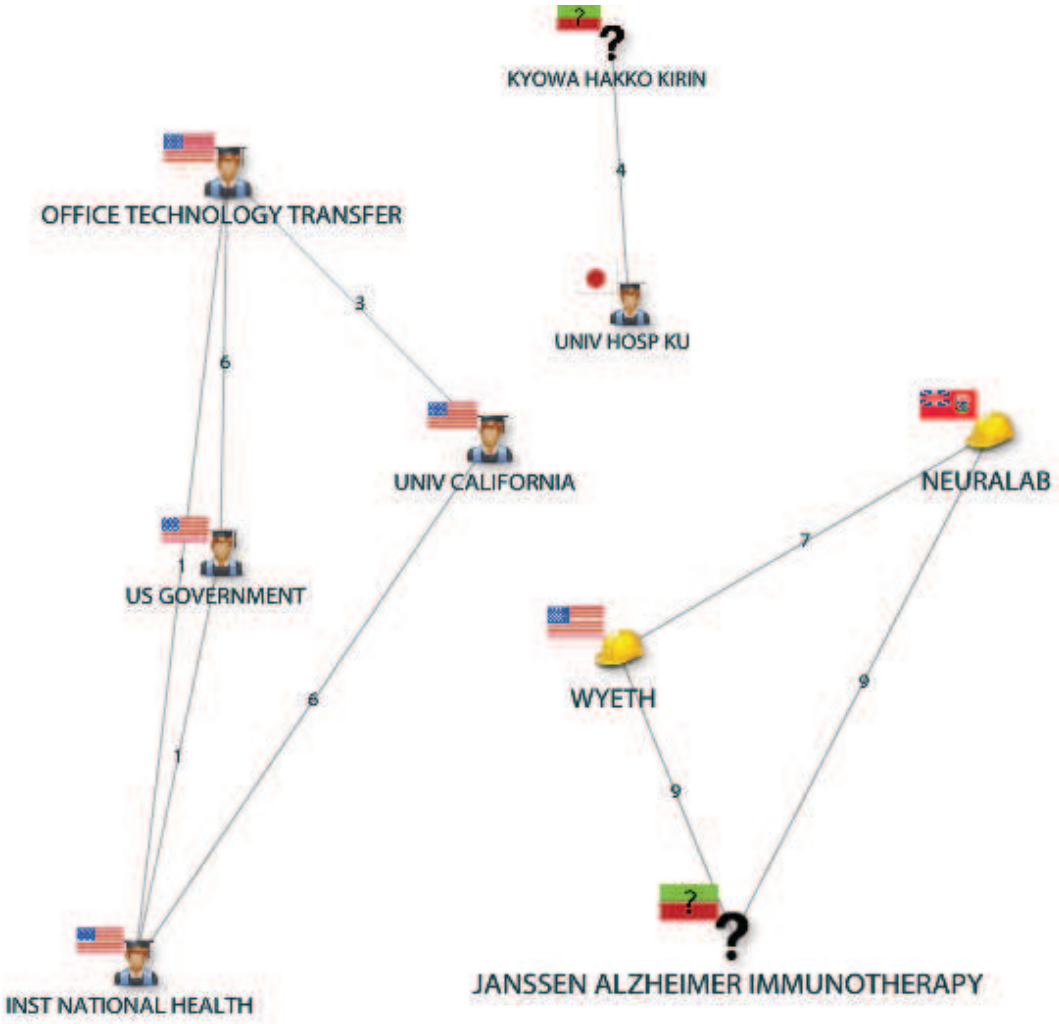

FIGURE 7.14 – Exemple de vue représentant un réseau de collaboration, sous forme de graphe nœud-lien.

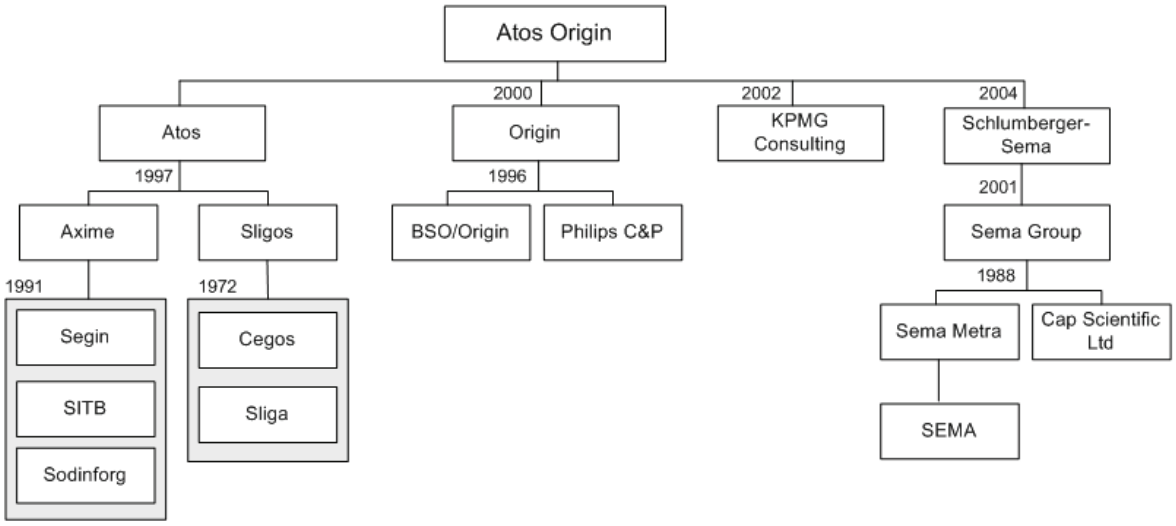

FIGURE 7.15 – Exemple de vue arborescente sous forme de graphe nœud-lien hiérarchisé

données abstraites ne sont donc pas les seules à intervenir dans le traitement des informations, et il convient d'en tenir compte. Cependant, une carte géographique peut également porter des données abstraites, si tant est qu'elles soient représentées en fonction d'un critère de spatialisation. Nous présentons en figure 7.16 page précédente un exemple de carte géographique sur laquelle les organisations propriétaires de documents sont réparties par pays d'origine. La taille des cercles est proportionnelle au nombre des organisations présentes. Cette vue est exploitable dans le cas où l'utilisateur aurait besoin d'identifier les pays les plus présents sur une technologie donnée.

Parmi les paradigmes de visualisation, les deux techniques peuvent *a priori* être envisagées, l'important étant avant tout de permettre à l'utilisateur autant de déplacements qu'il le souhaite au sein d'une carte, ainsi que des allers-retours entre niveau global et niveau local.

Cependant, dans un premier temps, nous nous attachons aux paradigmes de visualisation uniforme, de manière à ne pas déformer l'espace de la carte, et à laisser l'utilisateur évoluer de la manière la plus naturelle possible dans une carte donnée. L'utilisateur a donc la possibilité de faire évoluer son point de vue sur un espace informationnel représenté, à l'aide de transformations affines. Celles-ci permettent notamment l'application d'un facteur de zoom, la translation, la rotation, ou la sélection d'un espace plus restreint que celui initialement visualisé.

Les vues multiples sont également utilisées, et permettent d'éviter ou en tout cas de limiter la surcharge cognitive due au changement fréquent de point de vue sur une représentation. Dans ce cas, des vues globales et des vues locales sont présentes à l'écran simultanément, et reliées et coordonnées les unes aux autres. Grâce à cette co-présence, l'utilisateur est toujours en mesure de situer son point de focalisation au sein de la carte globale.

FIGURE 7.16 – Exemple de projection cartographique avec Google Earth

Enfin, les paradigmes d'interaction, que ce soit au niveau de l'espace informationnel structuré, représenté ou visualisé, sont pour la plupart exploités pour notre système d'immersion. Agir sur l'un des niveaux revient à agir en cascade sur les niveaux supérieurs.

Les interactions liées à l'espace informationnel structuré concernent notamment l'accès aux données elles-mêmes, et passent par des requêtes dynamiques et des activations de liens au sein d'une carte donnée. Celles qui sont appliquées à l'espace informationnel représenté permettent notamment de concentrer une vue sur des structures visuelles données. Enfin, celles qui concernent l'espace visualisé ont trait au mouvement de caméra et au contrôle de point de vue que nous avons évoqués plus haut, dans les paradigmes de visualisation uniforme.

Notre système d'immersion implique d'avoir à disposition un grand nombre de possibilités de cartographie, puisqu'il doit restituer des informations de manière adaptée en fonction d'un grand nombre de paramètres liés aux dimensions projetées et aux arguments de la requête.

Cependant, nous avons pu circonscrire les types de cartes privilégiés. En particulier, les représentations orientées relations et les arborescences sont celles qui se prêtent le mieux à la structuration de nos ensembles documentaires. A celles-ci s'ajoutent également des représentations géographiques, dans les cas où les localisations des organisations liées aux documents sont impliquées dans une requête. De plus, afin de permettre une exploitation maximale des documents, l'interaction entre utilisateur et système autorise la formulation de requêtes dynamiques, l'exploitation de liens, et des allers-retours entre global et local, vision synthétique et focalisation sur des données particulières. Dans le cas particulier de représentations géographiques, utilisées

dans le cas où des données doivent être visualisées en fonction de critères de localisation spatiale, des fonds de cartes sont utilisés.

Compte-tenu de cette analyse, nous présentons dans les figures 7.17 page ci-contre et 7.18 page précédente, une maquette d'interface du système d'immersion.

Dans ces interfaces, les trois arguments de la requête dynamique sont présentés sur un bandeau vertical à droite de l'écran. La partie principale est consacrée à la carte, visualisant un réseau d'organisations sous la forme d'un graphe nœud-lien. La seconde carte est plus locale que la première, et a été obtenue par un zoom à partir de celle-ci, effectué par l'utilisateur. Chaque nœud correspond à une organisation. Si aucun niveau spécifique n'a été sélectionné pour les organisations, alors activer un nœud donné permet par défaut de déployer toute sa hiérarchie. Les arcs reliant les nœuds représentent des relations de collaboration : sélectionner un arc permet en l'occurrence de visualiser le nombre de documents associés aux deux organisations unies par la relation, ainsi que d'accéder aux document correspondants. En bas du bandeau se trouve la liste des documents associés au nœud ou à la relation en cours de sélection. L'utilisateur peut finalement accéder au contenu de chaque document en cliquant sur les éléments de cette liste, atteignant ainsi le niveau le plus local.

7.3.2.4 Conclusion

Pour la visualisation d'informations au sein de notre système d'immersion, des ressources cartographiques sont nécessaires. Nous avons vu qu'un grand nombre de cartographies existaient. Cependant, seul un nombre restreint d'entre elles sont adaptées à notre système d'immersion. Dans notre cas, bien que nos besoins soient variés, toutes les cartes doivent être interactives, et permettre un certain nombre d'actions de la part de l'utilisateur. Celui-ci doit pouvoir jouer sur plusieurs niveaux de profondeur, et se déplacer à l'envie dans les vues proposées. De plus, les retours en arrière et les requêtes dynamiques doivent permettre la reformulation. Par ailleurs, certains paradigmes cartographiques sont plus pertinents que d'autres étant donnée la structure en dimensions du système : les représentations orientées relations, en particulier, sont pour nous des moyens d'expression visuelle privilégiées.

Quoi qu'il en soit, l'ensemble de ces cartes ne sont réalisables que grâce à l'apport de données externes, du point de vue de la représentation visuelle de l'espace informationnel, comme du point de vue de la visualisation effective des données.

En effet, des dispositifs cartographiques externes à l'ensemble documentaire et à ses informations doivent être intégrés, avec le formalisme permettant de transformer et abstraire les données structurées de l'espace informationnel en structures visuelles adaptées. D'autre part, leur visualisation nécessite également des processus exogènes permettant la projection effective de l'espace représenté dans le système d'immersion, par des vues présentées à l'utilisateur. Enfin, les outils

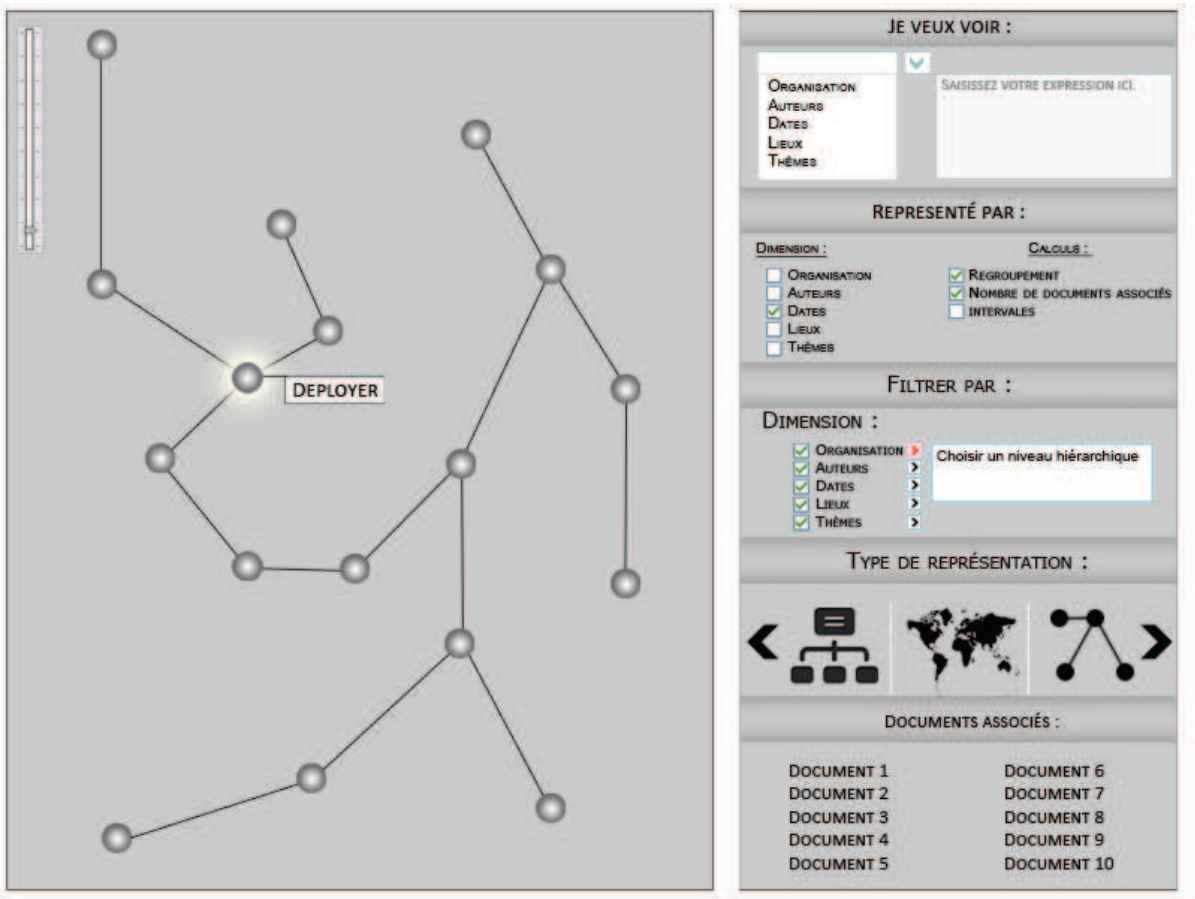

FIGURE 7.17 – Maquette d'interface pour le système d'immersion documentaire, avec une vue globale d'un graphe noeud-lien

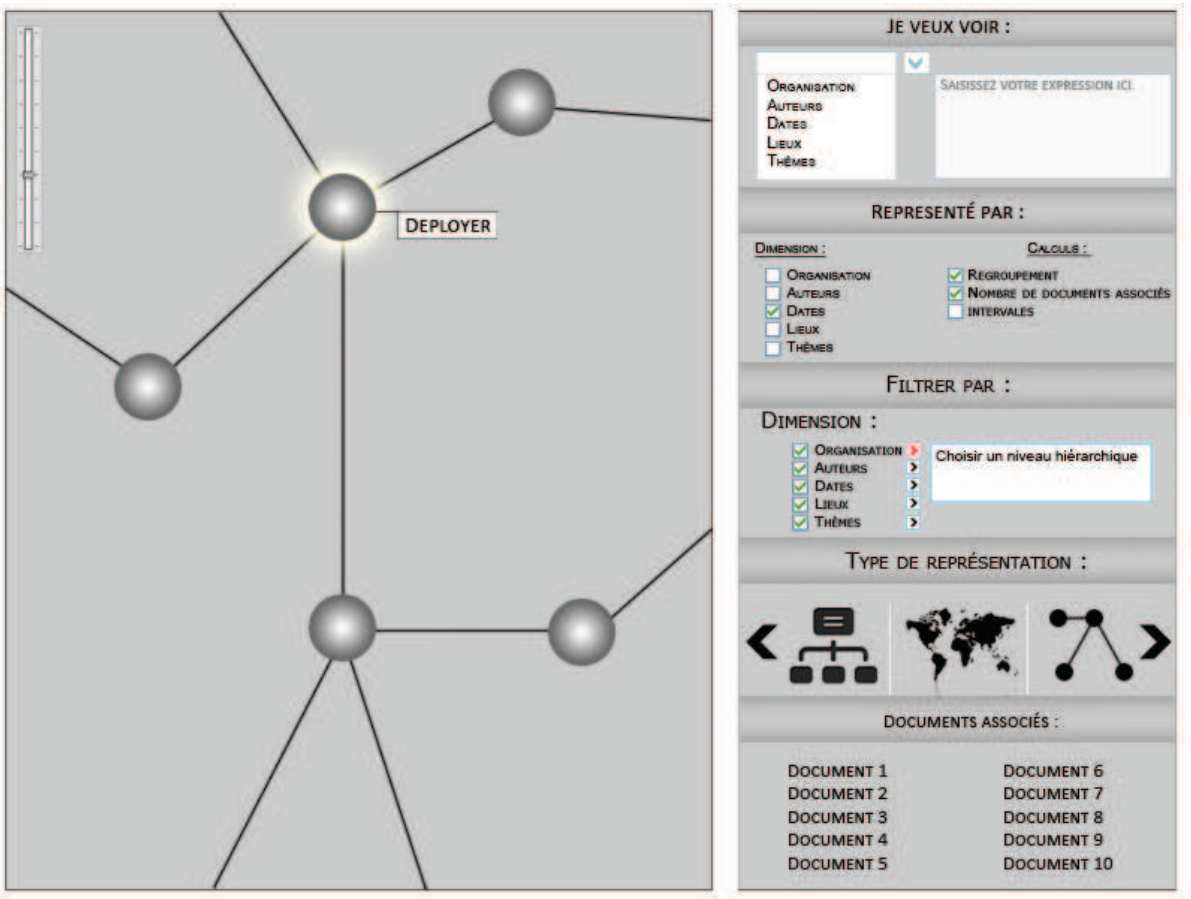

FIGURE 7.18 – Maquette d'interface pour le système d'immersion documentaire : vue locale d'un graphe noeud-lien par zoom avant

d'interaction rendus possibles par les paradigmes de visualisation choisis doivent pouvoir être utilisés.

7.4 Conclusion

Des méthodes exogènes ont été mises en place à tous les niveaux de granularité, et ont apporté, pour chacun d'entre eux, des informations permettant de mener à bien les traitements dédiés et de les optimiser. En ce sens, elles augmentent les données résultantes et les informations qu'elles véhiculent. Nous présentons dans la figure 7.19 page 315 les augmentations qu'elles permettent sur les différentes facettes, et la structuration qu'elles peuvent engendrer à partir des résultats endogènes.

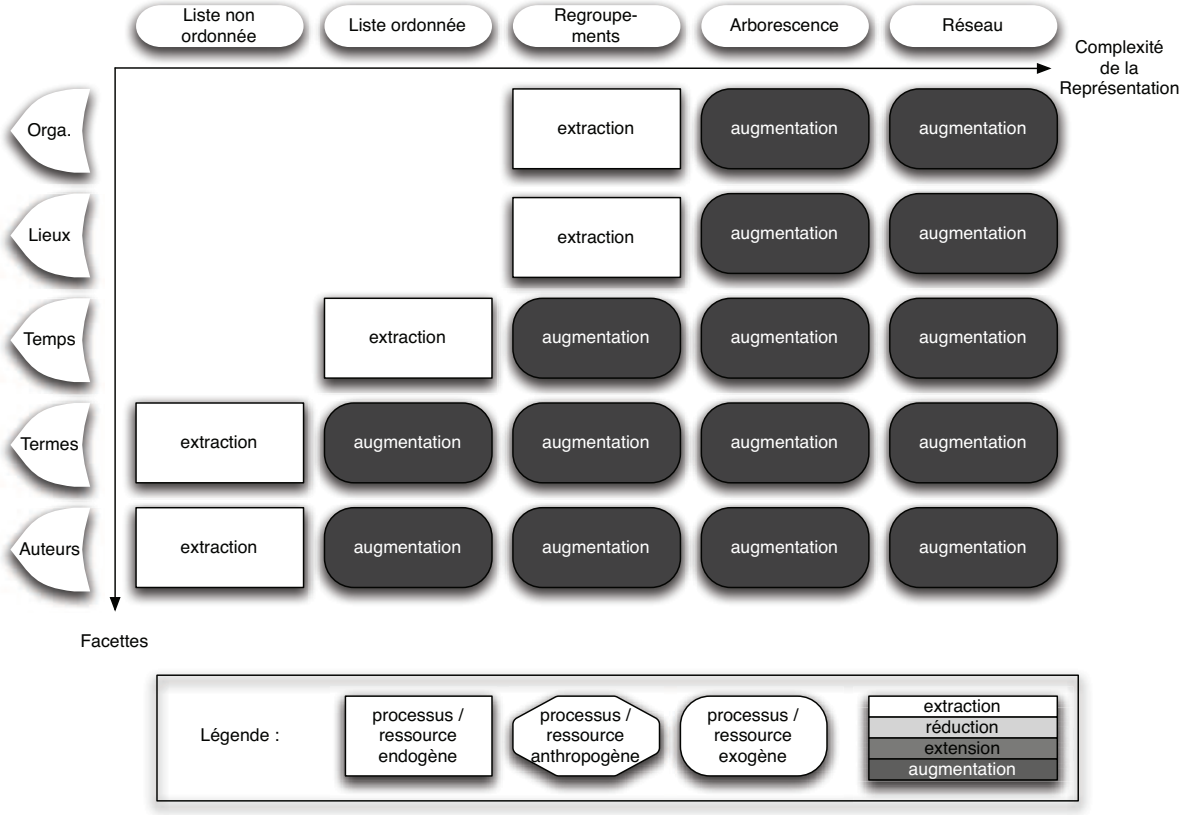

FIGURE 7.19 – Apport des ressources et processus exogènes pour chaque facette et structure résultante

Une ressource exogène permet d'augmenter chaque type d'information, et de les faire évoluer vers des structures de plus en plus riches, si tant est qu'elle-même contienne les informations nécessaires. Par exemple, une liste non ordonnée comme celle des auteurs peut devenir ordonnée, arborescente ou relationnelle.

La normalisation par système de règles, ressource dynamique fondée sur des patrons lexico-syntaxiques, permet d'aboutir à des résultats relativement bons sur les données objectivées, et d'obtenir des entités nommées propres, harmonisées entre elles et structurées hiérarchiquement. Ce système élimine en particulier certains types de variation, et complète ainsi les processus endogènes appliqués sur d'autres types de difficultés. Le système lui-même s'appuie sur des ressources exogènes statiques, plus précisément différents lexiques dont les items entrent en jeu dans les règles.

Dans le cadre de la constitution de la ressource termino-ontologique (RTO) multi-plans, des ressources statiques sous forme de lexiques viennent appuyer des traitements endogènes, et permettent d'améliorer leurs performances dans notre contexte. Les facettes des lieux et celle des thèmes, en particulier, bénéficient de données plus précises et complètes grâce à ces lexiques.

Enfin, de telles ressources sont exploitées dans le système d'immersion documentaire, à la fois du point de vue calculatoire de celui-ci et dans son aspect visuel : des ressources telles que des structures de représentations de données peuvent être injectées en parallèle des facettes endogènes et venir ainsi compléter et enrichir les dimensions de projection. D'autre part, des ressources cartographiques, statiques ou dynamiques, permettent la visualisation effective des informations.

Ainsi, si nous ne considérons pas comme souhaitable de réaliser l'ensemble des traitements à l'aide de ressources exogènes, celles-ci sont cependant indispensables pour certains aspects, et permettent de fortement améliorer les performances sur d'autres.

Notre système d'immersion est un outil devant permettre de passer de données brutes à des visualisations pertinentes des informations. Les traitements sont donc complexes, et impliquent un grand nombre d'étapes. Dans ce cadre, la combinaison de plusieurs types de méthodes et ressources, et en particulier de méthodes endogènes et exogènes entre autres, est à notre sens le meilleur moyen d'obtenir des résultats satisfaisants.

Chapitre 8

L'utilisateur comme agent interprétant : usage de processus et ressources anthropogènes pour chaque niveau d'unité

Sommaire

Nous avons énoncé, dans le chapitre 2 (sous-section 2.3.2 page 54), les principes du constructivisme auxquels nous rattachons en grande partie ces travaux. Les axiomes phénoménologique et téléologique, en particulier, mettent en avant le rôle du sujet humain dans la construction de la connaissance. Selon l'axiome phénoménologique, la connaissance est le résultat de l'expérience, qui n'est possible que par l'interaction du sujet humain avec l'objet observé. Le sujet est donc actif dans la construction de la connaissance, et celle-ci est un acte cognitif. Par ailleurs, d'après le principe téléologique, cet acte cognitif de connaissance est intentionnel : le sujet humain poursuit un but lorsqu'il construit une connaissance, et a donc une intention vis-à-vis de l'objet.

En conséquence, l'humain possède des connaissances du monde, riches et complexes, qui découlent de cette expérience. Or, l'expérience et la totalité des connaissances qui en découlent ne sont pas formalisables de manière réaliste par une machine. La source de connaissances la plus importante et la plus riche reste donc le sujet humain lui-même. Il peut utiliser ses connaissances pour en construire de nouvelles, notamment en vertu du principe selon lequel l'homme cherche à générer, à partir des modèles constitués grâce à ses connaissances, « un monde plus ou moins régulier et prévisible » [von Glasersfeld, 1994].

Chercher à modéliser entièrement ces connaissances humaines serait donc peu pertinent, voire utopique, à deux égards : l'exhaustivité n'est pas accessible, en particulier pour des ressources langagières, qui comportent une grande part d'implicite. D'autre part, puisque la construction d'une connaissance est intentionnelle, modéliser les connaissances humaines une fois pour toutes n'est pas envisageable : la pertinence d'une connaissance est mouvante en fonction du sujet qui l'utilise.

Cependant, à défaut de pouvoir intégrer toutes les connaissances humaines dans la machine, il est possible de placer l'humain lui-même dans le système. Il n'est certes pas question de faire reposer entièrement un outil sur ses utilisateurs. Néanmoins, nous pouvons tirer parti de leurs connaissances là où d'autres types de traitements, nommément les traitements endogènes et exogènes, atteignent leurs limites.

Pour cela, nous mettons en œuvre des processus anthropogènes.

Jusqu'ici, nous avons qualifié d'endogènes les processus qui utilisent l'information venant des données du corpus à traiter lui-même, et d'exogènes les processus exploitant l'information venant de sources externes. Nous appelons donc anthropogènes les processus utilisant l'information venant de l'humain. L'adjectif *anthropogène* n'est à notre connaissance pas usité dans le domaine des sciences du langage, ni dans celui de l'ingénierie des connaissances ; nous l'empruntons à la biologie.

De manière formelle, nous définissons l'adjectif *anthropogène* de la manière suivante :

Définition 6. *Anthropogène désigne ce qui influe sur l'objet, l'organisme, l'ensemble ou le système étudié, et qui tire son origine de l'humain. Anthropogène s'oppose à endogène et exogène.*

Dans le cadre de nos travaux, un processus anthropogène repose sur les compétences et l'intelligence de l'humain, et en particulier sur les compétences d'interprétation fondées sur ses connaissances. Il place l'humain au centre d'un système en tant qu'élément interprétant et concepteur de connaissance, qui possède un savoir qui ne peut être entièrement formalisé.

Ces compétences sont exploitables à tous les niveaux de granularité, soit en faisant participer l'utilisateur aux traitements de manière active, soit en utilisant les résultats de son intelligence pour modéliser certains éléments.

Dans ce chapitre, nous présentons d'abord les processus anthropogènes mis en place pour le traitement des entités nommées. Puis nous décrivons la manière dont l'utilisateur a influencé la conception et la structure de représentation des connaissances, c'est-à-dire la ressource termino-ontologique multi-plans. Enfin, nous abordons en détails le système d'immersion en tant que modèle éminemment anthropogène, ne pouvant fonctionner que par une immersion de l'utilisateur au sein des documents.

8.1 L'agent interprétant comme élément clé du traitement des entités nommées

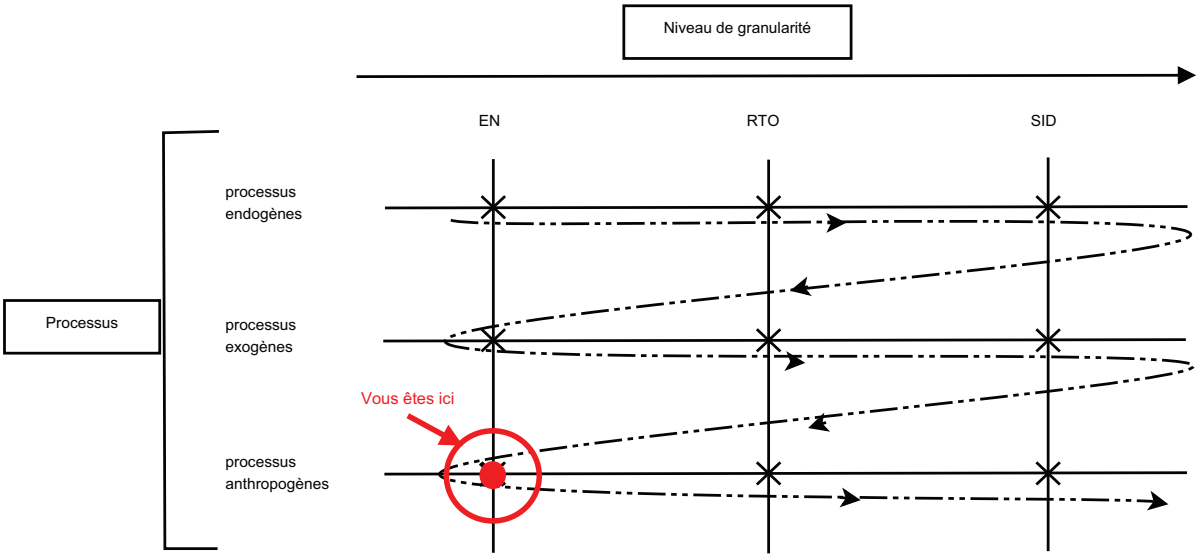

FIGURE 8.1 – Positionnement des traitements anthropogènes pour la normalisation dans l'ensemble des processus

La normalisation des entités nommées est fondée sur des processus endogènes et exogènes qui se complètent (voir les sections 6.1 page 159 et 7.1 page 241). Leur combinaison permet d'obtenir des données traitées de manière plus exhaustive et plus efficace, car les faiblesses des uns sont

partiellement compensées par les avantages des autres. Cependant, dans un certain nombre de cas, ces méthodes atteignent leurs limites sans qu'il soit possible de déterminer de manière sûre la normalisation « correcte » pour un nom. L'apport d'un utilisateur s'avère souvent pertinent pour résoudre de telles situations. Les processus guidés par cet apport de l'intelligence humaine sont des processus anthropogènes (figure 8.1 page ci-contre).

Réaliser l'ensemble des traitements de normalisation de manière anthropogène n'est pas réaliste : cela représente un travail beaucoup trop lourd pour l'analyste. De l'autre côté, se fonder uniquement sur la machine pour exécuter tous les traitements, sans intervention humaine, n'est pas suffisamment fiable dans un domaine d'activité où la précision et l'exactitude des données sont primordiales. Par conséquent, le moyen terme consiste à faire appel au travail humain uniquement sur ce qui n'a pas été jugé suffisamment fiable dans le travail effectué par la machine.

C'est en partie ici que se trouve l'intérêt des processus anthropogènes tels que nous les concevons : l'utilisateur intervient seulement dans certains cas, c'est-à-dire ceux qui sont problématiques pour la machine. Le deuxième avantage majeur de ces processus est que l'intervention de l'utilisateur permet la construction de connaissances qui sont adaptées spécifiquement à ses objectifs, et en principe, à ceux du groupe d'utilisateurs. En effet, d'après l'axiome téléologique du constructivisme, la connaissance est construite dans un but donné. Enfin, ils permettent de capitaliser, dans une seconde étape, ce qui est identifié par l'utilisateur comme « correct ».

Dans ce qui suit, nous présentons le processus de normalisation assistée des entités nommées, qui vient compléter les traitements endogènes et exogènes. Puis nous expliquons notre méthode de capitalisation des décisions prises lors de cette normalisation, qui prend place en deux temps.

8.1.1 La normalisation assistée des entités nommées

8.1.1.1 Hypothèse

Nous considérons l'utilisateur comme une source de connaissances, et comme une source d'interprétation des données qui lui sont présentées.

Ces connaissances sont difficilement formalisables dans des ressources exogènes, et il n'est pas réellement possible de les prendre en compte dans des traitements endogènes. Par exemple, nou avons vu dans la section 6.1 page 159 que deux amorces d'organisations comme *Inst* et *Center* n'étaient pas hiérarchisables par la machine, même à l'aide de ressources exogènes. Seul l'utilisateur, grâce à sa connaissance et son intelligence, peut décider de l'ordre hiérarchique entre deux noms particuliers comportant ces amorces. Par conséquent, faire appel à lui en tant que ressource pour les traitements est une manière de rendre ces derniers plus fiables, là où les autres types de processus atteignent leurs limites.

Pour ce qui concerne particulièrement la normalisation, utiliser des ressources anthropogènes permet de valider, ou d'invalider, des données issues de traitements exogènes et endogènes. Nous

faisons donc appel à elles pour améliorer les résultats de la normalisation par système de règles, mais aussi pour prendre des décisions à partir des suggestions fournies par les calculs endogènes de correction typo-orthographique et les calculs de découpage des noms à entités multiples.

Pour cela, nous nous fondons sur l'hypothèse suivante :

Hypothèse II.13. *Des processus et ressources anthropogènes, fondées sur l'utilisateur en tant qu'agent interprétant mais également comme concepteur de la connaissance, permettent d'optimiser les résultats de traitements endogènes et exogènes, tout en rendant les données ainsi traitées plus fiables.*

8.1.1.2 Application

Pour mettre en place les processus anthropogènes pour la normalisation, nous cherchons à placer les connaissances de l'utilisateur en entrée du système à un moment t, comme nous le ferions pour des ressources externes. Il s'agit donc, concrètement, de rendre possible l'interaction entre l'utilisateur et le système de normalisation.

La normalisation manuelle au début de nos travaux Avant le début de nos travaux, un moyen d'interaction humain-machine avait déjà été mis en place pour la normalisation. Son objet était à l'origine de permettre la normalisation, entièrement manuelle, des noms d'organisations et de lieux, et prenait la forme de fichiers de tableurs.

La normalisation réalisée grâce au système à base de règles permet de réduire la quantité de normalisations manuelles à effectuer. Nous rappelons qu'à l'issue du système à base de règles, 13,6% des noms d'organisations en moyenne, selon nos estimations objectivées, n'étaient pas normalisés correctement par rapport aux attendus. Si dans l'absolu, nous pouvons considérer ce taux d'erreurs comme faible, dans les faits, cela peut représenter un nombre conséquent de noms d'organisations, en particulier sur des ensembles documentaires très importants. Par exemple, sur une étude comportant 100 000 noms d'organisations, le nombre de noms à corriger s'élèverait donc à environ 13 600. De plus, nous avons expliqué que parmi les noms que nous avions considérés comme corrects puisque conformes aux standards attendus, certains d'entre eux pouvaient ne pas satisfaire les utilisateurs, en raison de cas particuliers, de structures hiérarchiques différentes de celles formalisées dans le modèle, ou de variantes sémantiques d'une même entité (voir sections 6.1 page 159 et 4.2 page 114 au sujet des variantes sémantiques).

Par conséquent, le travail de normalisation manuelle pour la correction peut parfois rester relativement lourd, en dépit de la normalisation effectuée automatiquement. Or, l'utilisation d'un tableur pour cette activité est loin d'être idéale : un tel fichier permet une interaction, mais pas d'une manière qui soit optimale pour l'analyste.

Il présente des inconvénients non négligeables, et en particulier celui de ne pas autoriser facilement l'ajout des fonctions conçues spécifiquement pour l'activité de normalisation. Par exemple, il ne permet pas la mise en place d'un calcul de distance de Levenshtein parmi les noms pré-normalisés.

D'autres problèmes se posent, du point de vue de l'utilisabilité : l'externalisation de cette normalisation manuelle est une source d'erreurs dans la manipulation des données, qui doivent être déplacées plusieurs fois et manuellement sur des supports différents. Ces manipulations sont autant de risques de ne pas placer ces données à l'endroit adéquat.

Par conséquent, cette manière de procéder à la normalisation manuelle peut être améliorée, à la fois dans sa forme, mais aussi dans son fonctionnement. Nous voyons dans la création d'un outil dédié et centralisé trois avantages :

- le traitement centralisé des données sans leur externalisation, donc un moyen de limiter les erreurs humaines lors de la récupération des fichiers sur le système global ;
- une plus grande souplesse dans les manipulations sur les noms d'organisations ;
- l'intégration de traitements permettant de passer d'une normalisation manuelle à une normalisation assistée.

Pour aboutir à la satisfaction de ces besoins, un outil doit être conçu, et permettre une interaction humain-machine plus simple. De manière applicative, cet outil prend la forme d'une interface web. Cette interface intervient en dernière étape du processus de normalisation. Elle doit permettre à l'utilisateur :

1. de corriger manuellement les résultats de la pré-normalisation automatique réalisée grâce au système à base de règles ;
2. de faire un choix parmi les suggestions de correction sur les paires de noms susceptibles d'être des variantes typo-orthographiques d'une même entité d'organisation ;
3. de sélectionner les découpages considérés comme corrects parmi les segmentations proposées sur les noms d'organisations à entités multiples.

Grâce à cette interface, l'utilisateur doit donc pouvoir transmettre ses connaissances au système, de façon à rendre les données plus fiables, mais aussi à enrichir le système lui-même.

Pour aboutir à ce résultat, et pour qu'il soit à la fois efficace et le plus satisfaisant possible pour les utilisateur, nous avons mis en œuvre une démarche ergonomique en trois facettes :

1. **Nous avons observé** les utilisateurs dans leur activité de normalisation manuelle ;
2. **Nous avons analysé** leur activité pour en tirer un modèle ;
3. **Nous avons conçu** l'outil de normalisation assistée à partir des analyses tirées de ces observations.

Cette démarche ne s'est pas réalisée en enchaînant les facettes : elle a consisté en plusieurs allers-retours entre les trois phases, qui ont permis d'intégrer l'évolution des usages au sein même

de la conception [Falzon, 2005]. Cette démarche est qualifiée de constructiviste (*ibid.*), puisqu'elle met en place un processus de construction des usages par l'interaction entre concepteurs et utilisateurs pendant la conception de l'outil.

Dans ce qui suit, nous présentons les trois cycles successifs qui ont permis de définir le modèle de l'interface. Chacun de ces cycles se décompose en trois phases distinctes, correspondant aux trois temps de la démarche ergonomique que nous avons décrite ci-dessus. Nous les présentons graphiquement dans la figure 8.2 page 324.

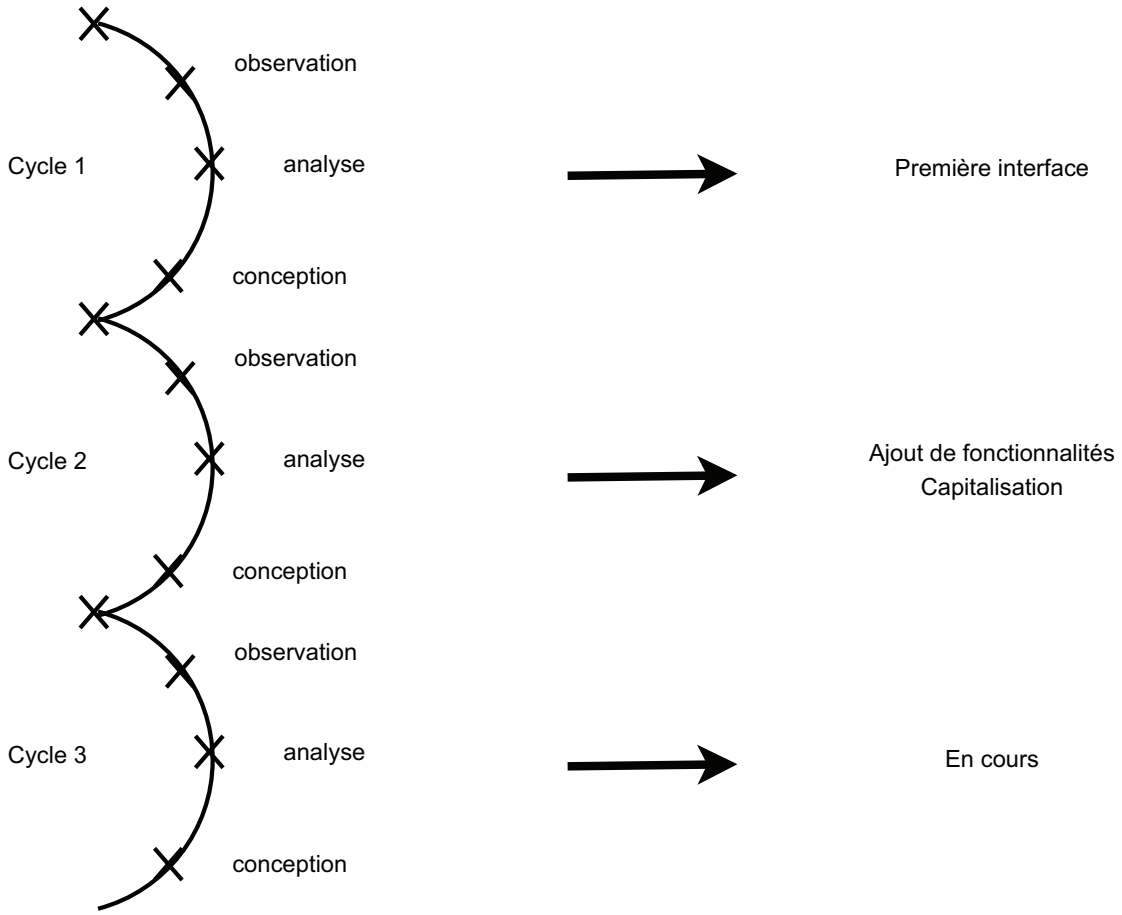

FIGURE 8.2 – Démarche cyclique de conception de l'interface de normalisation assistée

Premier cycle de la démarche

Première phase : l'observation de l'activité des utilisateurs Nous avons mené nos observations sur un échantillon d'utilisateurs, dont un quart étaient novices au début de ce travail, et les trois quarts restants des experts en normalisation. Tous les utilisateurs ont une utilisation intensive du système de normalisation. Le niveau d'étude de l'ensemble des sujets de

l'échantillon est sensiblement le même, et il s'agit souvent d'ingénieurs ; les domaines dans lesquels ils ont étudié sont divers, mais relèvent tous de sciences dures, comme la biologie, l'électrochimie, la physique des matériaux, etc. Leur niveau en informatique est également comparable. Globalement, à l'exception du critère d'expérience, les profils des sujets de l'échantillon sont donc homogènes.

Nos observations ont consisté dans un premier temps à observer physiquement les utilisateurs pendant leur activité réelle de normalisation à partir de fichiers de tableurs, et ce sur plusieurs études.

Les faits dégagés sont assez homogènes dans l'échantillon. Seuls les novices ont une activité différant en partie de celle des experts.

De manière générale, les analystes utilisent, de manière massive, les fonctions de tri et de filtres offertes par le logiciel de tableur. En effet, celles-ci leur permettent de ne sélectionner qu'un certain nombre de noms d'organisations, correspondant à certains critères, afin de les vérifier et, le cas échéant, de leur attribuer une nouvelle normalisation. De plus, ils peuvent visualiser côte à côte l'ensemble des occurrences d'une organisation donnée, grâce à des tris par exemple. Ils peuvent ainsi harmoniser l'ensemble de ces occurrences, en limitant le risque de normaliser deux entités semblables de deux manières différentes.

L'utilisation de ces fonctions n'est cependant pas la même en fonction du niveau d'expertise. Les experts font appel à des tris et filtres relativement complexes, incluant des chaînes ou sous-chaînes de caractères à rechercher dans les noms d'organisations bruts ou dans les noms pré-normalisés par le système à base de règles. Les novices, en revanche, se limitent souvent à des tris alphabétiques sur le début des chaînes de caractères, qu'ils font peu varier.

Des fonctionnalités comme les copier-coller, la possibilité d'étirer une sélection grâce à un clic maintenu, ou de coller sur plusieurs cellules le contenu d'une seule par un raccourci clavier, sont également très utilisées, en particulier chez les utilisateurs experts. Toutes ces manipulations concourent à accélérer la procédure de normalisation manuelle, afin de gagner du temps, sur des normalisations contenant de plus en plus de noms au fil des années d'activité de TKM [69].

D'autre part, dans le cas des utilisateurs novices, les allers-retours sont fréquents entre leur fichier de normalisation et une liste, construite par les analystes experts, contenant l'ensemble des universités françaises contenant un nom complet et standardisé. En effet, un grand nombre d'universités françaises ont plusieurs variantes pour leur nom. Par exemple, l'université de Paris 6 est aussi appelée *Université Pierre et Marie Curie*. L'intégration de telles listes pour les universités françaises est envisageable dans le système à base de règles. Cependant, à l'heure actuelle,

69. La taille moyenne d'une étude en nombre de documents n'évolue que légèrement sur la masse totale des missions, passant d'environ 5 000 il y a trois ans à environ 7 000 sur le premier semestre 2011. Cependant, la taille des corpus est extrêmement variable en fonction du type de mission. En outre, de plus en plus d'études comportent entre 10 000 et 20 000 documents, et des pics ont récemment été enregistrés à 100 000 et 200 000 documents.

l'harmonisation de ces noms doit être réalisée à la main. Face à de tels cas, les utilisateurs, et en particulier les novices, ont besoin de vérifier fréquemment les formes standards qui ont été définies.

Suite à ces observations, nous avons mené des entretiens, individuels ou collectifs, avec les utilisateurs, de façon à dégager leurs besoins et attentes vis-à-vis de la normalisation manuelle.

Si leurs attentes face à une nouvelle interface de normalisation peuvent parfois varier sur des détails, une tendance lourde se dégage de tous les entretiens : aucun d'entre eux ne serait prêt à utiliser une autre interface de normalisation si celle-ci n'offrait pas les mêmes fonctions de tri, de filtre et autres manipulations offertes par les logiciels de tableur, et ce même si d'autres fonctions étaient à leur disposition. Cependant, ils disent être prêts à un apprentissage, si celui-ci reste limité, concerne les nouvelles manipulations pour accéder aux mêmes fonctions, et facilite leur travail de normalisation à terme.

Deuxième phase : l'analyse de l'activité Ces éléments tirés des entretiens et de nos observations représentent autant de contraintes dans la conception d'une interface de normalisation assistée visant entre autres à remplacer les fichiers de tableurs.

Les utilisateurs, et en particulier les experts, sont habitués à ce format : ils y ont pris des repères, et en ont intégré le fonctionnement pour cette tâche de normalisation. Il convient donc, dans une certaine mesure, de respecter les grands principes de fonctionnement associés au tableur, afin de rendre l'outil résultant plus acceptable.

Les novices sont moins soumis à ce conditionnement. Toutefois, un novice devient rapidement expert : la normalisation est une tâche quotidienne de l'analyste. De fait, nous estimons que l'expertise est acquise en l'espace de trois mois, puisqu'au-delà de cette période, les utilisateurs ont acquis les mêmes habitudes que les plus expérimentés.

Ainsi, quel que soit le moyen mis en œuvre, la nouvelle interface devra nécessairement comporter les fonctionnalités suivantes :

- la possibilité d'effectuer des tris sur des chaînes de caractères déterminées par l'utilisateur ;
- la possibilité de mettre en place des filtres, là encore sur des chaînes de caractères saisies par l'utilisateur ;
- la possibilité d'étendre une normalisation à d'autres entrées de noms d'organisations dans l'interface ;
- de manière générale, la possibilité d'exécuter facilement dans l'interface toutes les fonctions réalisées dans un tableur par les nombreux raccourcis.

La présence de ces fonctionnalités, et leur facilité d'accès, sont des conditions *sine qua non* de l'utilité et de l'utilisabilité du système : s'il ne fournit pas du tout ces dernières, le système perdra son utilité, puisque des fonctions nécessaires à la réalisation de la tâche de normalisation ne seront pas disponibles ; s'il les fournit mais que leur mode d'exécution est trop complexe,

par exemple si leur réalisation nécessitait plus de manipulations qu'il n'est nécessaire dans un tableur, il en deviendrait beaucoup moins utilisable.

Troisième phase : la conception Si les fonctionnalités des tableurs que nous avons énumérées doivent être présentes dans le nouvel outil, leur réalisation n'a pas à suivre strictement celle de ces logiciels. Cependant, elles doivent être présentes, même *via* d'autres types de manipulations que des raccourcis claviers ou des glissements de fenêtre. De plus, leur exécution doit être aisée et rapide. Les choix de conception sont restreints par ces contraintes. Néanmoins, le choix des modes de manipulation n'est pas fermé, et nous disposons d'une certaine souplesse.

En effet, les utilisateurs se sont déclarés disposés à se former aux nouvelles manipulations, si tant est que cette formation soit courte et que le bénéfice soit tangible. Or, puisque notre public est habitué à l'informatique en général, et aux interfaces de type web en particulier, l'adaptation à des fonctionnalités simples est envisageable. Cette possibilité de souplesse est également ce qui motive la création de cette interface : offrir une plus grande souplesse d'utilisation, ainsi qu'une meilleure lisibilité, à l'activité de normalisation manuelle ou assistée.

Ainsi que nous l'avons mentionné plus haut, une nouvelle interface de normalisation doit également permettre l'application de fonctions de normalisation assistée, en particulier à travers les fonction de recherche de variantes typo-orthographiques, et d'aide à la segmentation des noms à entités multiples. Enfin, cette interface doit éviter les écueils dûs au traitement de cette normalisation à l'extérieur du système TKM.

Le type de l'interface a donc été choisi en fonction de ces critères : le plus pertinent est de faire appel à des interfaces de type web. Elles offrent de multiples possibilités, et donc une grande souplesse, à la fois dans la présentation de l'information et les fonctions qu'il est possible d'y associer. Enfin, elles peuvent être intégrées aux outils en place, et donc supprimer l'externalisation de la normalisation manuelle.

Puisque les utilisateurs ont l'habitude des tableurs, et qu'ils ont besoin d'un certain nombre de leurs fonctionnalités, il est important de concevoir une interface en accord avec la présentation des informations et les fonctionnalités de ces logiciels. Ainsi, l'interface doit reproduire la forme de table dans laquelle viennent se placer les noms d'organisations pré-normalisés.

Nous avons donc conçu une interface offrant ces caractéristiques. Nous en présentons un exemple dans la figure 8.3 page 328.

Cette interface permet de présenter l'information sur le modèle du tableur classique. A partir de cette représentation, il est possible d'enrichir l'interface, mais aussi de la moduler en fonction des besoins. Avec l'aide de l'équipe R&D de TKM, une première interface de test, contenant dans un premier temps l'ensemble des informations du fichier de normalisation, a été mise en place. Cette première version de l'interface permettait de simuler la normalisation manuelle, afin que les analystes puissent la tester. Dans ce cadre, et pour une première manipulation, seules les fonctions

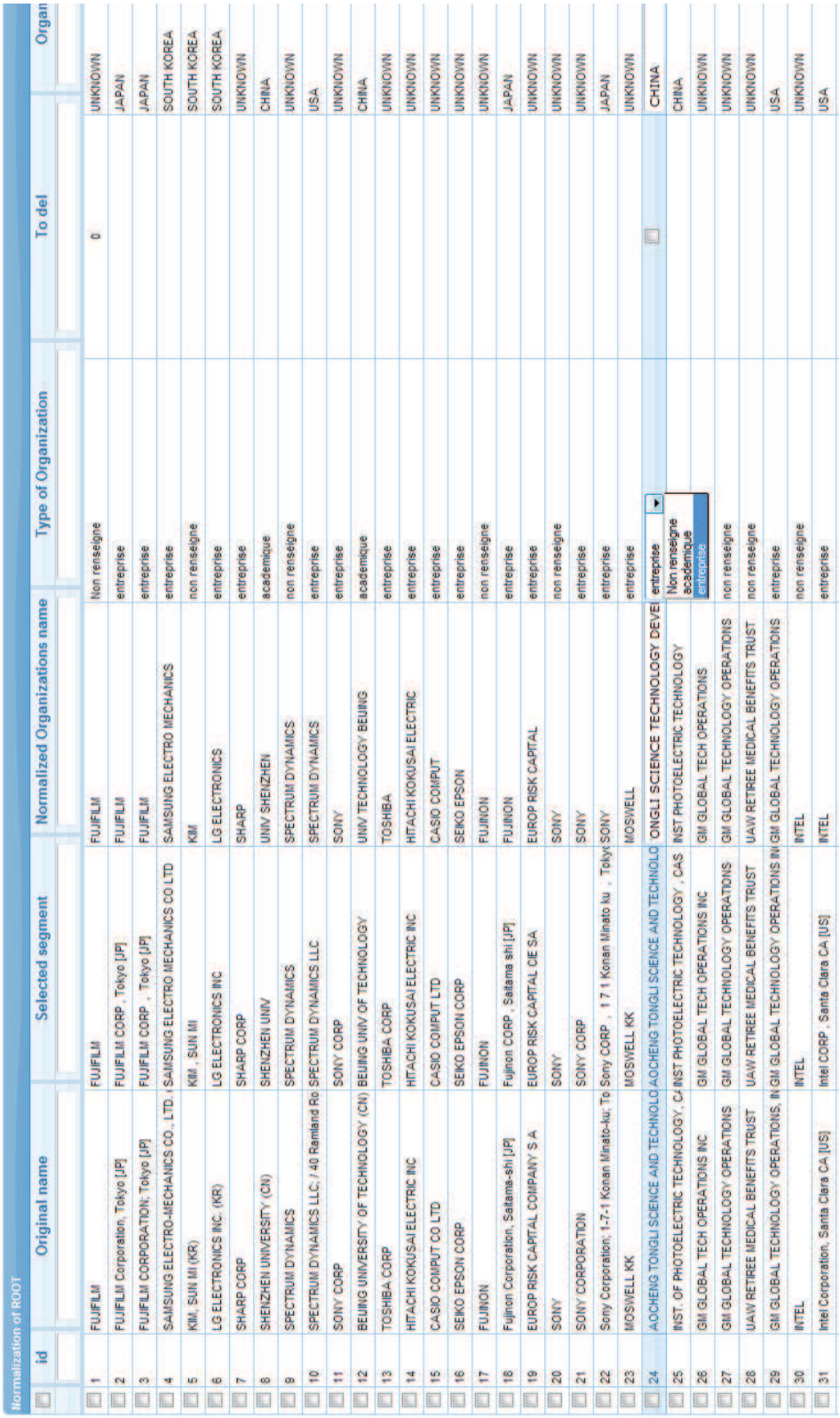

FIGURE 8.3 – Première version de l'interface de normalisation assistée

basiques du tableur ont été intégrées. Ces fonctions couvrent la possibilité de modification des noms d'organisations, ainsi que les fonctions de tri et de filtrage des données sur des chaînes de caractères déterminées par l'utilisateur.

Deuxième cycle de la démarche

Phase d'observation Cette première version de l'interface a été soumise à notre échantillon d'utilisateurs pour de premiers tests. Nos observations ont donc consisté, pour ce deuxième cycle, à collecter leurs retours.

Les résultats de ces tests rappellent les contraintes de fonctionnalités exprimées lors des entretiens avec les cinq analystes. Par exemple, les utilisateurs demandent à ce que les fonctions de type copier-coller soient facilitées, et à ce que la modification d'une ligne, et donc d'un nom d'organisation, puisse se faire d'un simple clic. Les paramètres par défaut de l'interface nécessitent en effet de double-cliquer sur une ligne pour l'éditer. Un double-clic au lieu d'un clic simple n'est pas problématique sur un ensemble restreint de données. En revanche, pour traiter des volumes très importants de données, cette manipulation devient lourde. L'objectif est donc, encore une fois, de gagner du temps lorsque les ensembles de noms à traiter sont très importants.

Par ailleurs, les remarques mettent l'accent sur la forme, et sur la visualisation des noms à normaliser. Les utilisateurs insistent sur l'intérêt de la mise en place de codes couleurs, à la fois pour distinguer les lignes entre elles, mais aussi pour mettre en valeur les entrées qui ont besoin d'être complétées, par exemple lorsque le pays n'a pas été identifié et/ou quand le type de l'organisation n'a pas été détecté.

Globalement, cette interface est cependant perçue unanimement comme « beaucoup plus agréable » que le fichier tableur de normalisation.

Phase d'analyse Les résultats de ces tests mettent en relief, de manière encore plus prégnante, la prépondérance du besoin de simplicité et de rapidité des manipulations. Cette simplicité et cette rapidité porte en particulier non sur la modification des noms en elle-même, mais sur l'accès et la gestion de cette modification. Tout doit donc être conçu pour faciliter le traitement humain de masses conséquentes de données.

Phase de conception Suite à ces tests, et sachant qu'un nombre important de fonctionnalités n'avait pas encore été implantées, une nouvelle version de cette interface a été développée.

Puisque la forme de l'interface a été jugée « agréable », et a été validée par l'ensemble des utilisateurs, le même modèle a été conservé.

Cette version intègre certaines des fonctionnalités présentes dans les tableurs, mais que nous n'avions pas encore implantées dans la première interface. De plus, elle prend en compte les

retours concernant la simplification des manipulations les plus utilisées.

Les améliorations de cette nouvelle interface peuvent être considérées en fonction des critères ergonomiques que nous avons présentés dans la section 2.2.3.2 page 44 de la première partie : l'utilisabilité, l'acceptabilité et l'utilité du système de normalisation assistée. Bien que les nouvelles caractéristiques ne relèvent pas exclusivement de l'un ou l'autre de ces critères, elles influencent de manière plus ou moins forte au moins l'un d'entre eux.

L'utilisabilité a été formalisée par [Nielsen, 1993] grâce à un ensemble d'éléments relativement objectivables ayant tous à voir avec la facilité d'utilisation d'un outil. En premier lieu, l'efficience renvoie au fait d'atteindre sans perdre de temps le but qui a été fixé. Dans notre interface, un ensemble de fonctions concourt à faire gagner du temps à l'utilisateur sur les manipulations et la normalisation des données. L'existence d'un menu principal et d'un menu contextuel optimise les actions de l'utilisateur, en fonction de leur cible : le menu général permet d'utiliser une fonction sur la totalité des noms à normaliser, tandis que le menu contextuel permet de se focaliser sur un nom en particulier.

Par ailleurs, qu'elles soient présentes dans le menu général (figure 8.4) ou dans le menu contextuel, certaines fonctionnalités participent au gain de temps sur les manipulations. Les fonctions *Edit and modify cells* et *Duplicate row* en particulier sont destinées à aider l'utilisateur à parvenir plus vite au résultat qu'il souhaite obtenir. La première permet d'agir sur plusieurs noms d'organisations à la fois, si tant est qu'au moins l'une des informations éditables doive être modifiée de la même manière. Elle consiste d'abord à sélectionner l'ensemble des noms qui doivent être normalisés de la même façon. La boîte de dialogue permet alors de saisir, pour la totalité des noms sélectionnés, la normalisation « correcte ». La seconde fonction, quant à elle, duplique une ou des lignes sélectionnées. De plus, certaines actions sont simplifiées, comme l'édition d'une ligne ou d'une cellule réalisée à l'aide d'un clic simple. Enfin, les lignes sont visualisées à l'aide de couleurs alternées, et les éléments manquants sont signalés par un code couleur spécifique, comme sur la figure 8.5 page 331.

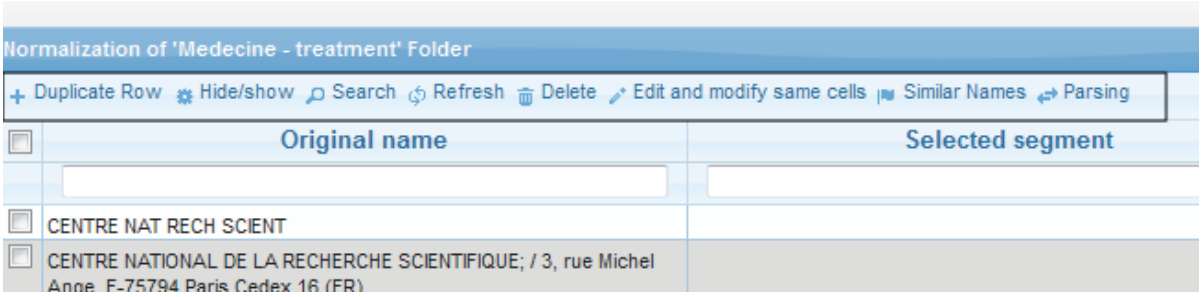

FIGURE 8.4 – Fonctions de normalisation assistée dans le menu général

Cette présentation des informations permet de réduire la charge cognitive des utilisateurs, qui

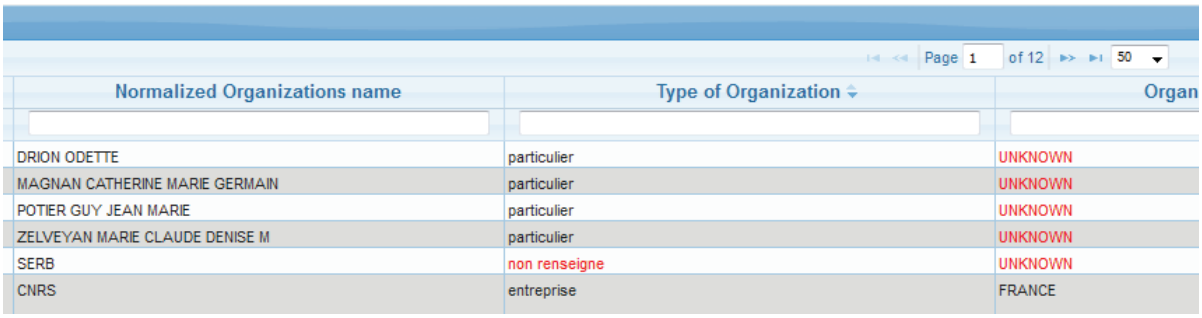

FIGURE 8.5 – Affichage par couleurs alternées et signalisation en rouge des informations manquantes (« Unknown »)

identifient immédiatement les lignes dans lesquelles il manque des informations. Cette réduction de la charge est également favorisée par la fonction *Hide/show* du menu général, qui permet d'afficher ou de masquer certains champs : ainsi, seuls les renseignements utiles à l'utilisateur à un moment t lui sont présentés.

L'apprenabilité concerne la facilité d'apprentissage du fonctionnement du système par l'utilisateur. Dans le cas de cette interface, puisqu'elle reprend un fonctionnement général déjà acquis par les analystes lors de la normalisation manuelle sur tableur, l'apprentissage est réalisé en un minimum de temps et sans difficulté.

Le critère de mémorisation est lui aussi satisfait, puisque l'utilisateur peut en permanence accéder à l'ensemble des noms d'organisations qu'il a normalisés, ce qui lui permet de retenir plus facilement ce qu'il a fait.

Le quatrième critère de [Nielsen, 1993] qui intervient ici concerne la fiabilité du système, et plus précisément la prévention et la gestion des erreurs par la machine. La fonction *Refresh* permet comme son nom l'indique de rafraîchir la page, c'est-à-dire d'annuler la ou les dernières modifications dans l'organisation des données qui ont été demandées par l'utilisateur. Ainsi, si celui-ci a fait un tri ou une sélection de lignes qui ne lui conviennent pas ou plus par exemple, ces dernières actions sont annulées et l'ordre initial des noms est rétabli. D'autre part, les erreurs de normalisation elles-mêmes sont également limitées, par exemple par le caractère non éditable de certains champs pour un nom d'organisation. Le nom brut par exemple ne peut pas être modifié, de façon à toujours garder un lien entre nom normalisé et nom d'origine.

Enfin, le critère de satisfaction est atteint lorsque les utilisateurs trouvent un système agréable à utiliser. Nous avons déjà vu, lors de la présentation de la deuxième phase, que c'était le sentiment général des utilisateurs ayant testé la première version de l'interface.

Les fonctionnalités et caractéristiques de l'interface ayant un impact fort sur sa facilité d'utilisation, et par là sur l'utilisabilité, sont donc les suivantes :

1. Efficience, dont réduction de la charge cognitive : *Edit and modify cells*, *Duplicate row*,

codes couleurs de la liste de noms, *Hide/show* ;

2. Apprenabilité : fonctionnement général déjà connu ;
3. Mémorisation : Accès aux organisations déjà normalisées ;
4. Fiabilité : *Refresh*, champs non éditables ;
5. Satisfaction : interface jugée agréable.

L'acceptabilité est également rempli puisque, ainsi que nous l'avons signalé, le mode de fonctionnement du tableur a été réutilisé pour l'interface. De cette manière, les utilisateurs acceptent le nouveau système de normalisation assistée, puisqu'il reprend le format, la présentation et les grandes fonctionnalités de ce qu'ils connaissent déjà. A ce titre, celle-ci s'intègre à leur pratique, à leurs outils et à leurs habitudes.

Par ailleurs, toutes les fonctions améliorant l'efficience influencent aussi l'acceptabilité, puisqu'elles « rentabilisent » le temps passé sur l'outil en restreignant le nombre de manipulations nécessaires.

Les caractéristiques influant sur l'acceptabilité de l'interface sont donc :

1. Fonctionnement intégré aux pratiques déjà en place ;
2. *Edit and modify cells* ;
3. *Duplicate row* ;
4. codes couleurs de la liste de noms ;
5. *Hide/show.*

Enfin, **l'utilité** est selon (citer Tricot, 2001) la possibilité d'atteindre un but visé à l'aide du système. A ce titre, notre interface est utile, puisqu'elle permet de normaliser les noms d'organisations et de lieux tout comme le permettaient les fichiers de tableur. Des possibilités telles que la fonction de recherche (*Search*), très utiles aux analystes, sont directement inspirées des fichiers de tableurs. De plus, les fonctions de normalisation assistée que nous avons ajoutées optimisent cette utilité, puisqu'elle permettent d'accéder à une information supplémentaire.

Les fonctionnalités « Similar names » et « Parsing » sont en cours d'implantation. Les suggestions de rapprochements de variantes typo-orthographiques, ainsi que les propositions de découpages de noms à entités multiples, sont présentées à l'utilisateur de manière optionnelle et adaptable. Leur présence dans le menu contextuel permet de les exécuter à partir d'un seul nom d'organisation, si l'utilisateur souhaite travailler spécifiquement sur ce nom. Elles sont également disponibles *via* deux boutons dédiés sur l'interface, et leur portée s'étend alors à l'ensemble des noms d'organisations présents.

L'intérêt de ces deux fonctions est qu'elles permettent de rassembler, en un seul traitement, des éléments de la liste des noms à normaliser présentant les mêmes types de problèmes, et qui

seraient autrement éparpillés : les tris et filtres ne permettent pas toujours de rapprocher visuellement tous les membres d'un tel ensemble. De cette manière, l'utilisateur n'a pas à parcourir toute la liste pour retrouver ces éléments, et surtout, la comparaison entre eux se fait automatiquement et non plus à la main.

Ces traitements sont mis en place de manière optionnelle, afin de ne pas contraindre l'utilisateur à patienter lors de l'exécution des calculs nécessaires à ces suggestions s'il ne les juge pas nécessaires au moment où il commence sa normalisation. Ce caractère optionnel agit donc à la fois sur l'utilité et sur l'utilisabilité du système, puisque l'utilisateur ne subit pas leur exécution et le temps de traitement qu'ils nécessitent, mais les décide. Ce choix est fonction du rapport bénéfice/coût évalué dans une situation donnée : l'utilisateur estime le ratio entre le bénéfice qu'il compte tirer de ces suggestions d'une part, et le coût qu'ils engendrent, notamment en temps, d'autre part.

Dans la version précédente de l'interface, la normalisation était simulée : le bouton de validation après correction n'était pas actif. Dans cette version en revanche, les décisions des utilisateurs sont capitalisées par défaut. Cette capitalisation permet de réinjecter, une fois l'interface en place, les connaissances ainsi acquises dans les futures normalisations. Cette capitalisation porte à la fois sur les corrections manuelles, et sur les corrections assistées de doublons et de découpage.

L'utilisateur a la possibilité de signaler au système, grâce à un champ spécifique, qu'il ne souhaite pas qu'une décision soit capitalisée. Dans certains cas très précis, étendre une décision propre à une étude à l'ensemble des données n'est pas pertinent. Mais de manière générale, la capitalisation des données étant un point fondamental des méthodes anthropogènes, la capitalisation sur un couple n'a pas lieu uniquement si cela est expressément demandé par l'utilisateur. Nous revenons sur cette capitalisation dans la sous-section suivante.

L'utilité est donc atteinte, à notre sens, par les caractéristiques et fonctions suivantes :

1. *Search* ;
2. *Similar names* ;
3. *Parsing* ;
4. Caractère optionnel de ces deux dernières fonctions ;
5. Capitalisation optionnelle des données.

Troisième cycle de la démarche La nouvelle version de l'interface est actuellement en fin de développement. Elle sera ensuite testée, à nouveau, par les utilisateurs.

Les résultats nous permettront de définir les derniers paramètres à respecter pour mettre en place l'interface définitive de normalisation assistée et éliminer totalement la normalisation manuelle par fichiers de tableur.

La mise en place finale de cette interface permettra la systématisation du processus anthropogène de normalisation, et fournira aux utilisateurs une assistance à leur travail. Cette aide sera apportée d'une part par le biais des manipulations permises, telles que les filtres et autres possibilités de classement ; d'autre part, par les fonctionnalités supplémentaires de repérage de couples proches et de suggestions de segmentation de noms à entités multiples.

Nous postulons que cette systématisation engendrera un gain de temps relativement important sur la normalisation. De plus, elle devrait également permettre un gain cognitif pour les utilisateurs : les nouvelles fonctionnalités, en particulier, font une partie de leur travail manuel de filtrage, et permettent de repérer des variantes difficilement détectables par d'autres moyens plus « classiques ». De cette manière, notre interface de normalisation assistée devrait réduire la charge cognitive des utilisateurs, qui d'une part se détourneront moins de cette tâche secondaire de normalisation, et d'autre part, dégageront plus de temps à consacrer à leur tâche principale d'analyse et de conseil.

Enfin, nous envisageons d'ores et déjà l'implantation dans l'interface d'un système permettant d'enrichir de manière anthropogène les lexiques utilisés pour les règles exogènes, de façon à optimiser ces dernières. Cette option ne sera cependant disponible que pour les utilisateurs les plus experts, afin de minimiser les risques d'erreurs.

Le fait de bénéficier d'une convention CIFRE pour nos travaux présente un intérêt majeur pour la conception de notre système d'immersion en général, et pour la mise en place de ces traitements anthropogènes en particulier. Le fait d'évoluer au sein même de l'entreprise présentant des besoins permet d'observer et de modéliser directement l'activité et l'utilisateur.

L'intervention humaine sur cette normalisation reste cependant relativement importante. Le contexte industriel de la tâche de normalisation oblige les utilisateurs à cette intervention régulière, qui reste le seul moyen de garantir que les informations capitales sont bien identifiées.

Les processus anthropogènes de normalisation ne sont donc pas voués à disparaître, mais au contraire à se pérenniser.

8.1.1.3 Conclusion

Les processus anthropogènes de normalisation sont fondés sur l'hypothèse selon laquelle leur utilisation optimise les résultats, puisque l'utilisateur est à la fois agent interprétant et concepteur de la connaissance. A ce titre, il interprète les résultats fournis par la machine, et peut à son tour fournir de l'information supplémentaire pour créer des connaissances fiables.

Puisque ces processus font appel à l'humain, il est nécessaire de fournir un moyen d'interaction entre celui-ci et la machine. Pour cela, nous avons conçu une interface de normalisation assistée. Elle permet de centraliser les traitements, offre une certaine souplesse dans les manipulations, et intègre les fonctionnalités de normalisation assistée que sont les suggestions de variantes et les

propositions de segmentation des noms à entités multiples.

Cette conception fait nécessairement intervenir des considérations ergonomiques, puisque l'interaction doit être adaptée à l'utilisateur et à ses besoins, de façon à ce qu'il communique le plus confortablement et efficacement possible avec le système. C'est pourquoi nous avons mis en place une démarche de conception cyclique, en trois temps pour chaque itération. L'activité des analystes est d'abord observée, puis analysée ; enfin, l'étape de conception tire de ces analyses des critères de modélisation pour l'interface. A la fin d'un cycle, un nouveau passage est itéré, de manière à prendre en compte les retours des utilisateurs jusqu'à atteindre un état satisfaisant de l'interface. Le caractère satisfaisant est atteint lorsque les critères d'utilisabilité, d'acceptabilité et d'utilité sont remplis le mieux possible.

La dernière itération de la démarche est en cours, et donnera lieu à de nouveaux tests utilisateurs qui permettront de finaliser l'interface et de l'implanter de manière pérenne dans notre système.

8.1.2 La capitalisation des décisions : un processus anthropogène en deux étapes

8.1.2.1 Hypothèse

Ainsi que nous l'avons rapidement évoqué plus haut, les décisions prises par l'utilisateur à l'égard de la normalisation sont capitalisées. Cette capitalisation est un principe fondamental des processus anthropogènes tels que nous les considérons. En effet, l'humain, considéré comme source de connaissances, peut en agissant sur un système le modifier, l'enrichir, le rendre plus fiable. Ses connaissances d'un domaine, et plus généralement ses connaissances du monde, sont difficilement formalisables *a priori* par la machine, et pas toujours repérables dans un corpus et ses données. De fait, l'humain lui-même est la source la plus fiable pour en retirer ces informations. Une fois ces informations exprimées d'une façon ou d'une autre et communiquées au système, il convient donc de les mémoriser et de les y intégrer, de façon à les réinjecter, au besoin, dans les traitements ultérieurs.

En effet, solliciter l'intervention humaine sans prendre en compte les résultats qui en découlent pour les études suivantes serait un non-sens : la perte d'information serait importante, et l'apport sur le long terme nul.

Concernant le point précis des méthodes anthropogènes qu'est la capitalisation des données, nous posons donc l'hypothèse suivante :

Hypothèse II.14. *La capitalisation des décisions des utilisateurs permet d'intégrer leurs connaissances dans le système. Elle rend les données d'un corpus déterminé plus fiables, et économise les traitements, qu'ils soient endogènes, exogènes ou anthropogènes, sur les corpus à venir.*

Les processus anthropogènes dédiés à la normalisation des entités nommées se divisent en deux niveaux de prise en compte, se matérialisant en deux étapes. Le premier niveau consiste à faire intervenir l'humain et à lui faire prendre des décisions concernant la validité ou la non validité des normalisations et suggestions du système. Le deuxième niveau concerne le choix de la capitalisation ou non de ces décisions. Dans les deux cas, c'est l'utilisateur qui est chargé de trancher. Dans la sous-section qui suit, nous présentons la capitalisation des décisions anthropogènes à travers ces deux étapes. Chacun de ces niveaux influence des éléments différents du système global.

8.1.2.2 Application

Premier niveau de prise en compte des processus anthropogènes Le premier niveau des processus anthropogènes, qui porte sur la validation, la correction ou l'invalidation des résultats du système, permet d'optimiser la normalisation automatique de l'ensemble documentaire en cours de traitement et d'analyse. Par conséquent, elle influe aussi sur la construction et l'enrichissement de la ressource termino-ontologique, toujours pour l'étude en cours.

Le processus de validation Nous venons de voir que les entités nommées, à l'issue de la normalisation exogène fondée sur un système à base de règles, n'étaient pas parfaitement normalisées : persistent des erreurs dues aux variations dans les structures hiérarchiques du monde réel, à l'existence de variantes typo-orthographiques ou sémantiques pour une même organisation, au fait que des filiales d'entreprises ne sont pas repérées, etc. L'intervention de l'utilisateur est donc nécessaire pour la finalisation de la normalisation, que celle-ci s'effectue de manière manuelle, ou assistée *via* les suggestions de correction.

Pour cela, les utilisateurs se servent de l'interface de normalisation que nous avons décrite dans la section précédente (8.1.1 page 321). Les analystes peuvent intervenir directement sur les noms pré-normalisés, puisque des zones éditables sont prévues à cet effet.

Ils peuvent également appliquer les calculs de détection de variantes et de segmentation, pour tout ou partie des noms présents. Si un utilisateur choisit de lancer le calcul de détection de variantes, un couple, ou une série de couples de variantes potentielles, lui sont présentés. Il doit donc décider pour chaque couple quelle variante correspond à l'entité normalisée, ou à saisir une nouvelle version normalisée pour les deux variantes. Il peut également signifier que le rapprochement effectué sur la base du calcul de distance d'édition est incorrect, et que les deux noms pré-normalisés réfèrent à des entités distinctes.

De même, s'il sélectionne les suggestions de segmentation, le ou les nom(s) concerné(s) lui est (sont) présenté(s) avec la ou les proposition(s) de découpage associée(s). Là encore, l'utilisateur sélectionne le découpage qu'il juge correct, ou saisit lui-même le découpage adéquat.

L'impact de cette validation Les résultats de l'ensemble de ces corrections, manuelles et assistées, sont enregistrés, et liés aux noms bruts dont les noms pré-normalisés avaient été extraits. Les noms pré-normalisés deviennent donc normalisés, et sont ceux qui sont restitués dans l'ensemble documentaire en cours d'analyse.

Au même titre que les processus endogènes et exogènes de normalisation, les processus anthropogènes ont donc également une influence sur la construction et l'enrichissement de la ressource termino-ontologique multi-plans, en ce qui concerne la facette des organisations et celle des lieux.

Les différents traitements effectués sur cet ensemble documentaire bénéficient donc de l'apport de la normalisation anthropogène, et les calculs statistiques qui en découlent sont de ce fait plus fiables.

Deuxième niveau de prise en compte des processus anthropogènes Le second niveau de prise en compte permet de capitaliser les décisions de validation, non seulement pour l'étude en cours, mais également pour les études ultérieures. En ce sens, la capitalisation permet de constituer, ou d'enrichir, de nouvelles ressources pour les traitements de normalisation portant sur de futurs ensembles documentaires.

Le processus de capitalisation Nous venons d'expliquer que la validation d'un nom d'organisation était prise en compte pour l'étude en cours, et intégrée aux données de l'ensemble documentaire correspondant.

En réalité, non seulement les noms peuvent être validés pour l'étude en cours, mais par défaut, ces validations sont prises en compte pour l'ensemble des études et ensembles documentaires associés à venir. Cette capitalisation globale est réalisée par défaut, puisque la capitalisation est un principe prépondérant des méthodes anthropogènes.

Cependant, l'utilisateur peut prendre la décision de ne pas capitaliser une validation ou une correction. Nous lui laissons donc le libre arbitre, et c'est à lui qu'il revient de déterminer l'intérêt de capitaliser ou non son action. Cette possibilité donne tout son sens à l'ensemble des traitements anthropogènes, puisque l'utilisateur garde le contrôle permanent sur les données et leur caractère valide.

S'il estime qu'une décision ne doit valoir que pour l'ensemble documentaire qu'il est en train de traiter, il a donc la possibilité de le signaler au système, par le biais d'une interaction à partir de l'interface de normalisation assistée. Dans ce cas, la décision est enregistrée pour un seul et unique nom d'organisation brut, et uniquement pour le document associé dans l'étude.

Dans le cas contraire, l'impact des corrections ne s'arrête pas à une seule variante de nom d'organisation, associé à une et une seule publication, et pour une étude donnée. Nous avons présenté, dans le chapitre 7 (section 7.1 page 241), la notion de segment associé à chaque nom d'organisation pré-normalisé. Ce segment est la portion d'un nom brut dont a été extrait le

nom normalisé, c'est-à-dire l'entité d'organisation globale, dans une version nettoyée mais pas à proprement parler normalisée, puisque n'ayant pas subi de réécriture ni de découpage interne. Par défaut, lorsqu'un nom normalisé est modifié par un utilisateur, le résultat de cette modification, et donc de la normalisation définitive et sûre, est gardé en mémoire et associé non seulement avec le nom d'organisation brut, mais également avec le segment qui lui correspond. Cette normalisation vaut donc pour l'ensemble des noms d'organisations partageant le même segment.

Dans le cas où l'utilisateur utilise les traitements de normalisation assistée, la décision de la capitalisation ou non d'une validation humaine revient là encore à l'utilisateur, sur le même mode : par défaut, toutes les décisions sont conservées pour être étendues à d'autres ensembles documentaires.

Pour les suggestions de correction de variantes, deux types d'informations sont retenues. Tout d'abord, un nom ayant subi une modification humaine, qu'elle soit faite entièrement manuellement ou de manière assistée, est capitalisé en tant que nom normalisé validé dans la base des entités nommées. De plus, chaque paire de variantes et la décision qui leur est associée est enregistrée. Dans le cas où l'utilisateur aurait signalé un rapprochement erroné, l'information est elle aussi mémorisée pour le couple, de façon à éviter que ces deux noms soient à nouveau présentés comme variantes lors des prochains calculs de repérage de doublons.

Le même fonctionnement sert à mémoriser les résultats des choix de découpages de l'utilisateur pour les noms à entités multiples : lorsque l'utilisateur choisit un découpage plutôt qu'un autre, ou qu'il saisit son propre découpage dans l'interface, ce choix est lui aussi conservé.

Pour les deux types de normalisation assistée, l'utilisateur peut également considérer qu'une décision ne vaut que pour l'étude en cours, et ne pas capitaliser sur cette décision.

Par exemple, dans le cas de la correction assistée de variantes, deux variantes telles que celles présentées dans le tableau 8.1 page précédente peuvent renvoyer à la même organisation, en l'occurrence la *Chinese Academy of Science*. L'utilisateur peut donc souhaiter identifier ces deux variantes comme semblables, et leur attribuer le nom normalisé *Chinese Acad Science*. Cependant, capitaliser cette décision et l'étendre à l'ensemble de la base documentaire n'est pas souhaitable : ce serait considérer que tous les noms pré-normalisés de la forme *Acad Science* renvoient à la même entité *Chinese Acad Science*. Or, d'autres entités peuvent prendre la même forme pré-normalisée, telles que la *Hungarian Acad Science*, la *Bulgarian Acad Science*, etc.

Nom pré-normalisé	Nom brut
Acad Science	Academy of Science, China
Acad Sciences	Academy of Sciences, China

Table 8.1 – Exemples de variantes non capitalisables après correction

L'impact de cette capitalisation Un nom pour lequel un utilisateur a capitalisé ses décisions acquiert dans les données globales un statut d'entité validée. En tant que telle, est est considérée comme fiable.

Cette capitalisation permet de réinjecter les connaissances ainsi acquises par normalisation assistée dans les études ultérieures. Nous avons expliqué, dans le chapitre 7 (section 7.1 page 241), que lors de l'import d'un nouvel ensemble documentaire pour une nouvelle étude, les noms d'organisations sont automatiquement normalisés par le système à base de règles. L'une des étapes de cette normalisation consiste à sélectionner puis à nettoyer partiellement le ou les segment(s) contenant une ou des entité(s) d'organisation(s), avant l'étape de réécriture.

Ce segment représente donc une version préliminaire de l'entité finale. A ce stade, nous cherchons à savoir, avant la phase de normalisation et de réécriture, si ce segment correspond à un segment similaire relié à une entité normalisée et validée par un utilisateur. Si tel est le cas, le segment nouvellement entré dans la base lors du dernier import est normalisé de la même manière que ce nom validé. Pour cela, les noms précédemment capitalisés et ayant acquis un statut de noms « valides », lors de corrections sur des ensembles documentaires précédents, sont exploités. Les corrections prises en compte sont à la fois celles qui ont été effectuées entièrement manuellement, ou qui ont été assistées par le repérage de variantes ou par suggestion de segmentation.

Cette fonction de capitalisation, fondamentale, permet de limiter les traitements manuels à effectuer par la suite par les analystes. De plus, en cherchant à faire correspondre les segments et non les noms d'organisations bruts, nous augmentons les chances de recoupements entre des unités déjà partiellement nettoyées, et donc moins soumises aux variations.

Dans un second temps, les décisions capitalisées de l'utilisateur concernant les couples de variantes typo-orthographiques sont elles aussi exploitées. Elles permettent en effet de constituer une liste de « fausses variantes », c'est-à-dire de couples proposés à l'utilisateur et que ce dernier a considéré comme référant à deux organisations distinctes. A chaque nouveau calcul de détection de doublons, l'une des étapes préliminaires au calcul lui-même consiste donc à examiner les couples de candidats entrants en les comparant aux paires de la liste. Si les deux membres candidats ont déjà fait l'objet d'une comparaison l'un avec l'autre, et que l'utilisateur les a signalés comme non liés par le passé, le couple est directement éliminé, sa distance d'édition n'est pas calculée, et il n'est pas proposé à l'utilisateur. Cela permet donc d'augmenter la précision du traitement, tout en gagnant en temps de calcul.

8.1.2.3 Validation

Les processus anthropogènes dédiés à la normalisation des entités nommées se divisent en deux niveaux de prise en compte, se matérialisant donc en deux étapes. Le premier niveau consiste à faire intervenir l'humain et à lui faire prendre des décisions concernant la validité ou

la non validité des normalisations et suggestions du système. Le deuxième niveau concerne le choix de la capitalisation ou non de ces décisions. Dans les deux cas, c'est l'utilisateur qui est chargé de trancher. Dans la sous-section qui suit, nous présentons la capitalisation des décisions anthropogènes à travers ces deux étapes. Chacun de ces niveaux influence des éléments différents du système global.

Nous venons de voir que la capitalisation des données s'effectue en deux étapes, toutes deux relevant de la décision et de l'action des utilisateurs du système.

La première étape, qui permet de valider une décision, permet de corriger les données normalisées et de les intégrer dans le système, pour le traitement d'un ensemble documentaire déterminé, c'est-à-dire celui qui fait l'objet d'une analyse pour la réalisation d'une étude en cours. Cette étape a donc une influence sur la totalité du système d'immersion, puisqu'elle rend les données plus fiables et les traitements qui en découlent plus performants. Cependant, à ce stade, la validation et ses conséquences ne valent que pour cet ensemble documentaire.

La deuxième étape permet de tirer parti des décisions d'un utilisateur pour un corpus donné, et de les étendre à la totalité des ensembles documentaires créés ultérieurement. Cette phase est donc celle qui permet de capitaliser, à proprement parler, les validations réalisées à l'étape précédente. Pour cela, l'utilisateur indique au système que les validations qu'il a effectuées doivent s'étendre aux données à venir. Ainsi, les noms d'entités acquièrent un statut particulier : ils sont considérés comme très fiables.

En tant que tels, ces noms forment un lexique qui est utilisé, au même titre que les autres ressources, endogènes ou exogènes, lors des traitements de normalisation automatique pour les études suivantes. La correspondance est établie, nous l'avons vu, à partir des segments des nouvelles entités d'organisations à normaliser et ceux des noms d'organisations capitalisées à l'étape anthropogène. Lorsqu'une identité est constatée par le système entre deux segments, le nom d'organisation entrant est normalisé d'après la forme du nom d'organisation validé par l'utilisateur lors d'une normalisation précédente, pour un autre ensemble documentaire.

Puisque ce lexique de noms validés est issu de traitements réalisés sur d'autres données que celles du corpus particulier à normaliser, il constitue une ressource exogène. En effet, ces noms validés ne sont pas tirés du corpus à traiter lui-même ; il ne s'agit donc pas d'une ressource endogène. De plus, l'utilisateur n'intervient plus sur le contenu de ce lexique, puisqu'il l'a déjà validé ; ce n'est donc plus une ressource anthropogène. La résultante d'un processus anthropogène dans une étude donnée devient donc ressource exogène dans les études suivantes.

Finalement, les processus anthropogènes de normalisation que nous venons de présenter, arrivant en aval des processus endogènes et exogènes, permettent de générer, en plus des entités normalisées elles-mêmes, une nouvelle ressource exogène. Le résultat du processus de capitalisation permet donc de compléter les ressources exogènes déjà en place que nous avons présentées dans la section 7.1 page 241. De plus, ce lexique est enrichi à chaque nouvelle normalisation as-

sistée, puisque les nouvelles normalisations validées sont intégrées au fur et à mesure des études, de manière automatique.

Dans l'état actuel de nos travaux, cette méthode de capitalisation enrichit les ressources exogènes statiques. La ressource exogène procédurale, c'est-à-dire le système à base de règles, n'est quant à elle pas concernée par ce processus. Cependant, étant donné le fonctionnement de la normalisation dans sa globalité, il est envisageable de mettre en place des processus d'apprentissage symbolique et/ou statistique permettant d'améliorer à son tour le système à base de règles. L'utilisateur pourrait, par exemple, dans l'interface de normalisation assistée, saisir de nouvelles règles, qui pourraient alors être intégrées au système pour de prochaines normalisations. Ce type d'apport anthropogène doit cependant être contrôlé, et être réservé aux utilisateurs les plus experts et les plus avertis, de façon à ne pas introduire des biais dus à des règles mal construites ou mal définies.

Quoi qu'il en soit, ce lexique exogène évite à l'utilisateur d'avoir à reprendre inlassablement les mêmes décisions sur les mêmes cas. Cela permet aux analystes de gagner du temps, de fiabiliser les résultats, et de limiter le caractère répétitif de la tâche.

L'usage de cette ressource est quantifiée grâce aux fichiers de log qui ont été mis en place il y a approximativement 6 mois, et qui recensent le nombre d'organisations normalisées grâce à lui lors d'un nouvel import de données. Les chiffres obtenus nous permettent de valider l'intérêt de la capitalisation pour la normalisation : en moyenne, sur les imports effectués lors des six derniers mois, 69% des noms d'organisations sont normalisés par ce biais, par l'appui sur les noms d'organisations bruts ou sur les segments.

Nous n'avons pu accéder aux chiffres antérieurs à cette période, puisque les fichiers de log n'existaient pas encore. Cependant, nous postulons que ce chiffre devrait aller en augmentant au fil des études et des imports associés.

Nous rappelons que dans ces données anthropogènes sont comptabilisées toutes les validations par défaut des noms normalisés : à moins que l'utilisateur ne signifie explicitement qu'il ne souhaite pas capitaliser un nom ou un ensemble de noms normalisés, tous, une fois la phase de normalisation assistée terminée, sont considérés comme validés.

Le fait que les correspondances entre noms se fassent sur le segment, et non pas sur la version brute d'un nom, permet des recoupements plus nombreux, avec une portée plus large sur l'ensemble des collections de documents.

En somme, plus le système de normalisation est utilisé, et plus le nombre de corrections manuelles ou assistées est important, plus les décisions capitalisées joueront un rôle dans l'amélioration du système. Les données qui en résultent sont considérées comme plus fiables que celles qui n'ont pas été validées d'une manière ou d'une autre par l'utilisateur.

Le système de normalisation dans sa globalité, incluant les trois types de processus - exogènes, endogènes et anthropogènes - est donc un système qui s'enrichit au fil des utilisations : plus il

est exploité, plus il est performant. Ce trait majeur permet de postuler qu'à partir d'un certain nombre d'utilisations, les cas de noms mal normalisés à l'issue du système à base de règles vont aller en diminuant, sans toutefois disparaître totalement. En effet, de plus en plus de cas seront traités en amont par le recoupement avec les décisions antérieures. Ainsi, les utilisateurs auront moins de données à traiter, même dans le cas d'ensembles documentaires importants.

8.1.3 Conclusion

La normalisation assistée des entités nommées, comme son nom l'indique, fait appel à l'utilisateur, à ses connaissances et à son intelligence. Elle implique également qu'il communique avec la machine par le biais d'une interface adaptée. La démarche que nous avons mise en place nous a permis de définir les besoins des utilisateurs relativement à cette interface, et qui concernent à la fois les fonctionnalités et la présentation des informations. Celles-ci ont en effet un impact important sur son utilisabilité.

Assister la normalisation, plutôt que de l'automatiser entièrement, permet de rendre les résultats plus fiables, et de les modeler en fonction des objectifs généraux des utilisateurs. D'autre part, une normalisation assistée plutôt que manuelle implique une plus grande rapidité d'exécution, et une optimisation du confort de travail pour l'utilisateur, qui lui permet d'être plus performant et de passer moins de temps sur des tâches répétitives et annexes.

8.2 Influence des méthodes anthropogènes sur la structure de représentation des connaissances

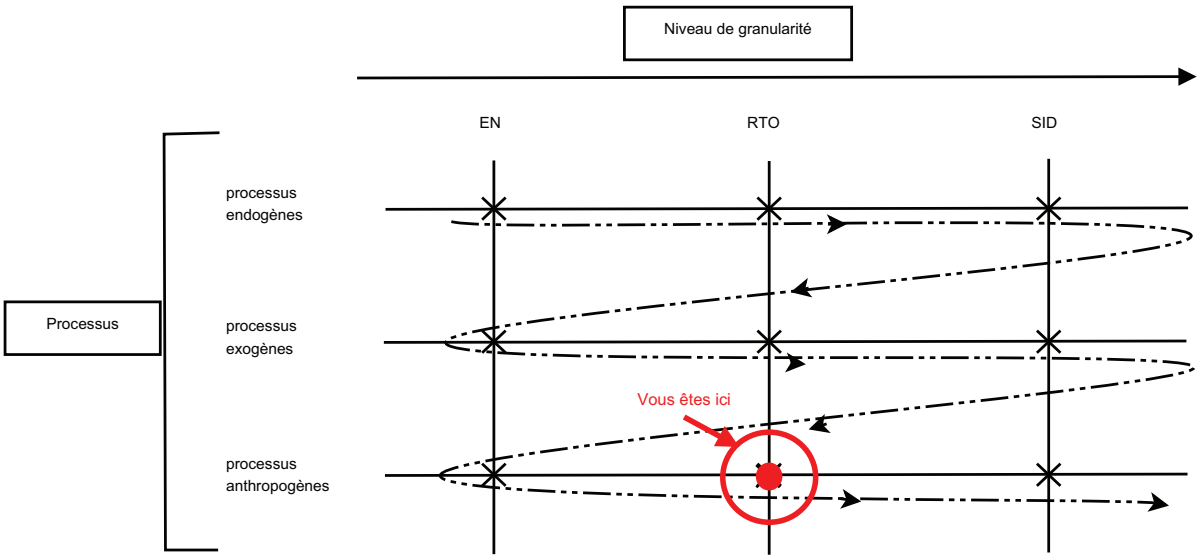

FIGURE 8.6 – Positionnement des traitements anthropogènes pour la ressource termino-ontologique multi-plans dans l'ensemble des processus

Nous avons abordé, dans les sections 6.2 page 199 et 7.2 page 284, les traitements endogènes et exogènes pour la constitution de la ressource termino-ontologique (RTO). Ces sections s'insèrent dans la structuration que nous avons choisie pour le présent document. Nous avons en effet segmenté ce dernier en fonction des différents types de processus appliqués aux niveaux de granularité que représentent les unités d'information en jeu, la RTO qui les englobe et les structure, et le système d'immersion documentaire qui s'appuie sur cette ressource. Or, ce découpage, dans le cas des ressources anthropogènes pour la RTO, ne correspond pas strictement au déroulement de l'application des processus. Cependant, celle-ci bénéficie bien d'apports et de ressources anthropogènes, ou pour le moins, « anthropomorphes » (figure 8.6 page précédente). Ils sont présents de manière relativement abstraite, par la prise en compte dans la phase de conception de la ressource de la représentation que se fait l'utilisateur d'un ensemble documentaire. D'autre part, un apport plus « concret » se fait de manière indirecte, par le biais des autres niveaux de granularité.

Nous présentons dans un premier temps l'aspect anthropomorphe de la structure de représentation qu'est la ressource termino-ontologique multi-plans. Puis nous décrivons les ressources anthropogènes qui sont employées à un niveau de granularité donné, mais qui ont une influence en cascade sur les niveaux plus globaux. Enfin, nous expliquons la façon dont la capitalisation et le report vers d'autres niveaux de ces processus anthropogènes permettent la limitation, voir

l'élimination, des actions concrètes de l'utilisateur dans le processus même de construction de la RTO.

8.2.1 « Anthropomorphisme » de la structure de représentation

L'utilisateur est à l'origine de la RTO. Celle-ci ne se place pas du côté des processus de construction de la ressource pour chaque nouvel ensemble documentaire, mais plutôt du côté de la conception de son modèle. Les utilisateurs n'interviennent pas directement, par des actions physiques, sur la conception, et leur influence n'est donc pas strictement anthropogène. Cependant, notre modèle de RTO se fonde en grande partie sur l'observation des pratiques des analystes. L'analyse de leur activité nous a permis de nous appuyer, pour la conception de la RTO, sur l'image - bien entendu partielle - que nous avons dégagée de la représentation mentale qu'ils pouvaient avoir d'un ensemble documentaire (ED). En l'occurrence, nous pouvons parler ici d'*anthropomorphisme*.

Le Petit Robert donne pour l'entrée *anthropomorphisme* la troisième définition suivante :

> Propriété d'un mécanisme dont la structure est à l'image du corps humain. *L'anthropomorphisme d'un robot.*

L'anthropomorphisme peut donc s'appliquer à un mécanisme, qui dans ce cas est conçu pour « ressembler » à l'humain. Selon nous, un mécanisme peut être informatique. Par ailleurs, le dictionnaire établit cette analogie entre mécanisme et humain du point de vue du corps de ce dernier. Or, nous estimons que l'imitation de l'humain par la machine peut également se situer sur le plan mental, et en particulier sur celui des représentations et raisonnements propres à l'homme.

Dans le cadre de nos travaux, nous définissons donc l'anthropomorphisme de la manière suivante :

Définition 7. *L'anthropomorphisme est une propriété d'un système informatique dont la structure reproduit partiellement les représentations et les raisonnements de l'esprit humain.*

Au vu de cette définition, la RTO est conçue par anthropomorphisme plutôt qu'elle n'est une ressource anthropogène : sa structure de représentation est dérivée de la structure de la représentation mentale des utilisateurs, ou en tout cas de ses caractéristiques les plus saillantes.

Pour accéder à cette représentation mentale, nous avons pris en compte la tâche principale des analystes de notre terrain d'application, et procédé à l'analyse de leur activité (voir le chapitre 2, section 2.2 page 28). Nous en avons dégagé des caractéristiques permettant de percevoir, partiellement, cette représentation, en fonction de la mission à réaliser. La prendre en compte est, à notre sens, une condition *sine qua non* de l'efficacité du système d'immersion.

Deux paramètres sont particulièrement saillants concernant ces représentations mentales : leur variété d'abord, qui correspond à la diversité des objectifs de missions confiées aux analystes.

Ensuite, le type des informations impliquées dans la grande majorité des missions forment une constante fiable. Les informations d'intérêt pour l'analyste renvoient en effet la plupart du temps :

- aux documents eux-mêmes ;
- aux organisations propriétaires des documents ;
- aux personnes qui en sont les auteurs ;
- au lieu où se situe l'organisation ;
- à la date de publication ;
- aux thèmes abordés dans les documents.

L'ensemble de ces informations ont donc été intégrées au modèle. Cependant, en raison de la disparité des objectifs, nous les avons scindées et distinguées au sein de facettes, toutes rattachées à la facette d'origine contenant les documents en tant qu'unités.

Ainsi, si les processus anthropogènes n'agissent pas directement sur la construction d'une RTO donnée, l'aspect anthropogène a présidé à la conception du modèle global, et donc antérieurement à une constitution particulière. La représentation des informations dans la ressource est donc à l'image de la représentation mentale des utilisateurs que nous avons dégagée, et de ce qu'ils considèrent comme pertinent. A ce titre, et de ce point de vue, nous qualifions la RTO multi-plans d'anthropomorphe, plutôt que d'anthropogène, en ce sens qu'elle est calquée sur une simplification de la représentation mentale que peuvent avoir les utilisateurs d'un ensemble documentaire.

L'intervention active de l'utilisateur lors de la construction même d'une RTO est donc limitée : les processus les impliquant sont évacués de ce niveau de granularité, et déplacés en amont ou en aval. En effet, des processus de ce type, ayant une influence sur la RTO, sont en particulier utilisés lors de la normalisation des entités nommées, au niveau de granularité le plus fin.

8.2.2 Les ressources anthropogènes multi-granularités

Nous avons exposé, dans la section précédente, la façon dont les processus anthropogènes interviennent dans la normalisation des entités nommées. Ils permettent de fiabiliser les unités d'information, mais aussi de contrôler et d'exercer une influence sur la structuration de ces unités, et en particulier sur les entités d'organisations. C'est en effet l'utilisateur qui, après les traitements du système à base de règles et en dernier ressort, détermine le niveau hiérarchique de chaque organisation détectée.

La structuration hiérarchique des entités nommées d'organisations est donc réalisée au moment de leur normalisation. Cette hiérarchie est ensuite exploitée dans la construction de la ressource termino-ontologique (RTO) multi-plans.

La facette des organisations tire donc profit de cette structuration, qu'elle restitue dans son agencement : chacune des organisations est associée au niveau qui lui a été attribué, mais

également à son entité mère la plus directe. Le résultat de cet agencement est donc un arbre : hormis pour les entités de niveau 1, toutes peuvent avoir une entité mère, et à l'exception des entités de plus bas niveau, toutes peuvent avoir une entité fille. Une entité mère peut avoir plusieurs filles, mais contrairement à une structure en réseau, comme celle des amorces qui permettent cette hiérarchisation par le biais du système à base de règles (voir la section 7.1 page 241), une entité fille ne peut avoir qu'une seule entité mère.

Les processus anthropogènes n'agissent donc pas directement sur la RTO une fois que la facette est constituée, pour venir en corriger les unités d'informations ou la structure, mais en amont de l'intégration des entités d'organisations.

De même, la sélection du niveau hiérarchique pertinent pour un ensemble documentaire déterminé, et donc pour une étude donnée, aurait pu dans l'absolu être réalisée au moment de la constitution de la RTO. La sélection de ce niveau permet à l'utilisateur de décider à quel niveau des organisations il souhaite travailler : par exemple, un état de l'art technique nécessitera de s'en tenir au niveau hiérarchique le plus haut, tandis qu'un niveau plus fin sera choisi pour une étude de positionnement d'un laboratoire donné par rapport à d'autres.

Ce point de vue hiérarchique adopté sur les organisations liées à une étude n'a toutefois pas été intégré au processus de constitution de la RTO, mais à celui de l'utilisation du système d'immersion.

8.2.3 Limiter les processus anthropogènes dans la construction de la RTO par la capitalisation et le report

Finalement, les traitements anthropogènes reliés à la ressource termino-ontologique ont été évacués du moment de sa constitution. Ce parti-pris permet d'éliminer de ce niveau de granularité deux écueils majeurs de ce type de méthodes.

Tout d'abord, elles peuvent s'avérer chronophages : employer trop souvent de tels processus fait perdre une partie du temps que l'utilisateur consacre à l'analyse des données, et en même temps peut décourager ce dernier de mener à bien la tâche qui permet d'intégrer ses connaissances au système. D'un point de vue théorique, il aurait été envisageable de placer l'utilisateur en position de « validateur » de la structure des facettes une fois constituées, comme nous l'avons fait pour la normalisation des entités nommées. Cependant, d'un point de vue pratique, cela aurait représenté une redite, au moins pour la facette des organisations, et n'aurait pas garanti une amélioration par rapport à la première validation effectuée lors de la normalisation anthropogène malgré le temps supplémentaire consacré. De fait, cela aurait réduit l'utilisabilité du système, ainsi que son acceptabilité.

D'autre part, cumuler des traitements anthropogènes au stade de la RTO pourrait être générateur de trop d'erreurs humaines. En particulier, la choix d'un niveau de manière définitive

à ce stade, impliquant que seul le niveau sélectionné serait intégré à la facette des organisations pour l'ensemble documentaire donné, serait un problème puisqu'il ne permettrait aucun retour en arrière en cas d'erreur ou même d'hésitation de la part de l'utilisateur. Repousser ce choix au moment de l'immersion le rend réversible à l'envi, limitant ainsi les risques d'erreurs qui ne seraient modifiables qu'en relançant le processus d'intégration des entités nommées à la facette, c'est-à-dire le processus de normalisation lui-même. Nous reportons donc à la granularité globale du système d'immersion les démarches anthropogènes de sélection, et réciproquement d'élimination, de certaines informations.

Cependant, par son action lors de l'étape de normalisation, l'utilisateur influence la structure de la ressource. Les données anthropogènes sont donc capitalisées à ce moment-là. La constitution d'une RTO donnée se fait donc par l'exploitation de la capitalisation réalisée à l'étape antérieure, sur le niveau de granularité le plus fin. Ces données issues de la normalisation sont réinjectées dans la RTO. Nous présentons ce processus dans la figure 8.7.

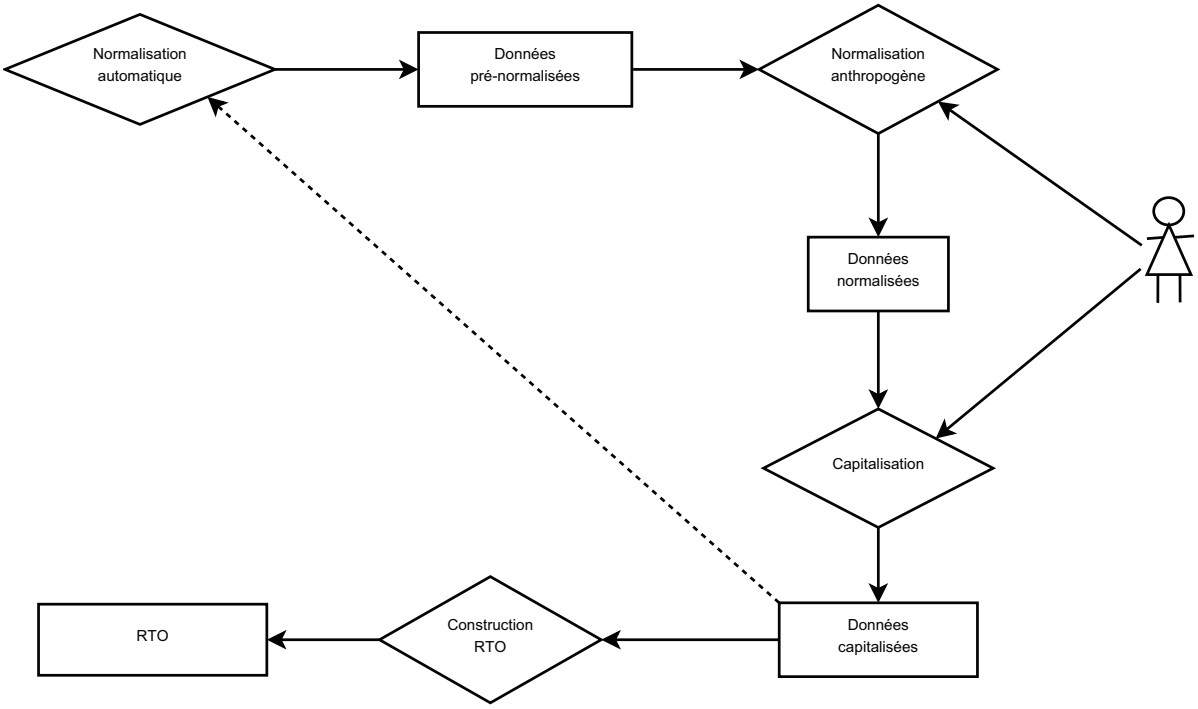

FIGURE 8.7 – Processus de capitalisation des données au niveau des entités nommées et de réinjection dans la RTO

Ici, les données pré-normalisées sont dans un premier temps normalisées de manière anthropogène ; leur capitalisation est elle aussi supervisée par l'utilisateur, qui décide de sa portée (son caractère modulable est représentée sur la figure 8.7 par la flèche en pointillés ; voir section 8.1.2 page 335). Ces données capitalisées, quelle que soit leur portée globale, sont réinjectées dans la RTO au moment de sa constitution.

L'utilisateur influence ensuite la résultante de la RTO par ses besoins informationnels, exprimés lors de l'utilisation du système d'immersion. Il modèle les inscriptions numériques que cette dernière contient pour n'en retenir que les informations pertinentes pour lui, c'est-à-dire celles qui lui permettront de construire des connaissances adéquates.

8.2.4 Conclusion

Reporter vers d'autres niveaux de granularité les traitements anthropogènes relatifs aux informations contenues dans les facettes de la RTO a donc pour effet d'améliorer les processus anthropogènes au niveau global. Ce report est une façon d'éviter l'accumulation d'interventions de l'utilisateur, qui doivent rester ponctuelles pour conserver leur efficacité. De plus, en fixant dès le départ le modèle de représentation d'après la représentation globale des utilisateurs, il n'est pas nécessaire d'intervenir par la suite pour apporter des modifications pour adapter la ressource aux besoins.

Dans l'absolu, il aurait été possible d'envisager un système universel, dans lequel l'utilisateur aurait lui-même défini les points d'intérêt pour tout type d'information, et tout type d'activité. Dans ce cas, la conception même de la RTO relèverait d'un processus anthropogène. Cependant, dans notre contexte, notre mission était de réaliser une ressource correspondant à un cadre donné, avec des utilisateurs déterminés et une activité globale récurrente. Ainsi, il a été possible de rassembler les types d'informations supposés pertinents, en raison de leur récurrence dans la représentation mentale des utilisateurs. Par conséquent, il n'est pas exclu, si la représentation mentale des utilisateurs se modifie de manière pérenne, que les nouveaux types d'informations soient intégrés à la RTO. La structuration en facettes permet en effet d'envisager de telles modifications dues à l'influence de l'utilisateur, si elles portent sur l'ajout de facettes supplémentaires.

8.3 L'utilisateur dans un système d'immersion documentaire anthropogène

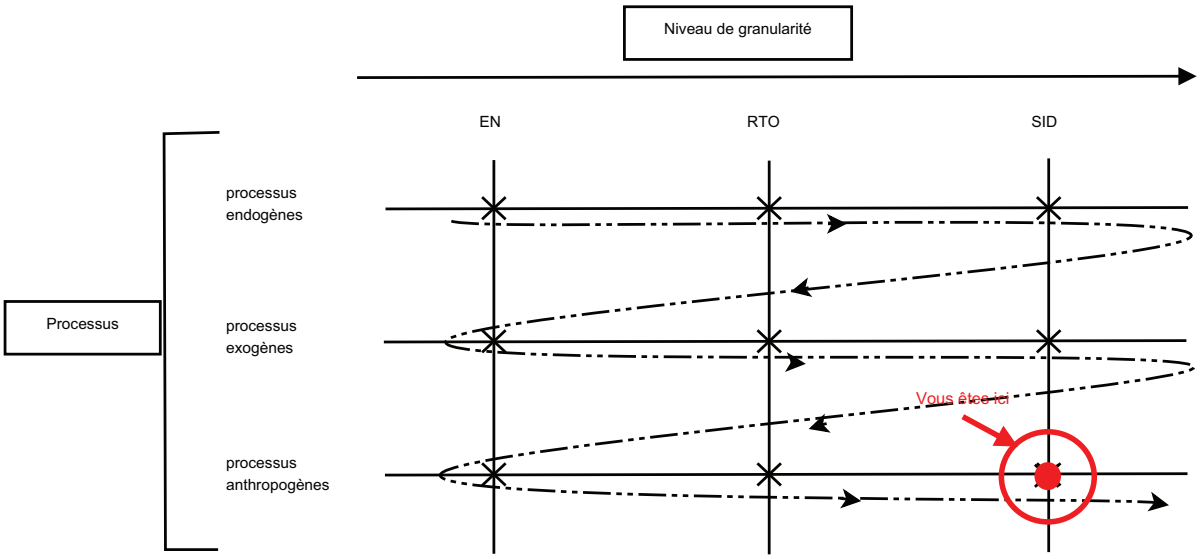

FIGURE 8.8 – Positionnement des traitements anthropogènes pour le système d'immersion documentaire dans l'ensemble des processus

Ce que je sais, c'est que je ne sais pas. Johnny Halliday [70]

Chaque niveau de granularité participe à l'élaboration du système d'immersion. Ainsi, les processus et ressources anthropogènes des niveaux plus fins (voir les sections 8.1 page 320 et 8.6 page 343) influencent en amont la structure du système global. Par ailleurs, du point de vue de son utilisation, celui-ci est exploité à l'aide de processus endogènes et exogènes (voir 6.3 page 222 et 7.2 page 284). Cependant, il ne peut fonctionner sans l'utilisateur, et n'aurait d'ailleurs pas lieu d'être sans lui ni son activité. Ceux-ci ont donc une place cruciale dans les processus d'immersion documentaire (voir figure 8.8).

Nous présentons à nouveau en figure 8.9 page 350 le point d'intervention de notre système d'immersion (cerclés sur la figure) dans le déroulement de l'activité de l'analyste pour la réalisation de la tâche générique de mission de conseil. Chaque phase correspond à une tâche secondaire prescrite par la tâche principale.

Le système d'immersion est utilisé pour la recherche d'information et l'analyse qui en découle. Par ricochet, il influence également l'étape de rédaction, puisque celle-ci s'effectue sur la base des analyses tirées de la recherche d'information.

70. Nous avons préféré cette citation à la maxime de Socrate : *Je ne sais qu'une chose c'est que je ne sais rien.*

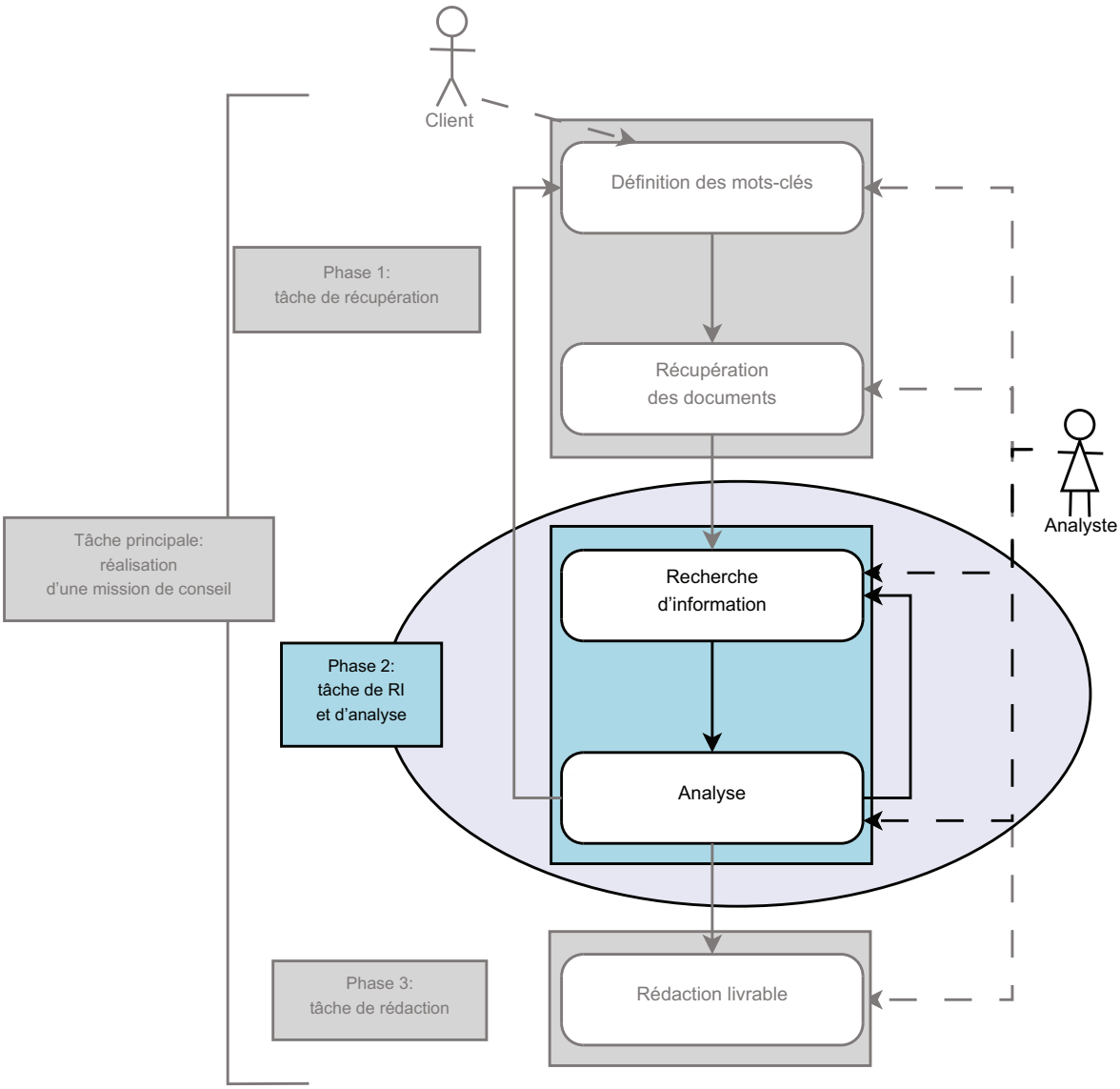

FIGURE 8.9 – Intervention de notre système d'immersion dans le déroulement par étapes de l'activité de l'analyste, pour une tâche de réalisation d'étude (figure 2.4 page 45 extraite du chapitre 2)

Puisque cet utilisateur est détenteur de connaissances, même « faibles » par rapport à une mission, et d'intelligence, nous ne le faisons pas seulement agir sur le système : nous l'immergeons dans les documents, de manière à intégrer son intelligence au système. De cette manière, il peut lui-même concevoir ses connaissances pertinentes à partir des possibilités de visualisation des inscriptions numériques de la ressource termino-ontologique.

Nous présentons, dans ce qui suit, la manière dont les processus anthropogènes interviennent dans le système d'immersion, et à partir desquels l'utilisateur peut construire sa propre connaissance pertinente grâce à la place qui lui est laissée. Tout d'abord, nous décrivons la manière dont l'utilisateur accède à l'immersion, en particulier par sa propre action, mais également, indirectement, par la modélisation qui a été faite de ses besoins et de ses buts de recherche d'information. Puis nous présentons les actions de l'utilisateur en tant qu'agent connaissant, d'abord lorsqu'il construit sa connaissance dans le système, puis dans la façon dont il interagit avec lui, et enfin, en tant que rédacteur d'une nouvelle source de connaissances.

L'ensemble de ces processus anthropogènes est représenté dans la figure 8.10 page précédente.

8.3.1 L'accès anthropogène à l'immersion : objectifs, matériau et critères de requête

L'accès anthropogène à l'immersion revient, pour l'utilisateur, à entrer littéralement en immersion : cet accès représente une porte vers le système et l'ensemble documentaire.

Le système est conçu de manière à répondre aux besoins des utilisateurs. Ces besoins se définissent en fonction d'objectifs liés à la tâche principale qu'ils doivent effectuer, qui peuvent être très variés.

8.3.1.1 Objectifs de recherche de l'utilisateur

Les objectifs des utilisateurs sont, nous l'avons expliqué dans le chapitre 2, extrêmement divers. Ils dépendent de plusieurs paramètres liés à leur activité, et inhérents à la tâche qu'ils doivent effectuer, ainsi qu'au contexte dans lequel il se place. Nous présentons ici certains éléments d'analyse que nous avons énoncés dans le chapitre 2 (section 2.2 page 28), et qui permettent de dégager les aspects de l'activité des analystes à prendre en compte dans le système d'immersion. Nous ne les reprenons pas *in extenso*, et nous concentrons sur les points utiles à notre développement.

Nous pouvons classer les caractéristiques de l'activité en fonction des réponses qu'elles apportent aux questions suivantes :

1. Pourquoi l'utilisateur a-t-il besoin d'accéder à l'information ?
2. Quel résultat cherche-t-il à obtenir ?
3. Dans quoi cherche-t-il ?

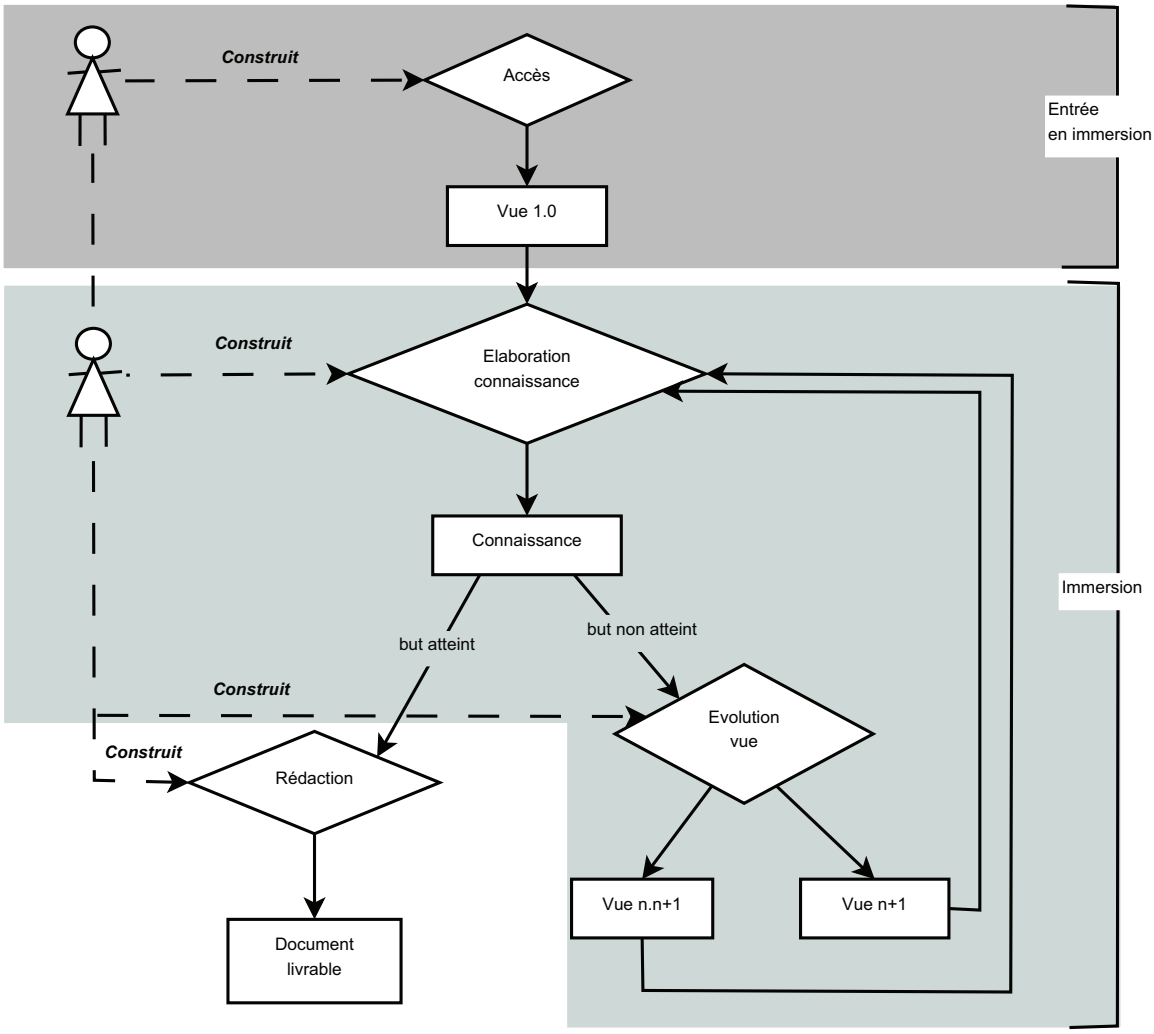

FIGURE 8.10 – Etapes anthropogènes d'immersion et de rédaction

4. Que cherche-t-il ?
5. Dans quel contexte a lieu sa recherche ?

1. Pourquoi l'utilisateur a-t-il besoin d'accéder à l'information ?

Parmi les six types de besoins d'information recensés par [Tricot, 2003], quatre apparaissent couramment dans l'activité des analystes :

1. Rechercher une connaissance qu'ils n'ont pas ;
2. Rechercher une confirmation d'une connaissance qu'ils ont ;
3. Rechercher une connaissance plus complète que celle qu'ils ont ;
4. Rechercher pour être conformes aux buts, aux contraintes, aux attentes de la situation.

Les trois premiers types sont de nature interne, propres à l'analyste, tandis que le dernier est externe, motivé par le cadre professionnel de leur activité. Nous avons établi que les besoins internes, dans ce cas, naissaient du besoin externe de devoir se conformer aux exigences de leur poste.

Bien que les besoins internes et externes d'information soient fortement liés, ils n'engendrent pas toujours les mêmes implications concrètes dans l'activité de recherche. Par conséquent, du point de vue du type de besoins, l'activité de recherche d'informations peut être très large.

Quel résultat cherche-t-il à obtenir ? Nous fondant encore une fois sur la classification des buts de recherche de [Tricot, 1993] (voir schéma 6.24 page 232 dans la section 6.3 page 222), nous avons constaté que l'ensemble des buts de recherche étaient poursuivis par les analystes, en fonction des études et/ou du stade d'avancement d'une étude donnée. Ils peuvent donc alternativement chercher un renseignement précis parmi les documents, les explorer, en collecter une partie, ou butiner parmi eux, en fonction du caractère flou ou précis de la représentation de leur besoin, et de la localisation de la cible contenant l'information.

Le but de recherche d'information évolue naturellement au cours de la recherche, puisque les résultats des premières recherches permettent de le réajuster au fil de la navigation.

Ainsi, là encore, la caractéristique dominante est la diversité des besoins, entre deux études, mais surtout au sein d'une même étude à des états d'avancements différents.

Dans quoi cherche-t-il ? Nous avons expliqué, dans le chapitre 2 (voir 2.2.3.1 page 38), qu'une étape de récupération des documents précédait celle de recherche d'information proprement dite. Cette étape de récupération se déroule sur des bases de données en ligne, à partir desquelles les documents répondant à des équations de recherche construites par les utilisateurs, semblables à des requêtes, étaient récoltés.

Ces documents forment l'ensemble documentaire dans lequel l'activité d'immersion a lieu. De fait, celui-ci est relativement homogène du point de vue thématique.

En revanche, les documents sont multi-sources, et représentent une masse importante de données. Nous ne prenons pas ici en compte la source secondaire que représentent les bases

de données en ligne, mais bien celle dont émane le document en premier lieu. Ces documents impliquent un traitement spécifique de l'information qu'ils véhiculent. En effet, ils font l'objet de la part des analystes d'une représentation (mentale) multi-documentaire, mettant en jeu des stratégies de lecture spécifiques. En particulier, cette représentation implique une indexation sélective par recoupements partiels entre contenus et sources. Cette représentation doit donc être facilitée par la visualisation des informations dans l'immersion.

Par ailleurs, les documents sont considérés alternativement par les analystes, nous l'avons exposé dans le chapitre 3 (sous-section 3.1.2 page 76), à travers leur fonction de medium et leur fonction de signe, selon le besoin d'information. Le document considéré comme medium s'inscrit dans une pratique sociale déterminée, et véhicule par conséquent des informations relatives à sa situation d'énonciation, et en particulier au versant de son contexte de production. D'un autre côté, sa fonction de signe a trait à son contenu propositionnel, et donc à l'énoncé qu'il porte. Ces deux caractéristiques, fondamentales pour l'activité, doivent être prises en compte et être visibles.

Que cherche-t-il ? L'utilisateur peut rechercher différents types d'informations dans l'ensemble documentaire. Cependant, ces types sont de manière générale récurrents au fil des études, reflétant une certaine régularité de l'activité sur cet aspect.

Tout d'abord, en tant que preuves d'une invention, ou d'un travail scientifique, les documents peuvent être recherchés pour eux-mêmes, puisque leur simple existence a valeur de preuve. Celle-ci renvoie à la fonction de medium du document que nous avons évoquée ci-dessus. Autour de cette fonction gravitent un ensemble d'informations véhiculées par les documents, et qui concernent leur situation de production : les organisations propriétaires, les auteurs ou inventeurs, les dates de publication et les lieux des organisation permettent de caractériser un document.

Par ailleurs, le contenu propositionnel des articles ou brevets est lui aussi d'intérêt pour les analystes, qui s'en servent pour dégager les thèmes traités dans un ensemble documentaire.

Les utilisateurs peuvent donc avoir besoin de rechercher l'ensemble de ces informations dans le système d'immersion : il convient de leur offrir cette possibilité, ainsi que celle de croiser entre eux ces types ainsi que les documents eux-mêmes, tout en en permettant des visualisations adéquates.

Dans quel contexte a lieu sa recherche ? Comme toute recherche d'information, l'activité de recherche des analystes s'inscrit dans un contexte déterminé. En l'occurrence, il s'agit d'un contexte professionnel, et industriel.

En conséquence de quoi, les contraintes temporelles sont fortes, d'autant plus que le milieu du conseil en innovation est extrêmement concurrentiel. Or, une recherche d'information réalisée avec la conscience d'un délai restreint est de manière générale moins efficace [Proctor, 2001].

Enfin, le degré d'expertise des utilisateurs, qu'elle soit procédurale ou conceptuelle, est la plupart du temps élevée : à l'exception des utilisateurs novices, c'est-à-dire des nouveaux analystes

entrant dans l'entreprise, les utilisateurs sont experts dans l'utilisation des outils actuels, et sont habitués à travailler sur des domaines scientifiques et techniques de pointe. Les utilisateurs novices, par l'utilisation intensive des outils actuels de recherche d'information, deviennent experts du point de vue procédural en l'espace de quelques semaines. Cependant, puisque le système d'immersion diffère des outils en place, un apprentissage sera nécessaire.

En somme, les besoins de l'utilisateur sont variés. Cependant, certaines constantes se dégagent, en particulier en ce qui concerne les documents exploités, et les types d'informations d'intérêt.

Les besoins et buts évolutifs des analystes, couplés au caractère multi-sources et multi-facettes des documents à traiter, impliquent une contrainte fondamentale : celui de séparer les données potentiellement significatives en autant de dimensions, permettant un accès distinct et personnalisé à différentes types d'informations en fonction des besoins.

En pratique, ces dimensions permettent de jouer sur trois types d'interactions avec l'utilisateur en tant qu'acteur de son immersion, en particulier pour la dimension des thèmes, ainsi que nous l'avons mentionné dans les sections 4.1.2.3 page 112 et 6.2.3.3 page 218. Les collocations brutes porteuses de thèmes permettent le filtrage perceptif, l'utilisateur perçoit principalement les termes qui ont du sens pour son étude. D'autre part, elles peuvent être utilisées en tant qu'axes de représentation. Dans ce cas, elles autorisent une navigation heuristique, grâce à l'exploitation de la logique du réseau par l'utilisateur : il peut naviguer de lien en lien à partir des relations thématiques entre documents, matérialisées par des collocations communes. Enfin, nous rappelons que c'est un objectif d'exhaustivité qui a présidé à la constitution de cette facette. En effet, nous avons pris le parti de ne pas préjuger de la pertinence des collocations extraites, puisque nous savons que celle-ci est mouvante en fonction des buts des utilisateurs. C'est pourquoi, lors de la navigation heuristique, un utilisateur peut être amené à découvrir des informations qu'il ne cherchait pas spécifiquement, en particulier si la représentation de son besoin de recherche était floue, et qui sont pourtant pertinentes pour lui. L'exhaustivité de la facette des thèmes est donc, couplée à l'intelligence de l'utilisateur, génératrice de sérendipité.

L'utilisateur est donc un acteur de son immersion dans les documents, puisqu'il filtre des données, il suit des idées à travers sa navigation, découvre des éléments proches de ceux qu'il connaît et qui lui apportent de l'information supplémentaire. Ces nouveaux éléments lui permettent alors de suivre un autre cheminement que celui qu'il avait envisagé *a priori*.

8.3.1.2 Matériau : des dimensions bien cernées

La contrainte principale ayant présidé à la conception du modèle de système d'immersion documentaire est la souplesse, puisque celui-ci doit répondre à une variété de besoins. Pour cela, nous avons défini cinq dimensions informationnelles distinctes. Elles sont formées par les cinq

facettes de la ressource termino-ontologique (RTO), et sont constituées à partir des informations tirées des documents de l'ensemble à traiter (voir la section 6.3 page 222).

Ces cinq dimensions renvoient respectivement aux organisations, aux auteurs, aux dates et aux lieux impliqués dans la situation de production des documents, mais également aux thèmes de ces documents, et donc à leur énoncé.

Le choix de ces dimensions n'est pas anodin : elles comportent les types d'informations que l'utilisateur maîtrise. Elles lui fournissent des éléments qui correspondent à la représentation mentale qu'il peut se faire d'un ensemble documentaire. A ce titre, la construction de la RTO multi-plans, et par là la possibilité de projeter les dimensions correspondantes, est un facilitateur du processus anthropogène d'immersion.

La distinction des types d'informations en dimensions offre un potentiel important de combinaisons entre informations. Dès lors que les types sont bien cernés, il devient possible de sélectionner uniquement ceux qui paraissent pertinent à l'utilisateur, et de les croiser afin d'en dégager des points de vue spécifiques en fonction des besoins. Par exemple, une étude de positionnement pour un institut fera intervenir *a minima* la dimension des organisations, pour positionner l'institut par rapport à elles, mais également celle des thèmes, afin de dégager les forces et faiblesses de l'institut client en fonction des domaines abordés. De manière à cibler les informations, il est également possible de se concentrer par exemple sur un certain intervalle de temps, et/ou de focaliser l'information sur les instituts européens.

La ressource structurant et représentant les connaissances impliquées permet donc une exploitation des informations au sein de dimensions.

En pratique, il manque, pour en tirer le bénéfice potentiel, un moyen d'attribuer un rôle à chacune de celles qui rentrent en jeu dans un point de vue adopté.

8.3.1.3 Moyen d'interaction : les critères de requête

Nous avons posé, dans le chapitre 2 (voir la section 2.2 page 28), que la tâche globale de l'utilisateur, soit, génériquement, la réalisation d'une étude, était un *problème mal défini*. Cette notion est utilisée par [Simon, 1973] et [Falzon, 2005], entre autres, pour désigner des problèmes dans lesquels les données sont peu claires et le but flou. La recherche d'information est l'une des solutions possibles pour résoudre un tel problème, et peut elle-même représenter, en tant que sous-tâche de la tâche globale, un problème mal défini.

La première condition à la recherche d'information par un individu pour l'exécution une tâche plus globale, est que ce dernier ait conscience qu'il manque de connaissances pour cette tâche. Pour atteindre cette prise de conscience, il doit tout de même posséder des connaissances, et notamment celles qui lui permettent de savoir qu'il ne dispose pas des connaissances nécessaires [Tricot, 2003]. Ces connaissances « faibles » lui permettent également de formuler la requête

complexe lui permettant d'accéder à l'ensemble documentaire, puisqu'elles lui permettent de penser, formuler, puis formaliser dans le système d'immersion son besoin d'information.

Tenir compte des types de connaissances « faibles » qui président à la construction de la requête est nécessaire pour laisser le champ libre à l'utilisateur, sans toutefois le perdre dans sa formulation. Ces connaissances sont de trois ordres : la connaissance de la tâche à réaliser, celle de son but de recherche, même si celui-ci est flou, et enfin, les connaissances procédurales nécessaires à l'utilisation du système d'immersion.

Lorsqu'il s'apprête à réaliser une recherche d'information, l'utilisateur le fait en fonction d'une tâche principale. Or, un état de l'art technique ne porte pas prioritairement sur les mêmes types d'informations qu'une étude de positionnement. L'utilisateur est conscient de cette différence, et sait *a priori* quels seront les types d'informations les plus pertinents, et quel rôle ils doivent jouer.

De même, le but de recherche influence la construction de la requête : un besoin flou sur des cibles distribuées ne donnera pas lieu à la même requête qu'un besoin précis sur cible localisée.

Enfin, la connaissance que peut avoir l'utilisateur de l'outil d'immersion est impliquée dans la création de la requête : un analyste ayant une bonne maîtrise de l'outil aura une représentation précise du rôle de chaque argument de la requête, ce qui ne sera pas le cas d'un utilisateur novice. Nous rappelons cependant qu'étant donnée l'utilisation intensive des outils TKM, un utilisateur novice devient expert rapidement.

Les connaissances de l'utilisateur, lui permettant de savoir qu'il a besoin d'effectuer une recherche d'information, sont donc plus ou moins importantes. Cependant, elles sont suffisantes pour lui permettre de former sa représentation mentale de l'ensemble documentaire (ED) auquel il doit accéder. Les moyens d'entrée dans cet ED, c'est-à-dire les arguments de la requête et leur contenu, doivent donc offrir une représentation adéquate à celle de l'utilisateur, afin que celui-ci puisse formuler la requête, en adéquation avec les informations qu'il a déterminées *a priori* comme pertinentes pour exécuter la tâche qui lui est dévolue.

Puisqu'il est celui qui possède les connaissances lui permettant d'en construire de nouvelles, il convient à notre sens de lui laisser prendre la main sur le système afin que lui-même décide de la manière dont il accède au système. Pour cela, il construit une requête correspondant à sa tâche globale et à son but de recherche. De cette manière, nous lui restituons son libre arbitre quant aux moyens d'accès à l'ensemble documentaire.

Nous avons présenté, en 6.3 page 222, les critères d'accès à l'immersion que nous avons mis en place. Chacun des critères correspond à l'un des arguments de la requête structurée de l'utilisateur pour accéder aux documents. Nous rappelons ces informations dans le tableau 8.2 page 357 résumant ces critères d'accès et la requête formulée.

L'utilisateur sélectionne lui-même les dimensions, ainsi que leur rôle de point d'accès, de modalité de visualisation ou de critère de filtrage. L'accès aux documents se fait donc de manière

Critère d'accès	Argument de la requête	Information nécessaire
Point d'entrée	Je veux voir...	Type(s) de données
Modalité de visualisation	Calculé et/ou visualisé par...	Type(s) de données/ regroupement(s)/ calcul(s)
Critère de filtrage	Filtré selon...	valeur(s)

TABLE 8.2 – Requête générique pour l'immersion définie dans la section 6.3 page 222 : critères d'accès, arguments de requête et informations en jeu

anthropogène, puisque le système sollicite l'utilisateur qui, avec ses connaissances, conçoit la requête dans le détail.

A partir de l'éventail de possibilités de requêtes virtuelles, c'est justement l'utilisateur qui a pour rôle, en tant qu'agent connaissant, d'actualiser certaines de ces possibilités. Pour cela, il décide des dimensions à prendre en compte, et de quel rôle elles doivent jouer, puis du paramétrage des dimensions sélectionnées.

La formulation générique de la requête (tableau 8.2 page 357) permet donc à l'utilisateur d'agir sur les inscriptions numériques par leur sélection, mais aussi en leur attribuant un rôle correspondant à un besoin spécifique. C'est à la fois la structuration de la requête et de la ressource utilisée, mais aussi la souplesse qu'elles permettent, qui autorisent réellement l'utilisateur à participer activement au fonctionnement du système, et qui évitent son cantonnement à un rôle d'utilisateur passif, qui ne pourrait par exemple agir que par le choix de chaînes de caractères à chercher dans les documents, sans autre distinction sur le rôle de ces chaînes.

En l'occurrence, nous laissons à l'utilisateur son libre arbitre, puisqu'il est concepteur de la question. Ses moyens humains de la formuler reposent sur la représentation qu'il se fait de l'ensemble documentaire par rapport à une tâche. Cette représentation trouvant écho dans la représentation numérique que nous avons mise en place, il a la liberté de poser toutes les questions qu'il souhaite. En pratique, les critères de recherche, l'agencement et le classement des résultats, leur filtrage, et leur visualisation sont laissés à son appréciation. A ce titre, le processus d'entrée est anthropogène, puisque l'utilisateur définit la requête, en fonction des paramètres que sont ses besoins spécifiques. De l'autre coté, nous donnons au système d'immersion les moyens de répondre à ces questions. Il entre donc par cette requête dans le système, et commence son immersion dans les documents de l'ensemble étudié.

8.3.2 L'agent connaissant immergé comme concepteur de la connaissance

Une fois que l'utilisateur a accédé à l'ensemble documentaire grâce à sa requête, des représentations visuelles, ou vues, lui sont présentées. Elles correspondent, dans leur forme comme dans leur contenu, aux critères que ce dernier a définis. A partir de ce point, l'utilisateur est immergé

dans les documents de l'ensemble, *via* le système.

Par le modèle d'immersion documentaire que nous avons élaboré, nous cherchons à mettre en place un système « qui donne à penser, et non [un système] qui pense » [Charlet, 2004]. Notre objectif est donc que l'utilisateur, en tant que sujet humain, construise ses propres connaissances à partir des propositions du système.

L'utilisateur, dans un premier temps, perçoit les informations présentes sur la vue. Dans un deuxième temps, à partir de cette perception, il peut les interpréter. Idéalement, la visualisation qui lui est présentée correspond effectivement à la représentation mentale de son besoin face à l'ensemble documentaire. Les éléments des dimensions informationnelles, *via* les arguments de la requête, forment alors des clés d'interprétation fournies à l'utilisateur. L'interprétation dépend pour partie de la manière dont les informations sont agencées au sein de la vue : en fonction de leur répartition, mais aussi des relations qui les unissent, elles peuvent signifier différentes choses pour l'utilisateur.

Cette interprétation donne lieu à la construction de connaissances, grâce à l'expérience vécue par l'utilisateur face à l'objet observé. En l'occurrence, cet objet est double : il est à la fois unitaire, puisque de la vue elle-même dans sa globalité peuvent être tirées des interprétations. D'autre part, cet objet est multiple : les éléments représentés dans la vue, ainsi que les relations les unissant, peuvent tous donner lieu à une interprétation et donc à des connaissances.

L'utilisateur ne fait donc pas que recevoir l'information. Il élabore, construit des représentations mentales de ce qui lui est présenté, en fonction de son objectif de recherche. L'aspect téléologique de la construction des connaissances joue donc ici aux deux niveaux de l'immersion documentaire : la finalité de l'utilisateur a permis de produire la requête d'interrogation de l'ensemble documentaire. Elle est aussi impliquée dans l'interprétation des résultats renvoyés. Pour des objectifs différents, les éléments d'informations présents ne seront pas exploités de la même façon.

Les connaissances construites à l'issue de l'interprétation d'une vue donnée peuvent être réutilisées dans la suite du cheminement au sein de l'ensemble documentaire. Elles guident en effet l'utilisateur dans la définition ou le réajustement de son but de recherche, puisqu'elles modifient sa représentation mentale.

8.3.3 L'agent connaissant immergé en interaction avec le système

Nous venons d'expliquer que l'humain, en tant qu'être intelligent, construit sa propre connaissance à partir de l'expérience qu'il tire de l'objet ou des objets observés. Il y a donc une interaction de fait entre objet et sujet, et celle-ci résulte en une expérience. L'interaction, selon l'axiome téléologique du constructivisme (voir la sous-section 2.3.2 page 54), est par ailleurs guidée par l'objectif du sujet par rapport aux objets. Dans notre cadre, l'utilisateur consulte l'ensemble

documentaire par immersion afin de répondre à une question, en relation avec la tâche principale qu'il doit accomplir.

Cet objectif global peut se décomposer en plusieurs éléments, correspondant à des étapes dans la construction de ses connaissances. Si l'interaction entre sujet et objet est toujours présente à partir d'une vue donnée, elle peut aussi se matérialiser différemment en fonction de ces étapes et des connaissances qui en ont été tirées.

En effet, sans interaction, l'utilisateur ne pourrait orienter les informations et leur visualisation sur les éléments qui l'intéressent en fonction de son but. Pour lui laisser la possibilité d'interagir avec les objets, le système d'immersion propose, au-delà d'une vue donnée considérée pour elle-même, des vues dynamiques et évolutives, plutôt des vues pré-déterminées.

Les vues sont évolutives et interactives selon deux modes :

1. le point de vue sur une visualisation donnée peut être précisé, élargi ou déplacé, en fonction du sous-objectif immédiat, impliquant alors toujours cette même visualisation à des échelles et sur des points divers ;
2. une vue particulière peut laisser place à une autre vue, construite sur des critères différents.

8.3.3.1 L'agent connaissant immergé comme acteur du cheminement

A partir d'une requête donnée, une vue particulière sur l'ensemble documentaire est proposée à l'utilisateur. Elle contient un certain nombre d'éléments d'informations, agencés et filtrés selon certains critères qu'il a déterminés. Partant de cette vue, prenant la forme d'une unique image dynamique, celui-ci peut tirer un certain nombre de connaissances par son interprétation.

Il peut cheminer dans la vue, par la sélection de points, de zones, ou de liens entre points/zones particuliers. Chaque sélection est un « pas » dans le cheminement de l'utilisateur, au sein du monde représenté (la vue) correspondant à une requête.

A chaque pas, il peut construire une nouvelle connaissance, qui peut être utilisée, sans que cela soit obligatoire, pour décider de la direction et de la destination du prochain pas dans le parcours.

Le chemin n'est donc pas déterminé à l'avance : l'intelligence de l'utilisateur, à travers le processus anthropogène de parcours, construit l'itinéraire à chaque étape.

Les pas effectués sont de deux sortes :

– les zooms avant ou arrière ;

– les déplacements latéraux/linéaires.

Les zooms avant et arrière permettent l'alternance de vues plus ou moins globales, ou plus ou moins locales, pour un même monde représenté.

Par exemple, une vue peut consister à représenter un « monde » d'auteurs répartis par organisations sur la totalité de l'ensemble documentaire ; si l'utilisateur est intéressé plus précisément

par une organisation parmi elles, il peut alors la sélectionner afin d'obtenir des détails. L'objectif pour lui peut être d'accéder aux auteurs de cette organisation, et à leur répartition précise en fonction de différentes filiales s'il s'agit d'une maison-mère, ou en fonction de laboratoires s'il s'agit d'une université. Son action consiste dans ce cas à réaliser un zoom avant pour préciser la vue, et à la focaliser sur un élément d'intérêt.

Inversement, il peut souhaiter passer d'une vue focalisée à une vue plus large. S'il est arrivé par exemple à une visualisation des thèmes abordés en fonction des pays à l'échelle européenne, il peut souhaiter étendre les informations à la totalité des pays du monde. Dans ce cas, le pas qu'il effectue consiste à élargir la vue pour atteindre le niveau mondial, à la place du niveau européen.

Son cheminement peut également l'amener à sélectionner des documents particuliers pour les consulter dans le détail. Puisque le lien est systématiquement conservé entre documents et informations qui en sont tirées, il peut accéder à ces documents complets à la demande.

Les déplacements linéaires sont une manière de déplacer le point de focalisation de la vue. Pour cela, l'utilisateur peut utiliser les liens entre éléments d'informations visualisés, jusqu'à atteindre les points souhaités.

8.3.3.2 L'agent connaissant capable de reformulation

La deuxième possibilité d'évolution dans la visualisation consiste à générer une autre vue que la vue actuelle, construite sur une requête différente. Cette possibilité est notamment utilisée lorsque la vue obtenue en premier lieu ne correspond pas à l'attente de l'utilisateur.

Il peut alors souhaiter reformuler son besoin à travers sa requête, si la vue obtenue ne lui renvoie pas les informations auxquelles il avait l'intention d'accéder. Cette vue inadéquate peut venir du fait qu'il a mal formulé sa requête, en raison par exemple de connaissances procédurales limitées s'il s'agit d'un utilisateur novice ; ou bien, la vue qu'il perçoit et interprète ne lui permet pas d'accéder à ce qu'il pensait obtenir au départ. Dans ce cas, deux possibilités s'offrent à lui : il peut reformuler sa requête, en changeant tout ou partie des critères afin d'obtenir une nouvelle vue potentiellement pertinente pour lui. Il peut également choisir de retourner à la vue précédente, afin de s'en servir à nouveau comme point d'étape dans l'immersion. Nous représentons ces deux actions possibles sur la figure 8.11 page 362.

La possibilité de revenir en arrière dans le parcours ou de reformuler la requête est une composante fondamentale du système d'immersion. Les erreurs de l'utilisateur participent en effet à l'évolution de son but : elles permettent de le réajuster, grâce aux informations, mêmes erronées dans son contexte, qui lui sont présentées. Que l'échec soit complet, si aucun aspect de la vue ne correspond à son besoin, ou partiel, si seuls certains aspects sont pertinents, des connaissances peuvent en être déduites. En effet, selon les positions constructivistes, l'échec en

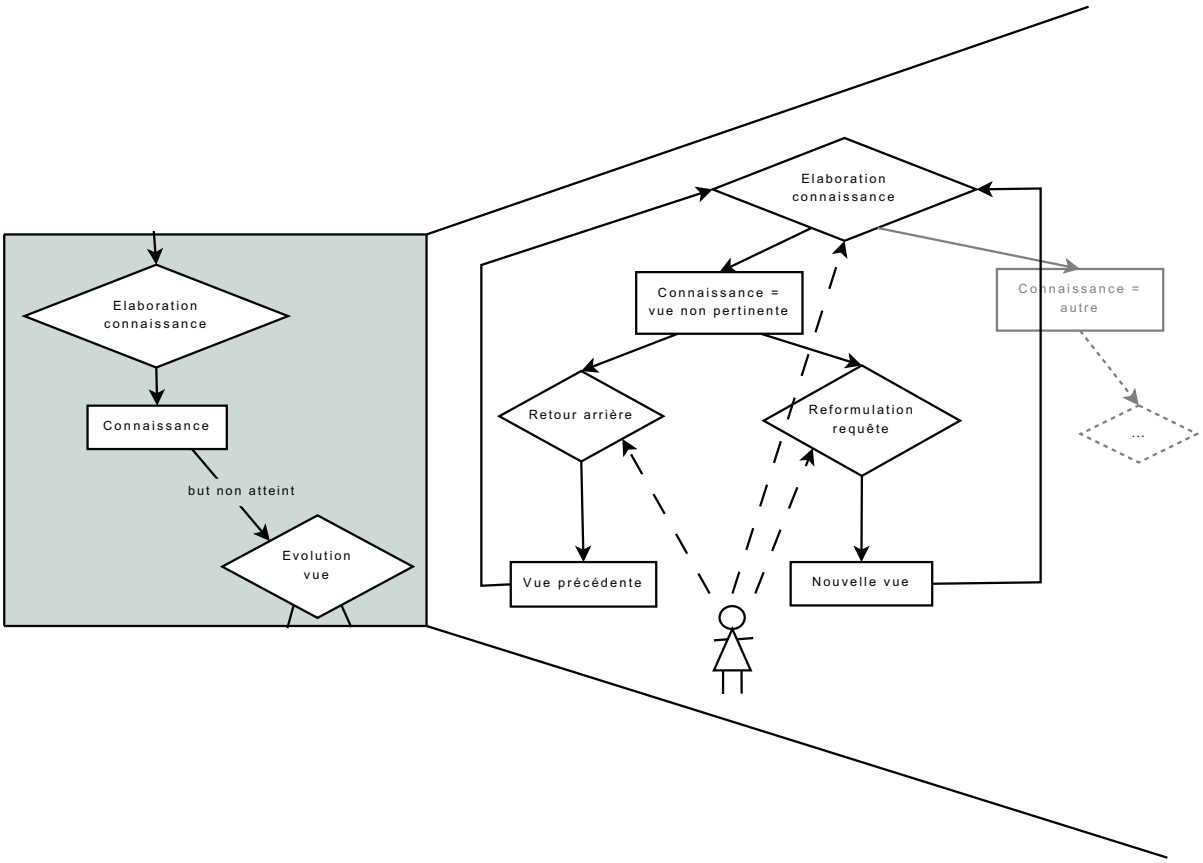

FIGURE 8.11 – Processus de reformulation de demande utilisateur

soi est une expérience source d'apprentissage [von Glasersfeld, 1994] : il permet d'acquérir des connaissances sur des actions ou opérations qui ne sont pas pertinentes. Il convient donc de prendre en compte ces erreurs et de permettre à un utilisateur de les corriger de la manière la plus simple possible pour lui. Des retours en arrière sur un clic, ou la reformulation de requête, sont des solutions.

Le même processus de reformulation peut être appliqué lorsqu'un utilisateur a parcouru la vue, pertinente pour lui, et qu'il souhaite passer à un autre point de vue sur d'autres informations venant compléter les connaissances acquises par la première visualisation.

8.3.4 L'agent connaissant émergé comme rédacteur

Le cheminement de l'utilisateur dans l'ensemble documentaire prend fin lorsqu'il estime que son besoin d'information global, par rapport à son activité, est satisfait, et donc que son but global de recherche est atteint. Il « émerge » alors de l'ensemble documentaire.

Il dispose à ce stade, en principe, des connaissances permettant de répondre à la question qu'il se posait au départ, dérivée de la question posée par le client. Il détermine alors, grâce à

elles, les informations qu'il doit utiliser dans son document livrable. L'utilisateur devient donc rédacteur, puisqu'il va rédiger en se fondant sur les connaissances acquises le document résultant de son activité.

La tâche de rédaction, tâche secondaire par rapport à la tâche principale de conseil, et intervenant après celle de recherche d'information, implique l'utilisation des connaissances acquises grâce à l'ensemble documentaire de plusieurs manières.

D'une part, elle intègre *a minima* une synthèse des informations tirées de l'ensemble documentaire, permettant de répondre à la question, ou au moins à une partie de la question.

D'autre part, si la réponse attendue consiste entre autres en éléments précis et ciblés, alors des éléments spécifiques peuvent aussi être nécessaires. Il peut s'agir de données chiffrées, comme le nombre de documents parus sur un domaine donné à une série de dates précises par exemple. La réponse peut également consister en une liste d'éléments d'informations, comme une liste des acteurs présents dans un domaine donné. Enfin, certains documents eux-mêmes peuvent représenter une information cruciale, du fait même de leur existence, ou bien par le contenu qu'ils véhiculent. Tous ces éléments doivent alors être inclus dans le document final de l'analyste. Ils peuvent l'être de manière textuelle et linéaire, mais aussi de manière graphique, par l'inclusion de vues tirées du système d'immersion.

Ainsi, à partir d'éléments de connaissances, l'analyste en construit un autre - le document. Il permettra au lecteur, c'est-à-dire au client, de construire lui-même ses connaissances pertinentes, pour sa propre tâche, à partir des informations délivrées. En effet, la tâche principale de l'analyste représente, pour le client, une tâche secondaire rattachée à sa propre tâche principale.

8.3.5 Conclusion

Nous avons vu que le processus d'immersion était en grande partie anthropogène : c'est en effet l'intelligence de l'utilisateur qui lui permet d'abord de pénétrer dans l'ensemble documentaire, puis lors de l'interprétation et de la génération des connaissances à partir des vues proposées en résultat de la requête. D'autre part, l'interaction entre humain et système lors de l'immersion dans ces résultats utilise également les capacités de raisonnement de l'utilisateur, puisque c'est lui qui détermine son chemin au sein des documents. Enfin, c'est encore elle qui est mise à contribution pour la rédaction du document final à livrer au client, et qui se fonde sur les connaissances tirées de l'immersion.

A certains égards, une requête lancée sur un moteur de recherche comme Google pourrait être considérée comme anthropogène. Cependant, ce type d'outils laisse en réalité un champ d'action très limité à son utilisateur. En effet, à peu de chose près, la suite de mots choisis pour la requête est la seule liberté accordée à ce dernier. Les critères de recherche, l'agencement et le classement des résultats, leur visualisation, sont imposés de fait par le moteur de recherche. De

plus, il permet peu d'interaction et pas d'alternance entre vues globales et locales.

A l'inverse, tout le système d'immersion repose sur un enrichissement mutuel entre le système et l'utilisateur : celui-ci injecte d'abord ses connaissances « faibles » dans le système par le biais d'une requête. Les résultats lui permettent alors de construire de nouvelles connaissances, qu'il peut réutiliser pour cheminer dans l'ensemble documentaire, cheminement qui renverra de nouveaux résultats. Ce processus prend fin lorsque l'utilisateur a dégagé les connaissances qui lui permettent de répondre à la question posée par le problème défini que représente sa tâche globale.

La construction des connaissances s'inscrit donc dans une démarche cyclique, liée à l'évolution du but de recherche d'information de l'utilisateur. Celui-ci se positionne donc comme concepteur de sa propre connaissance pertinente, grâce à l'interaction existant avec l'ensemble documentaire. Dans ce processus interactif, les éléments des dimensions informationnelles, *via* les arguments de la requête, peuvent être considérés comme des « panneaux de signalisation » permettant à l'utilisateur de se diriger pour atteindre son but, qui est la constitution de connaissances suffisantes pour répondre à sa question.

8.4 Conclusion

L'ensemble des niveaux de granularité incluant les unités en jeu dans nos travaux intègrent l'intelligence de l'utilisateur, en faisant appel à des processus et ressources anthropogènes.

La normalisation des entités nommées, au niveau le plus fin, appelle la participation active de l'utilisateur dans un processus de normalisation assistée, après que les autres types de traitements ont été appliqués. Ses décisions sont gardées en mémoire, selon un processus de capitalisation lui aussi anthropogène. Lors de la normalisation assistée, les données sont capitalisées pour l'étude en cours, et donc pour l'ensemble documentaire correspondant, mais également pour la totalité des données disponibles dans les bases de données internes de la société TKM. Cependant, l'utilisateur peut déterminer lui-même la pertinence d'une telle capitalisation globale : si elle ne lui paraît pas opportune, il peut signifier au système qu'une décision particulière, ou un ensemble de décisions, ne doivent impacter que l'ensemble documentaire pour lequel elles ont été prises.

Au niveau des unités d'informations, le processus de normalisation anthropogène se place donc à un niveau global et à un niveau local, sollicitant l'intelligence de l'utilisateur selon deux points de vue différents.

La ressource termino-ontologique (RTO) multi-plans, quant à elle, exploite non pas des processus dédiés, mais des processus anthropogènes multi-granularités, dont les résultats ont une portée sur sa constitution pour une étude donnée. Les résultats des processus anthropogènes du niveau des entités nommées, en particulier, sont capitalisés et réinjectés dans la RTO. Parallèlement, le modèle même de cette RTO a été conçu de manière anthropomorphique, puisqu'il

reproduit, certes très partiellement, la représentation mentale que l'utilisateur peut avoir d'un ensemble documentaire multi-sources. Prendre en compte cette représentation dans le modèle, et capitaliser les résultats anthropogènes du niveau inférieur, concourent à limiter l'intervention de l'utilisateur à ce niveau, ce qui en soi représente une amélioration des processus anthropogènes : minimiser l'intervention humaine limite les risques d'erreurs, mais aussi le risque de lasser l'utilisateur et de le détourner de traitements trop longs et trop complexes.

Enfin, le système d'immersion bénéficie, au niveau de sa construction, de l'ensemble des résultats de ces processus anthropogènes. De plus, il place l'utilisateur au centre des documents, le rendant partie prenante du système. Les processus anthropogènes sont donc fondamentaux pour son fonctionnement et son exploitation, puisque l'humain a la mainmise sur toutes les décisions relatives aux traitements, endogènes et exogènes, réalisés sur les documents à l'entrée et durant l'immersion. Il est décisionnaire des moyens d'accès aux documents, *via* les arguments de la requête, mais aussi du chemin qu'il parcourt pas à pas une fois immergé. A ce titre, les processus anthropogènes influencent l'usage des facettes de la RTO à travers les dimensions informationnelles de l'immersion. Ils permettent d'abord de réduire certaines structures, par leur « appauvrissement » raisonné et motivé par les besoins d'un analyste. Cette réduction est figurée dans la figure 8.12.

La réduction informationnelle peut être souhaitée par un utilisateur dans un contexte donné. Par exemple, s'il ne souhaite accéder qu'aux organisations de niveau hiérarchique supérieur, par exemple des universités, il peut mettre de côté toutes les organisations dépendant d'elles. Dans ce cas, il transforme une structure arborescente, obtenue par des processus endogènes puis exogènes, en une liste énumérative non ordonnée, qui ne contient plus que des unités se situant sur un même plan.

Inversement, les méthodes anthropogènes peuvent enrichir des structures de facettes, par extension. Nous les représentons sur la figure 8.13 page 366, qui résume l'ensemble des apports des trois méthodes - endogène, exogène et anthropogène - par type d'information.

Ces extensions portent plus particulièrement sur les facettes non hiérarchisées après extraction des thèmes et des auteurs, et sur les dimensions correspondantes. Pour toutes, l'utilisateur peut, lors de sa requête, décider de regrouper les unités informationnelles qu'elles contiennent, en particulier par le biais d'expressions régulières, ou par la sélection d'intervalles pour le cas particulier des dates. Les auteurs et les thèmes, facettes endogènes non ordonnées, peuvent également être ordonnés par ce biais.

L'utilisateur est donc actif dans le processus de construction de ses propres connaissances, et peut ensuite endosser son rôle de rédacteur du document à livrer au client, et contenant de nouvelles informations destinées à permettre au lecteur de construire à son tour de nouvelles connaissances.

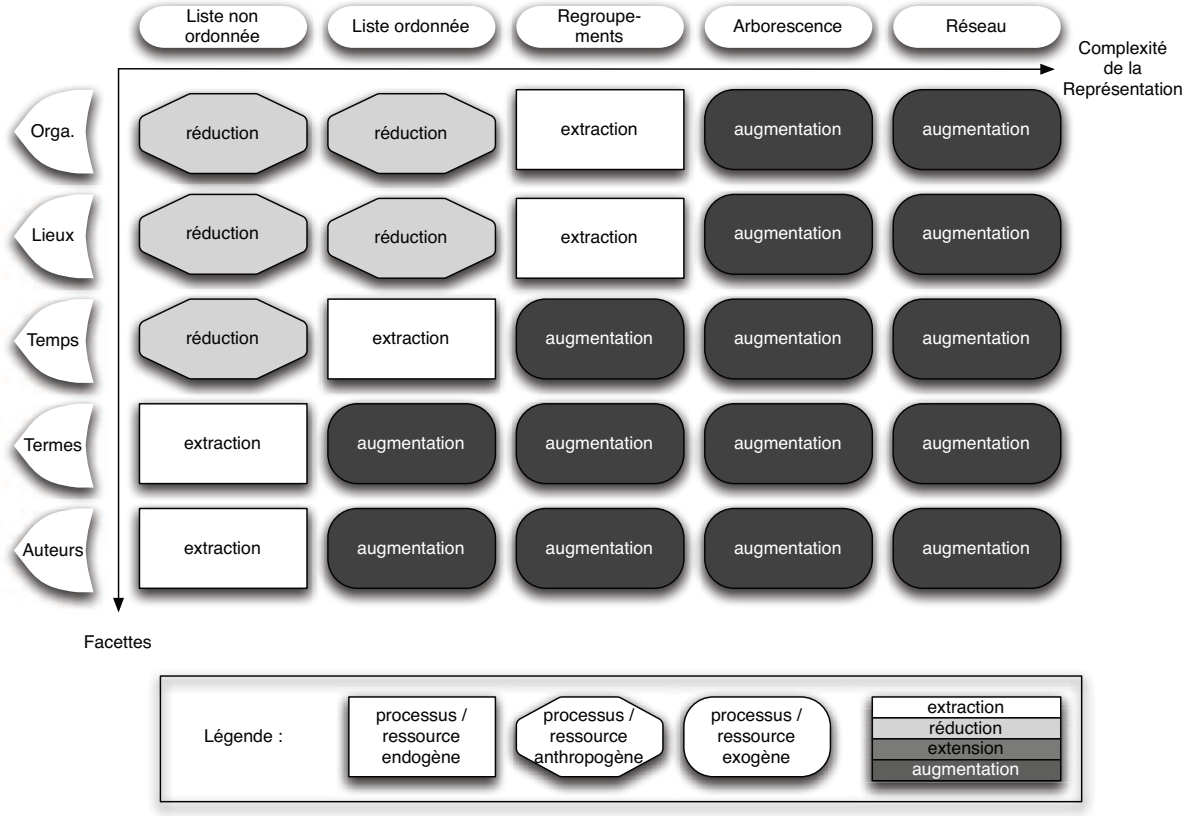

FIGURE 8.12 – Apport par réduction des ressources et processus anthropogènes sur chaque facette et structure résultante

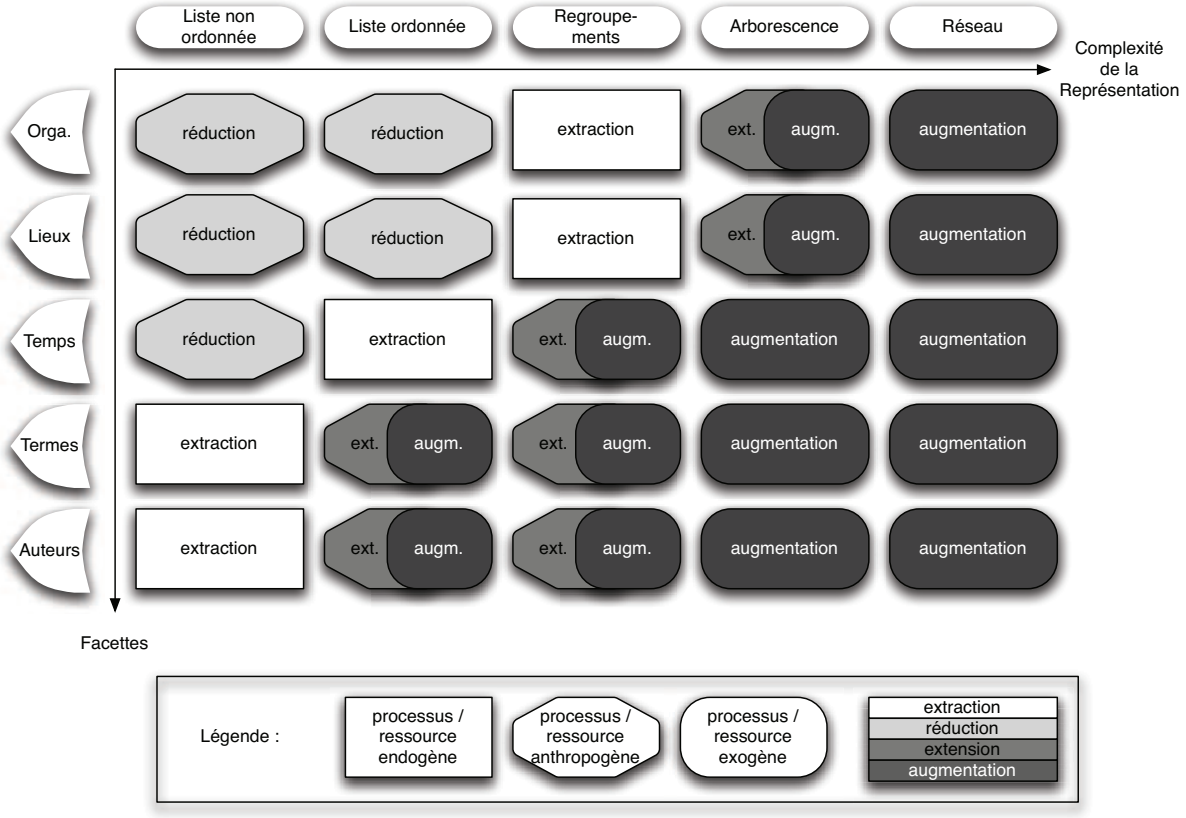

FIGURE 8.13 – Apport par extension des ressources et processus anthropogènes sur chaque facette et structure résultante

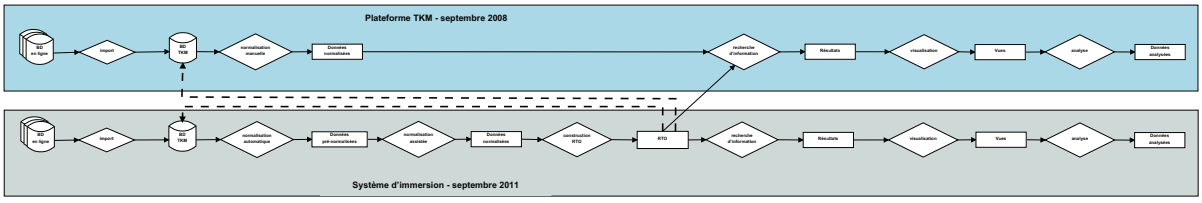

FIGURE 8.14 – Avant après

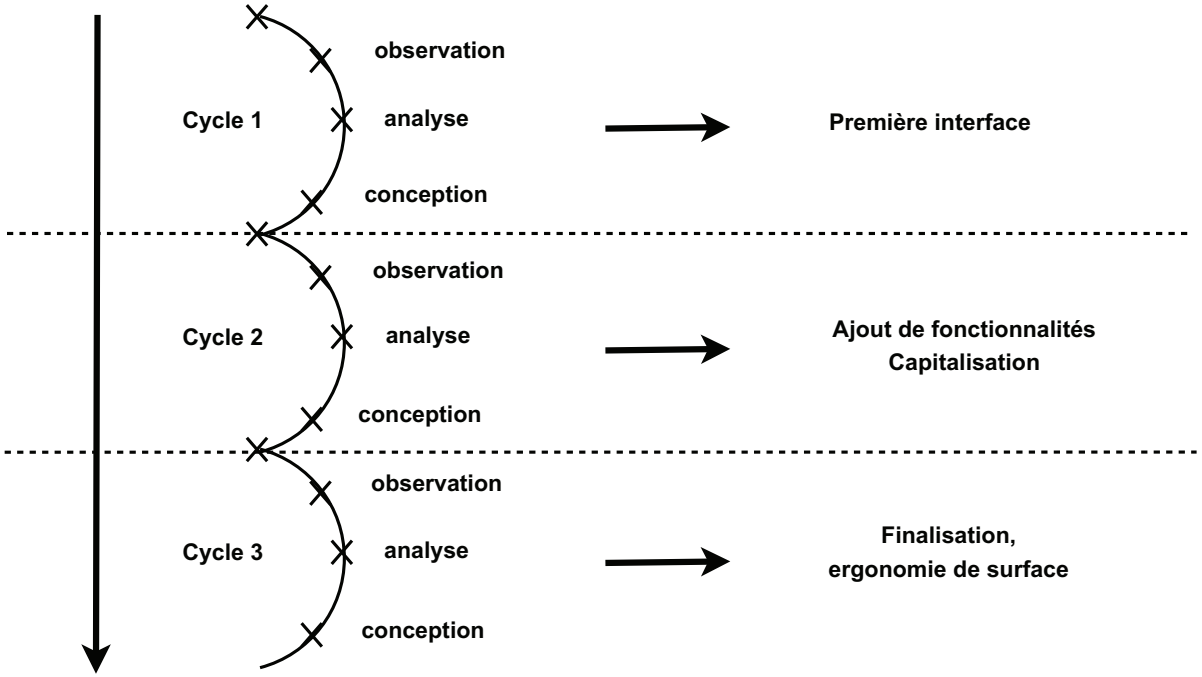

FIGURE 8.15 – Avant après

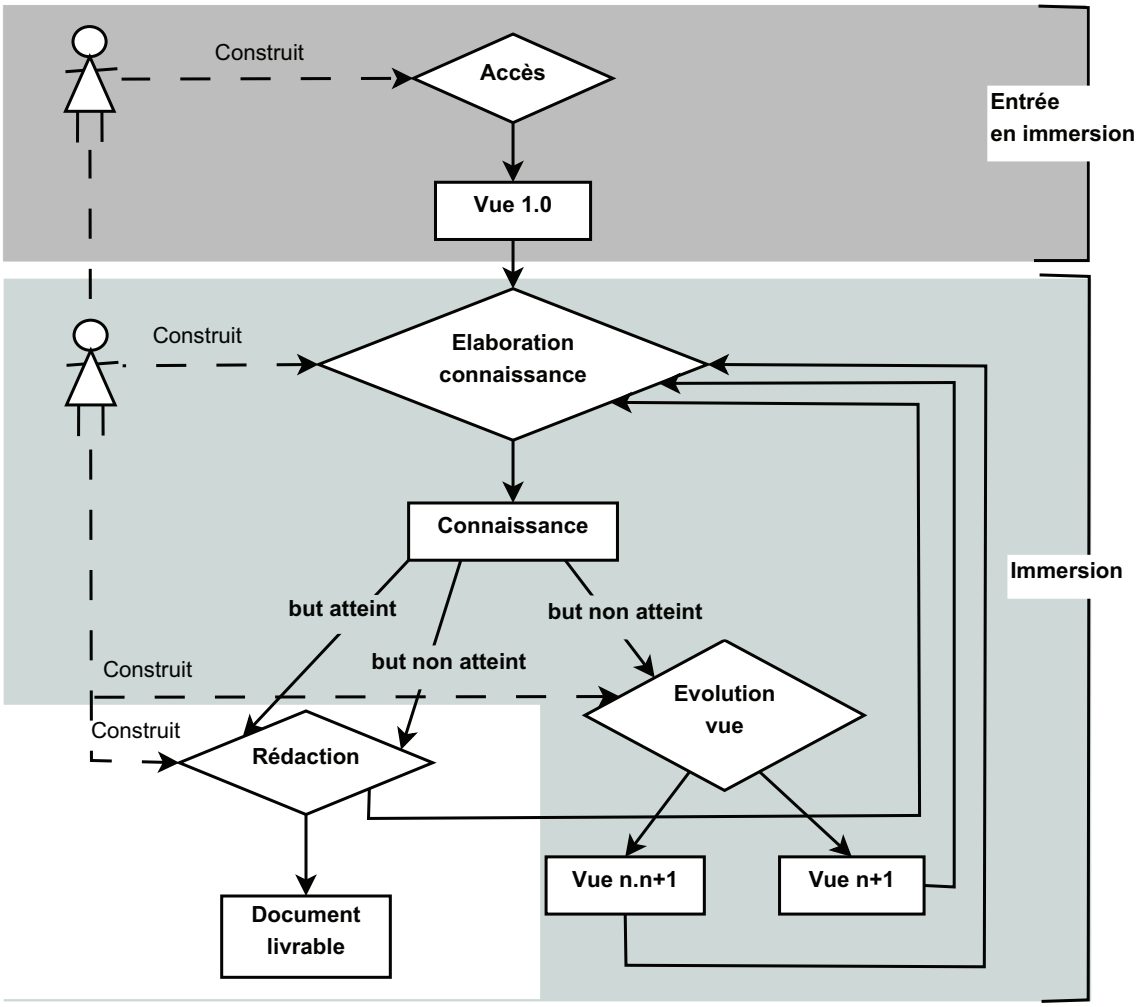

FIGURE 8.16 – Avant après

Conclusion

Ce n'est pas la fin. Ce n'est même pas le commencement de la fin.
Mais, c'est peut-être la fin du commencement.
Winston Churchill

La problématique centrale de cette thèse a porté sur l'accès à l'information scientifique et technique dans de grands ensembles documentaires textuels, dans un objectif d'activité ciblée en milieu industriel.

Nous avons appuyé, pour tenter d'y répondre, sur la prépondérance de la notion de pertinence (chapitre 1). Celle-ci est relative, et dépend de la tâche principale pour laquelle un utilisateur doit accéder à l'information, mais aussi du contexte de la recherche et de son thème. Tous ces paramètres, associés à ceux qui sont rattachés à la nature et au contenu des documents (chapitre 2), ont déterminé la modélisation de tous les éléments du système final. Ils influencent la façon dont les documents et les informations qu'ils véhiculent sont modélisés (chapitre 3), à partir des unités textuelles les plus saillantes (chapitre 4). Les choix de représentation cartographique des unités et des relations du modèle obtenu sont également fondés sur ces critères. Enfin, ils ont également présidé à la définition de la forme du système final : un système d'immersion documentaire.

Cette modélisation a été réalisée par l'apport théorique et pratique de plusieurs champs de recherche, distincts bien que convergents. En particulier, l'ingénierie des connaissances et l'ergonomie, deux disciplines orientées sur les épistémologies constructivistes, mais également la linguistique et le traitement automatique des langues, ont jeté les fondations de notre structure de représentation des connaissances et de l'aspect calculatoire du système d'immersion, à l'intérieur duquel l'utilisateur prend place. Puis, plus spécifiquement consacrée à l'aspect cartographique des données, le champ de la visualisation d'information a également apporté une contribution importante.

Pour aboutir à un système d'immersion documentaire efficace, nous avons formulé l'hypothèse générale selon laquelle combiner trois types de méthodes, endogènes, exogènes et anthropogènes, était un moyen de parvenir à obtenir de meilleurs performances dans un système d'accès documentaire (chapitre 5).

Pour chacun des types de méthodes, nous avons postulé qu'elles apporteraient de l'information, difficile à obtenir par le biais d'autres types de ressources. Les méthodes endogènes tirent l'information directement de l'ensemble documentaire à traiter. Par là, elles sont adaptées à l'ensemble documentaire sur lequel se fondent les missions des analystes (chapitre 6). Les méthodes exogènes apportent de l'information externe à l'ensemble documentaire. Elles enrichissent les traitements endogènes, voire les augmentent, et permettent de limiter leurs coûts calculatoires (chapitre 7). Les méthodes anthropogènes, enfin, exploitent l'intelligence de l'utilisateur en tant que source de connaissances dynamique capitalisables. Elles rendent les données traitées par les

processus précédents plus fiables, par la correction de biais éventuellement introduits par les méthodes précédentes, et plus riches. Elles sont appliquées en aval des méthodes endogènes et anthropogènes, de manière à limiter une sollicitation trop importante de l'utilisateur (chapitre 8).

Des outils ont été mis en place à partir de chacune de ces méthodes : notre hypothèse était que l'apport de chacun d'eux serait d'optimiser les résultats des autres. Cela se vérifie pour les aspects intra-granulaires. Au niveau de granularité le plus fin, celui des entités nommées notamment (section 4.2 page 114), nous avons postulé que des méthodes de normalisation par mesure de similarité et par exploitation des récurrences du corpus donnaient les moyens de déceler et de résoudre des variations d'ordre typo-orthographique et parfois syntaxique, mais aussi de suggérer des segmentation pertinentes (section 6.1 page 159). Notre deuxième hypothèse pour ce niveau consistait à considérer qu'un système exogène à base de règles permettait quant à lui de résoudre d'autres types de variations, et de structurer les entités nommées hiérarchiquement (section 7.1 page 241). Enfin, nous avons supposé que des méthodes anthropogènes laissant l'utilisateur superviser les résultats des précédentes méthodes, à l'aide d'une interface de normalisation assistée, produiraient des données plus fiables, mieux structurées et plus élaborées (section 8.1 page 320).

Au niveau intermédiaire, celui de la ressource termino-ontologique (RTO) multi-plans, d'autres outils permettent de constituer une RTO pour chaque nouvel ensemble documentaire. Ils sont là encore fondés sur plusieurs hypothèses. Dans ce cadre, la ressource endogène employée est là encore constituée de l'ensemble documentaire lui-même et des différents types d'informations qu'il véhicule. Plus précisément, elle contient les informations contextuelles qui en ont été tirées lors de la normalisation des entités nommées, mais aussi les informations thématiques sous forme de collocations brutes, extraites à partir de calculs de segments répétés (section 6.2 page 199). Une autre de nos hypothèses a consisté ici à considérer que des méthodes exogènes passant par des ressources statiques, et plus précisément des lexiques, viendraient compléter ces calculs et les informations contextuelles : pour celles-ci, ils permettent de réaliser des appariements. Pour ceux-là, les lexiques affinent les calculs d'extraction des collocations brutes (section 7.2 page 284). Enfin, nous avons postulé que les processus anthropogènes réalisés lors du niveau précédent pourraient structurer et fiabiliser les données intégrées à la RTO. De plus, la conception même du modèle de la structure de représentation est dérivée d'une représentation mentale partielle de l'ensemble documentaire par les utilisateurs. En ce sens, la RTO multi-plans est non pas anthropogène, puisqu'aucune action directe de l'utilisateur n'a présidé à la conception de son modèle (contrairement à sa construction pour un ensemble documentaire donné), mais anthropomorphe. Or, nous sommes partie du principe que cet anthropomorphisme ouvrirait un champ de possibilités qui seraient adaptées à l'utilisateur pour l'immersion (section 8.2 page 343).

Enfin, au niveau de granularité le plus large, celui du système d'immersion documentaire, les trois types de méthodes fournissent des outils, pour sa construction comme pour son exploi-

tation. Les dimensions informationnelles sur lesquelles se fondent les calculs, les requêtes et les projections d'informations sont issues des facettes de la RTO multi-plans, endogène par nature (section 6.3 page 222). D'autre part, des ressources exogènes sont exploitées par d'autres processus qui, avons-nous postulé, permettent éventuellement d'augmenter les dimensions, puisqu'elles superposent à leurs informations d'autres informations de même type, souvent englobantes. Les ressources et processus cartographiques, qui fournissent les informations nécessaires à la représentation visuelle des données de l'ensemble documentaire, sont également exogènes par leur nature même (section 7.3 page 294). Enfin, une grande part des procédés de navigation sont anthropogènes, puisque nous nous sommes fondée sur l'hypothèse selon laquelle l'utilisateur, grâce à son intelligence, peut construire lui-même son itinéraire d'immersion, et par là sa propre connaissance pertinente. Fort de cette nouvelle connaissance, celui-ci peut alors la réinjecter lorsqu'il endosse son rôle de rédacteur du document livrable pour son client (section 8.3 page 349).

Ainsi, si les outils sont dédiés à une tâche précise pour un niveau de granularité donné, leurs résultats se complètent également au niveau inter-granulaire, grâce à une capitalisation : les résultats des traitements des unités informationnelles, quelle que soit leur nature, sont réinjectés dans les niveaux suivants, d'abord dans les traitements de construction de la RTO, puis dans le système d'immersion *via* les dimensions informationnelles projetées dans le système.

L'ensemble de ces hypothèses locales permet de mettre en place un outil global avec différents degrés de connaissances, c'est-à-dire des connaissances intrinsèques, extrinsèques et humaines, à chaque niveau de granularité. De leur combinaison résulte l'efficacité de l'outil d'immersion.

Dans l'ensemble des outils conçus, et de manière plus appuyée encore dans le système d'immersion, nous mettons l'utilisateur en avant, en tant qu'être intelligent et en tant qu'agent connaissant (chapitre 8 notamment). Tout d'abord, nous le plaçons face à de l'information divisée, scindée en plusieurs catégories, grâce à la structuration dont elle fait l'objet. En effet, nous représentons les informations au sein de facettes, qui nourrissent les dimensions informationnelles du système d'immersion. Ces dimensions correspondent à celles que nous avons mis au jour lors de l'analyse de l'activité des utilisateurs (chapitres 2 et 3 notamment), et qui structurent au moins partiellement la représentation mentale qu'ils peuvent avoir des ensembles documentaires à partir desquels ils travaillent. De fait, les informations telles qu'elles sont restituées sont adaptées à leur activité.

De plus, nous immergeons l'utilisateur dans les documents, grâce à la projection de ces informations au sein d'un espace multidimensionnel, structuré grâce aux informations contextuelles portées par les entités nommées des métadonnées des documents et par les collocations brutes vectrices de thèmes. Il peut alors parcourir cet espace, par un chemin qu'il construit lui-même et pas à pas. Il tire de chaque étape de son itinéraire des connaissances, qu'il peut utiliser pour définir la direction ou la profondeur du pas suivant. Lorsque la connaissance qu'il a construite lui paraît suffisante, son besoin d'information est satisfait, et il peut alors sortir de l'immersion.

A partir de là, il peut réutiliser cette connaissance dans la réalisation de sa tâche principale de conseil.

Ce système d'immersion a été conçu dans le cadre d'une activité industrielle de conseil en innovation, exploitant des documents scientifiques et techniques. Dans ce contexte, il apporte un accès facilité et adapté à cette information. Pour autant, sa portée est limitée par certaines conditions. Il est conçu pour des utilisateurs spécialisés, qui travaillent sur des ensembles documentaires spécifiques, dans lesquels les unités de traitement sont identifiées, ou pour le moins identifiables. Un tel système ne pourrait par exemple pas être intégré à un moteur de recherche web, visant un public large et non spécialisé. En revanche, il pourrait être utilisé dans d'autres contextes spécialisés, même dans le cas où les domaines sont très différents de ceux que nous avons abordés.

Nous pensons en particulier au projet des Manuscrits de Stendhal [Lebarbé & Meynard, 2009], qui rassemble plus de vingt mille feuillets numérisés et en cours de transcription de l'auteur du XIXème siècle. Il est possible d'appliquer à leur version transcrite un tel système d'immersion, d'autant plus facilement qu'une grande partie des types d'informations constitutifs des facettes se recoupent. Les informations contextuelles n'ont pas à être des métadonnées : elles sont à l'heure actuelle identifiées et délimitées dans le corps des textes. Il s'agit alors de les intégrer aux plans correspondants. Le seul plan absent des manuscrits est celui des organisations. Par la structure en facettes et dimensions du modèle d'immersion, il est parfaitement envisageable d'en éliminer une dimension non pertinente, et d'en créer de nouvelles en fonction des besoins des utilisateurs relativement à d'autres types d'informations, à partir du moment où elles sont clairement identifiées. De même, nous avons délibérément écarté les processus complexes de la construction de la facette thématique. Cependant, d'autres structurations et représentations thématiques sont possibles, et peuvent être intégrées en lieu et place des collocations brutes, si le besoin est exprimé par les utilisateurs.

Le point crucial dans l'immersion est donc avant tout, et quelle que soit l'application spécialisée, de placer l'utilisateur au cœur du système, comme interprétant de l'information, acteur de son propre cheminement, et concepteur de la connaissance.

Bibliographie

[Abney, 1991] Abney, S. (1991). Parsing by chunks. *Studies in Linguistics and Philosophy*, 44, 257–278.

[AFNOR, 1987] AFNOR, Ed. (1987). *Vocabulaire de la documentation*. Comité de terminologie de l'AFNOR.

[Alphonse et al., 2004] Alphonse, E., Bessières, P., Bisson, G., Hamon, T., Lagarrigue, S., Nazarenko, A., Nédellec, C., Vetah, M. O. A., Poibeau, T., & Weissenbacher, D. (2004). Extraction d'information appliquée au domaine biomédical - apprentissage et traitement automatique de la langue. In *Actes de CIFT : Colloque International sur la Fouille de Texte*.

[Atinternet, 2011] Atinternet (2011). Baromètre des moteurs - décembre 2010. `http://www.atinternet.com/ressources/etudes/barometre-des-moteurs/barometre-des-moteurs-decembre-2010/index-1-1-6-218.aspx` [page consultée le 25 mars 2011].

[Aussenac-Gilles & Condamines, 2004] Aussenac-Gilles, N. & Condamines, A. (2004). Documents électroniques et constitution de ressources termniologiques et ontologiques. *Information-Interaction-Intelligence*, 4(1), 75–93.

[Bachimont, 2000] Bachimont, B. (2000). Engagement sémantique et engagement ontologique : conception et réalisation d'ontologies en ingénierie des connaissances. In J. Charlet, M. Zacklad, G. Kassel, & D. Bourigault (Eds.), *Ingénierie des connaissances, évolutions récentes et nouveaux défis*. Paris : Eyrolles.

[Bachimont, 2001] Bachimont, B. (2001). Dossier et lecture hypertextuelle : problématique et discussions. *Les cahiers du numérique*, (numéro spécial sur l'information médicale numérique), 105–123.

[Baleydier, 2004] Baleydier, L. (2004). Google ne satisfait pas le cerveau droit. *Recherche et référencement*, (50), 16–18.

[Beghtol, 1986] Beghtol, C. (1986). Bibliographic classification theory and text linguistics : aboutness analysis, intertextuality and the cognitive act of classifying documents. *Journal of Documentation*, (42), 84–113.

[Beguin & Rabardel, 2000] Beguin, P. & Rabardel, P. (2000). Concevoir pour les activités instrumentées. *Intelligence artificielle, numéro spécial, Interaction homme-système*, 14(1-2), 35–54.

[Bernhard & Ligozat, 2011] Bernhard, D. & Ligozat, A. (2011). Analyse automatique de la modalité et du niveau de certitude : application au domaine médical. In *Actes de TALN 2011*, volume 1 (pp. 433–444). Montpellier.

[Berthoud, 1996] Berthoud, A. C. (1996). *Paroles à propos. Approche énonciative et interactive du topic.* Paris : Ophrys.

[Bey, 2009] Bey, A. (2009). *Quantification de la variation de la norme - Calcul, modélisation et accessibilité des variations de l'arsenal législatif par une granularité variable.* Mémoire de master 2 recherche, Université Stendhal - Grenoble 3, Grenoble.

[Bilhaut, 2004] Bilhaut, F. (2004). Analyse automatique de la structure thématique du discours pour la navigation documentaire. In *Semaine du Document Numérique. Journée ATALA* La Rochelle.

[Bilhaut, 2005] Bilhaut, F. (2005). *Extraction automatique d'axes sémantiques pour l'analyse thématique du discours.* Rapport interne du GREYC, Université de Caen, Caen.

[Bonhomme et al., 1996] Bonhomme, P., Bruneseaux, F., & Romary, L. (1996). Codage, documentation et diffusion de ressources textuelles. *Cahiers GUTenberg*, 24(spécial TEI), 177–180.

[Bouaud et al., 1999] Bouaud, J., Seroussi, B., & Antoine, E. (1999). ONCODOC : une approche documentaire de l'aide à la décision. *Document numérique*, 3(3-4), 61–79.

[Bouaud et al., 1998] Bouaud, J., Séroussi, B., Antoine, E., Gozy, M., Khayat, D., & Boisvieux, J. F. (1998). Hypertextual navigation operationalizing generic clinical practice guidelines for patient-specific therapeutic decisions. *Journal of the American Medical Informatics Association*, (5 (suppl)), 488–492.

[Bouillon et al., 2000] Bouillon, P., Fabre, C., Sébillot, P., & Jacquemin, L. (2000). Apprentissage de ressources lexicales pour l'extension de requêtes. *Traitement automatique des langues pour la recherche d'information, TAL*, 41(2), 367–393.

[Boulaknadel, 2008] Boulaknadel, S. (2008). *Traitement automatique des langues et recherche d'information en langue arabe dans un domaine de spécialité : apport des connaissances morphologiques et syntaxiques pour l'indexation.* Thèse de Doctorat, Université de Nantes, Nantes.

[Bourigault, 1993] Bourigault, D. (1993). An endogeneous Corpus-Based method for structural noun phrase disambiguation. In *Proceedings of EACL'93 : 6th conference of the European chapter of the Association for Computational Linguistics* (pp. 81–86).

[Bourigault & Aussenac-Gilles, 2003] Bourigault, D. & Aussenac-Gilles, N. (2003). Construction d'ontologies à partir de textes. In *Actes de TALN 2003* (pp. 25–50).

[Bourigault et al., 2005] Bourigault, D., Fabre, C., Frérot, C., Jacques, M., & Ozdowska, S. (2005). Syntex, analyseur syntaxique de corpus. In *Actes de TALN 2005* Dourdan, France.

[Bourigault & Slodzian, 2000] Bourigault, D. & Slodzian, M. (2000). Pour une terminologie textuelle. *Terminologies Nouvelles*, (19), 29–32.

[Britt et al., 1999] Britt, M. A., Perfetti, C. A., Sandak, R., & Rouet, J. F. (1999). Content integration and source separation in learning from multiple texts. In S. Goldman, A. Graesser, & P. van den Broek (Eds.), *Narrative comprehension, causality, and coherence : Essays in honor of Tom Trabasso* (pp. 209–233). Mahwah, NJ : Erlbaum.

[Card et al., 1999] Card, S., Mackinlay, J., & Shneiderman, B. (1999). Information visualization. In *Readings in information visualization : using vision to think* (pp. 1–34). Morgan Kaufmann Publishers.

[Charlet, 2002] Charlet, J. (2002). *L'ingénierie des connaissances - Développements, résultats et perspectives pour la gestion des connaissances médicales.* Mémoire d'Habilitation à diriger des recherches, Université Paris 6, Paris.

[Charlet, 2004] Charlet, J. (2004). L'ingénierie des connaissances, entre science de l'information et science de gestion. In P. Lorino & R. Teulier (Eds.), *Entre connaissance et organisation, l'activité collective* (pp. 306–329).

[Charlet et al., 1999] Charlet, J., Bachimont, B., Brunie, V., El Kassar, S., Zweigenbaum, P., & Boisvieux, J. F. (1999). L'ingénierie documentaire au service du dossier patient électronique. In *L'informatisation du cabinet du futur, Informatique et Santé* Paris : Springer Verlag.

[Chinchor, 1998] Chinchor, N. (1998). Overview of MUC-7. In *Proceedings of the Seventh Message Understanding Conference* Fairfax, Virginia.

[Cho, 2001] Cho, M. H. (2001). *The role of prior knowledge, need for information and credibility of information in tourists' information search behavior.* Thèse de Doctorat, The Pennsylvania State University, University Park, PA.

[Church & Hanks, 1990] Church, K. & Hanks, P. (1990). Word association norms, mutual information and lexicography. *Computational Linguistics*, 16(1), 23–29.

[Coates-Stephens, 1993] Coates-Stephens, S. (1993). The analysis and acquisition of proper names for the understanding of free text. *Computers and the Humanities*, 26(5), 441–456.

[Cohen, 2005] Cohen, A. M. (2005). Unsupervised gene/protein named entity normalization using automatically extracted dictionaries. In *Proceedings of ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases* (pp. 17–24).

[Cohen, 1998] Cohen, W. W. (1998). Integration of heterogeneous databases without common domains using queries based on textual similarity. In *Proceedings of the 1998 ACM SIGMOD International COnference on Management of Data* (pp. 201–212).

[Condamines, 2005] Condamines, A. (2005). Linguistique de corpus et terminologie. *Langages*, (157), 36–47.

[Condamines & Rebeyrolle, 1997] Condamines, A. & Rebeyrolle, J. (1997). Construction d'une base de connaissances terminologiques à partir de textes : expérimentation et définition d'une méthode. In *Actes des Journées Ingénierie des Connaissances Apprentissage Automatique* (pp. 191–206). Roscoff.

[Cori & Léon, 2002] Cori, M. & Léon, J. (2002). La constitution du TAL. *TAL*, 43(3), 21–55.

[Cucerzan, 2007] Cucerzan, S. (2007). Large-scale named entity disambiguation based on wikipedia data. In *Proceedings of EMNLP-CoNLL'07 : Empirical Methods on Natural Language Processing and Computational Natural Language Learning* (pp. 708–716).

[Cunningham et al., 1996] Cunningham, H., Wilks, Y., & Gaizauskas, R. (1996). GATE - a general architecture for text engineering. In *Proceedings of the 16th Conference on Computational Linguistics* (pp. 1057–1060). Copenhague, Danemark.

[Déjean, 1998] Déjean, H. (1998). *Concepts et algorithmes pour la découverte des structures formelles des langues*. Thèse de Doctorat, Université de Caen Basse-Normandie, Caen.

[Dervin & Nilan, 1986] Dervin, B. & Nilan, M. (1986). Information needs and uses. *Annual review of information science and technology*, 21, 3–33.

[Ducrot, 1972] Ducrot, O. (1972). *Dire et ne pas dire*. Paris : Hermann.

[Ehrmann, 2008] Ehrmann, M. (2008). *Les Entités Nommées, de la linguistique au TAL : statut théorique et méthodes de désambiguïsation*. Thèse de Doctorat, Université de Paris 7, Paris.

[Ehrmann & Jacquet, 2006] Ehrmann, M. & Jacquet, G. (2006). Vers une double annotation des entités nommées. *TAL*, 47(3), 63–88.

[Elmagarmid et al., 2007] Elmagarmid, A. K., Ipeirotis, P. G., & Verykios, V. S. (2007). Duplicate record detection : A survey. *IEEE Transactions on Knowledge and Data Engineering*, 19, 1–16.

[Ericsson & Kintsch, 1995] Ericsson, K. A. & Kintsch, W. (1995). Long term working memory. *Psychological Review*, (102), 211–245.

[Fairthorne, 1969] Fairthorne, R. A. (1969). Content analysis, specification and control. *Annual review of information science and technology*, (4), 73–109.

[Falzon, 2005] Falzon, P. (2005). Ergonomie, conception et développement. In *40ème Congrès de la SELF* Saint-Denis, La Réunion.

[Farabet, 2010] Farabet, M. (2010). *Etat de l'art : les systèmes d'extraction terminologique*. Rapport interne, Université Stendhal - Grenoble 3 / TKM, Voiron.

[Fekete, 2010] Fekete, J. (2010). Visualiser l'information pour la comprendre vite et bien. In *L'usager numérique - séminaire INRIA* (pp. 161–194). Anglet : ADBS éditions.

[Fidel, 1994] Fidel, R. (1994). User-centered indexing. *Journal of the American Society for Information Science*, 45(8), 572–576.

[Folch & Habert, 2000] Folch, H. & Habert, B. (2000). Constructing a navigable topic map by inductive semantic acquisition methods. *Extreme Markup Languages 2000 - The Expanding XML/SGML Universe*, (pp. 55–61).

[Frege, 1892] Frege, G. (1892). Sens et dénotation. In *Ecrits logiques et philosophiques* (pp. 102–126). Paris : Le Seuil.

[Frérot et al., 2003] Frérot, C., Bourigault, D., & Fabre, C. (2003). Marier apprentissage endogène et ressources exogènes dans un analyseur syntaxique de corpus. le cas du rattachement verbal à distance de la préposition "de". *Revue t.a.l.*, 44(3).

[Geffroy et al., 1973] Geffroy, A., Lafon, P., & Tournier, M. (1973). L'indexation minimale. plaidoyer pour une non-lemmatisation. In *Actes du Colloque sur l'Analyse des corpus linguistiques : Problèmes et méthodes de l'indexation minimale* Strasbourg.

[Gravano et al., 2003] Gravano, L., Ipeirotis, P. G., Koudas, N., & Srivastava, D. (2003). Text joins in an RDBMS for web data integration. In *Proceedings of the 12th International World Wide Web Conference* (pp. 90–101).

[Grice, 1975] Grice, H. P. (1975). Logic and conversation. In *Studies in syntax, Speech acts*, volume 3 New York : Academic Press.

[Grishman & Sundheim, 1996] Grishman, R. & Sundheim, B. (1996). Message understanding conference-6 : a brief history. In *Porceedings of the 16th conference on Computational Linguistics* Morristown, NJ, USA : Association for Computational Linguistics.

[Guillemin-Lanne & Six, 2006] Guillemin-Lanne, S. & Six, A. (2006). La normalisation : nouveau challenge en extraction d'information. In *Actes de VSST 2006* Toulouse.

[Haddad, 2002] Haddad, H. (2002). *Extraction et impact des connaissances sur les performances des systèmes de recherche d'information.* Thèse de Doctorat, Université Joseph Fourier - Grenoble 1, Grenoble.

[Halliday, 1967a] Halliday, M. A. (1967a). Notes on Transitivity and Theme in English. part 1. *Journal of Linguistics*, 1(3), 37–81.

[Halliday, 1967b] Halliday, M. A. (1967b). Notes on Transitivity and Theme in English. part 2. *Journal of Linguistics*, 1(3), 199–244.

[Halliday, 1968] Halliday, M. A. (1968). Notes on Transitivity and Theme in English. part 3. *Journal of Linguistics*, 2(4), 179–215.

[Halliday & Hasan, 1976] Halliday, M. A. & Hasan, R. (1976). *Cohesion in English.* Longman.

[Hanisch et al., 2005] Hanisch, D., Fundel, K., Mevissen, H., Zimmer, R., & Fluck, J. (2005). ProMiner : organism-specific protein name detection using approximate string matching. In *Proceedings of BioCreative : Critical Assesment for Information Extraction in Biology.*

[Hascoët & Beaudouin-Lafon, 2001] Hascoët, M. & Beaudouin-Lafon, M. (2001). Visualisation interactive d'information. *I3 : Information, Interaction, Intelligence*, 1(1), 77–108.

[Hearst, 2008a] Hearst, M. (2008a). Flamenco home. `http://flamenco.berkeley.edu/` [page consultée le 12 janvier 2011].

[Hearst, 2008b] Hearst, M. (2008b). UIs for faceted navigation : Recent advances and remaining open problems. In *Workshop on Computer Interaction and Information Retrieval* Redmond.

[Heitz, 2006] Heitz, T. (2006). Modélisation du prétraitement des textes. In *Actes de JADT 2006 : Journées internationales d'Analyse statistique des Données Textuelles*, volume 1 (pp. 499–506).

[Hernandez, 2005] Hernandez, N. (2005). *Ontologies de domaine pour la modélisation du contexte en recherche d'information.* Thèse de Doctorat, Université de Toulouse 3, Toulouse.

[Hirschman et al., 2005] Hirschman, L., Colosimo, M., Morgan, A., & Yeh, A. (2005). Overview of BioCreAtIvE task 1B : normalized gene lists. *BMC Bioinformatics*, 6(Suppl 1).

[Hooper & Paice, 2005] Hooper, R. & Paice, C. (2005). The Lancaster Stemming Algorithm. `http://www.comp.lancs.ac.uk/computing/research/stemming/index.htm` [page consultée le 5 décembre 2010].

[Howarth, 1998] Howarth, P. (1998). International organization for standardization. In A. Cowie (Ed.), *Phraseology : Theory, Analysis, and Applications* (pp. 161–186). Oxford : Clarendon Press.

[Hudon, 1994] Hudon, M. (1994). *Le thésaurus : conception, élaboration, gestion.* Montréal : ASTED.

[Hume, 1963] Hume, D. (1963). *An enquiry concerning human understanding.* New York, NY : Washington Square Press.

[Hutchins, 1978] Hutchins, W. J. (1978). The concept of "aboutness" in subject indexing. In *Aslib Proceedings* (pp. 172–181).

[Ide & Véronis, 1995] Ide, N. & Véronis, J. (1995). *Text Encoding Initiative : Background and Context.* Dordrecht : Kluwer Academic Publishers.

[IEA, 2010] IEA (2010). *What is Ergonomics.* International Ergonomics Association. `http://www.iea.cc/01_what/What%20is%20Ergonomics.html`. [page consultée le 18 janvier 2011].

[ISO, 1999] ISO, Ed. (1999). *ISO 12620 :1999 Aides informatiques en terminologie - Catégories de données.* International Organization for Standardization.

[ISO, 2000] ISO, Ed. (2000). *ISO/IEC 13250 :2000 Topic Maps*. International Organization for Standardization.

[Jacquemin, 2000] Jacquemin, C. (2000). Traitement automatique des langues pour la recherche d'information. *Revue TAL*, 41(2).

[Jacquemin, 2001] Jacquemin, C. (2001). *Spotting and Discovering Terms through Natural Language Processing*. MIT Press.

[Jacquemin & Zweigenbaum, 2000] Jacquemin, C. & Zweigenbaum, P. (2000). Traitement automatique des langues pour l'accès au contenu des documents. In J. Le Maitre, J. Charlet, & C. Garbay (Eds.), *Le document multimédia en sciences du traitement de l'information* (pp. 71–110). Toulouse : Cépaduès Editions.

[Jakobson, 1963] Jakobson, R. (1963). *Essais de linguistique générale*, volume 1. Paris : Minuit.

[Jaro, 1976] Jaro, M. A. (1976). *Unimatch : A record linkage system : User's manual*. Technical report, U.S. Bureau of the Census, Washington.

[Jdey, 2010] Jdey, A. (2010). Bonne nouvelle : Kartoo ferme ses portes ! Demain la veille. `http://www.demainlaveille.fr/2010/01/25/bonne-nouvelle-kartoo-ferme-ses-portes/` [page consultée le 4 février 2011].

[Jijkoun et al., 2008] Jijkoun, V., Khalid, M. A., Marx, M., & de Rijke, M. (2008). Named entity normalization in user generated content. In *Actes de SIGIR 2008 - Workshop on Analytics for Noisy Unstructured Text Data*.

[Juignet, 2008] Juignet, P. (2008). L'objet de la connaissance scientifique. `http://www.philosciences.com/Articles/Gobjet.html` [page consultée le 10 janvier 2011].

[Kant, 1787] Kant, E. (1787). *Kritik der reinen Vernunft*, volume 3 - Werke, Koenigliche Preussische Akademie der Wissenschaften. Berlin : Reimer.

[Karkaletsis et al., 1999] Karkaletsis, V., Spyropoulos, C., & Petasis, G. (1999). Named entity recognition from greek texts : the GIE project. In S. Tzafestas (Ed.), *Advances in Intelligent Systems : Concepts, Tools and Applications* (pp. 131–142). Kluwer Academic Publishers.

[Kerbrat-Orecchioni, 1980] Kerbrat-Orecchioni, C. (1980). *L'énonciation - de la subjectivité dans le langage*. Paris : Armand Colin.

[Khalid et al., 2008] Khalid, M. A., Jijkoun, V., & de Rijke, M. (2008). The impact of named entity normalization on information retrieval for question answering. *LNCS*, 4956, 705–710.

[Kleiber, 1981] Kleiber, G. (1981). *Problèmes de référence. Descriptions définies et noms propres*. Paris : Klincksieck.

[Kleiber, 2004] Kleiber, G. (2004). Peut-on sauver un sens de dénomination pour les noms propres ? *Functions of language*, 11(1), 115–145.

[Korenius et al., 2004] Korenius, T., Laurikkala, J., Jarvelin, K., & Juhola, M. (2004). Stemming and lemmatisation in the clustering of finnish text documents. In *Proceedings of ACM conference on Information and knowledge management* (pp. 625–633).

[Kripke, 1972] Kripke, S. (1972). *La logique des noms propres (Naming and Necessity).* Paris : Minuit.

[Labbé & Labbé, 2006] Labbé, C. & Labbé, D. (2006). La diachronie dans le discours politique : Le général de gaulle. *Actes des Nouvelles journées de l'ERLA*.

[Labbé, 1990a] Labbé, D. (1990a). *Le vocabulaire de François Mitterrand.* Paris : Presses de la Fondation nationale des sciences politiques.

[Labbé, 1990b] Labbé, D. (1990b). Normes de saisie et de dépouillement des textes politiques. *Cahiers du CERAT*, 7, 1–135.

[Labbé, 2006] Labbé, D. (2006). Analyse des données textuelles et statistique lexicale. In *Actes de JADT 2006 : Journées internationales d'Analyse statistique des Données Textuelles.*

[Lafon et al., 1985] Lafon, P., Lefèvre, J., Salem, A., & Tournier, M. (1985). *Le Machinal. Principes d'enregistrement informatique des textes.* Paris : Klincksieck.

[Lafon & Salem, 1983] Lafon, P. & Salem, A. (1983). L'inventaire des segments répétés d'un texte. *Mots*, (6), 161–177.

[Lahlou, 1996] Lahlou, S. (1996). Le projet Scriptorium : prospecter dans la mémoire sociale de l'entreprise. In *Actes du Séminaire de Sciences Cognitives* (pp. 104–106). Université de Technologie de Compiègne.

[Le Moigne, 1995] Le Moigne, J. (1995). *Les épistémologies constructivistes.* Presses Universitaires de France.

[Lebarbé, 2007] Lebarbé, T. (2007). LexTract : Extraction semi-automatique de termes à portée juridique. *Actes des Journées de la Linguistique de Corpus*, (pp. 197–205).

[Lebarbé, 2009] Lebarbé, T. (2009). Du corpus littéraire au corpus linguistique : dématérialisation, restructuration, lectures rhizomatiques et analyses linguistiques des manuscrits. *Corpus*, (n°8), 221–239.

[Lebarbé, 2010] Lebarbé, T. (2010). *Fonctions interdisciplinaires intrinsèques et extrinsèques du traitement automatique des langues.* Habilitation à diriger des recherches, Université Stendhal - Grenoble 3, Grenoble.

[Lebarbé & Breese, 2001] Lebarbé, T. & Breese, P. (2001). Computer assisted TradeMark infringement evaluation (CATMInE). In *3rd American-French Conference on Technology and Legal Practice* Syracuse, USA.

[Lebarbé & Meynard, 2009] Lebarbé, T. & Meynard, C. (2009). Nouvelles pratiques éditoriales, nouvelles lectures : les enjeux de l'édition électronique de manuscrits littéraires. *Mémoire du livre*, (1).

[Lebart & Salem, 1994] Lebart, L. & Salem, A. (1994). *Statistique textuelle*. Paris : Dunod.

[Leplat & Hoc, 1983] Leplat, J. & Hoc, J. (1983). Tâche et activité dans l'analyse psychologique des situations. *Cahiers de psychologie cognitive*, 3(1), 49–63.

[Levenshtein, 1966] Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10, 707–710.

[Ligozat, 2006] Ligozat, A. (2006). *Exploitation et fusion de connaissances locales pour la recherche d'informations précises*. Thèse de Doctorat, Université Paris 11 - Orsay, Paris.

[Machado, 1917] Machado, A. (1917). Chant XXIX. In *Proverbios y Cantarès. Campos de Castilla*. DeuxiÃ¨me edition.

[Magdy et al., 2007] Magdy, W., Darwish, K., Emam, O., & Hassan, H. (2007). Arabic cross-document person name normalization. In *Proceedings of CASL Workshop '07 : Computational Approaches to Semitic Languages* (pp. 25–32).

[Maisonnasse, 2008] Maisonnasse, L. (2008). *Les supports de vocabulaires pour les systèmes de recherche d'information orientés précision : application aux graphes pour la recherche d'information médicale*. Thèse de Doctorat, Université Joseph Fourier - Grenoble 1, Grenoble.

[Maniez, 2001] Maniez, F. (2001). La traduction du nom adjectival en anglais médical. *Meta : journal des traducteurs / Meta : Translators' Journal*, 46(1), 56–67.

[Mann, 1987] Mann, T. (1987). *A Guide to Library Research Methods*. Oxford University Press, USA, First edition.

[Manning et al., 2008] Manning, C., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

[Mel'cuk & Wanner, 1996] Mel'cuk, I. & Wanner, L. (1996). Lexical functions and lexical inheritance for emotion lexemes in german. In L. Wanner (Ed.), *Lexical Functions in Lexicography and Natural Language Processing* (pp. 209–278). Amsterdam/Philadelphie : Benjamins.

[Merzeau, 2010] Merzeau, L. (2010). L'intelligence de l'usager. In *L'usager numérique - séminaire INRIA* (pp. 9–37). Anglet : ADBS Editions.

[Mill, 1843] Mill, J. (1843). *Système de logique déductive et inductive*. Londres.

[Millon, 2011] Millon, C. (2011). *Acquisition automatique de relations lexicales désambiguïsées à partir du Web*. Thèse de Doctorat, Université de Bretagne Sud, Lorient.

[Mizzaro, 1997] Mizzaro, S. (1997). Relevance : the whole history. *Journal of the American Society for Information Science*, 48(9), 810–832.

[Mizzaro, 1998] Mizzaro, S. (1998). How many relevances in information retrieval ? *Interacting With Computers*, 10(3), 305–322.

[Monge & Elkan, 1996] Monge, A. E. & Elkan, C. P. (1996). The field matching problem : Algorithms and applications. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (pp. 267–270).

[Moreau & Claveau, 2006] Moreau, F. & Claveau, V. (2006). Extension de requêtes par relations morphologiques acquises automatiquement. *I3 : Information, Interaction, Intelligence*, 6(2), 31–50.

[Morris, 1938] Morris, C. W. (1938). *Foundations of the theory of signs*. Chicago : The University of Chicago Press.

[Muller, 1967] Muller, C. (1967). *Etude de statistique lexicale. Le vocabulaire du théâtre de P. Corneille*. Paris : Larousse.

[Nielsen, 1993] Nielsen, J. (1993). *Usability engineering*. Morgan Kaufmann.

[Otman, 1995] Otman, G. (1995). *Les représentations sémantiques en terminologie*. Thèse de Doctorat, Université Paris 6, Paris.

[Paice, 1990] Paice, C. (1990). Another stemmer. In *SIGIR Forum*, volume 24 (pp. 56–61).

[Patry & Langlais, 2005] Patry, A. & Langlais, P. (2005). Corpus-based terminology extraction. In *Proceedings of the 7th International Conference on Terminology and Knowledge Engineering* (pp. 313–321). Copenhague, Danemark.

[Pédauque, 2003] Pédauque, R. T. (2003). *Document : forme, signe et relation, les reformulations du numérique*. Réseau Thématique Pluridsciplinaire "Document et contenu : création, indexation, navigation" (RTP-33), Département STIC du CNRS, Document de travail.

[Perfetti et al., 1999] Perfetti, C. A., Rouet, J. F., & Britt, M. A. (1999). Towards a theory of documents representation. In H. van Oostendorp & S. Goldman (Eds.), *The construction of mental representations during reading* (pp. 99–122). Mahwah, NJ : Erlbaum.

[Pernin, 2007] Pernin, J. (2007). Mieux articuler activités pour l'apprentissage, artefacts logiciels et connaissances : vers un modèle d'ingénierie centré sur le concept de scénario. In *Environnements informatisés et ressources numériques pour l'apprentissage : conception et usages, regards croisés* (pp. 161–190). Paris : Hermès.

[Péry-Woodley, 2000] Péry-Woodley, M. (2000). *Une pragmatique à fleur de texte : approche en corpus de l'organisation textuelle*. Mémoire d'Habilitation à diriger des recherches, Université Toulouse 2 - Le Mirail, Toulouse.

[Philips, 1990] Philips, L. (1990). Hanging on the metaphone. *Computer Language Magazine*, 7(12), 39–44.

[Piaget, 1967] Piaget, J. (1967). *Biologie et connaissance*. Paris : Gallimard.

[Pichon & Sébillot, 1999] Pichon, R. & Sébillot, P. (1999). Différencier les sens des mots à l'aide du thème et du contexte de leurs occurrences : une expérience. In *Actes de TALN 1999* (pp. 279–288).

[Pirolli & Card, 1999] Pirolli, P. & Card, S. (1999). Information foraging. *Psychological Review*, (106), 643–675.

[Poibeau, 2001] Poibeau, T. (2001). Deconstructing Harry, une évaluation des systèmes de repérage d'entités nommées. *Revue de la société d'électronique, d'électricité et de traitement de l'information*.

[Poibeau, 2003] Poibeau, T. (2003). *Extraction automatique d'information : du texte brut au web sémantique*. Paris : Hermès.

[Porter, 1980] Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3), 130–137.

[Porter, 2006] Porter, M. F. (2006). Porter stemming algorithm. `http://tartarus.org/~martin/PorterStemmer/`. [page consultée le 10 décembre 2010].

[Proctor, 2001] Proctor, T. Y. (2001). *The effects of time pressure and accountability on hypothesis generation and information search strategies : An experimental study of internal revenue agents*. Thèse de Doctorat, University of Memphis, Memphis.

[Rastier, 1995] Rastier, F. (1995). Le terme : entre ontologie et linguistique. *Texto !. http://www.revue-texto.net/index.php?id=568 [page consultée le 3 novembre 2010].*

[Rastier, 2001] Rastier, F. (2001). *Arts et sciences du texte*. Paris : Presses Universitaires de France.

[Renals et al., 1999] Renals, S., Gotoh, Y., Gaizauskas, R., & Stevenson, M. (1999). Baseline IE-NE experiments using the SPRACH/LaSIE system. In *Proceedings of DARPA Broadcast News Workshop* (pp. 47–50). Virginia.

[Rey-Debove, 1979] Rey-Debove, J. (1979). *Sémiotique*. Paris : Presses Universitaires de France.

[Roselli, 2010] Roselli, M. (2010). Formes de réception et d'appropriation des ressources numériques en milieu étudiant. Enquête ethnographique en bibliothèque universitaire. *TIC et Société*, 4(1 : Interactivité et lien social).

[Rouet, 2000] Rouet, J. F. (2000). *Les activités documentaires complexes. Aspects cognitifs et développementaux*. Mémoire d'Habilitation à diriger des recherches, Université de Poitiers, Poitiers.

[Rouet et al., 1996] Rouet, J. F., Britt, M. A., Mason, R. A., & Perfetti, C. A. (1996). Using multiple sources of evidence to reason about history. *Journal of Educational Psychology*, 88(3), 478–493.

[Rouet & Tricot, 1995] Rouet, J. F. & Tricot, A. (1995). Recherche d'informations dans les systèmes hypertextes : des représentations de la tâche à un modèle de l'activité cognitive. *Sciences et Techniques Educatives*, 2(3), 307–331.

[Rouet & Tricot, 1996] Rouet, J. F. & Tricot, A. (1996). Task and activity models in hypertext usage. In H. van Oostendorp & S. de Mul (Eds.), *Cognitive aspects of electronic text processing*, volume 58 of *Advances in Discourse Processes* (pp. 239–264). Norwood, NJ : Ablex Publishing.

[Rouet & Tricot, 1998] Rouet, J. F. & Tricot, A. (1998). Chercher de l'information dans un hypertexte : vers un modèle des processus cognitifs. *Hypertextes et hypermédias*, (hors-série), 57–74.

[Roy, 2007] Roy, T. (2007). *Visualisations interactives pour l'aide personnalisée à l'interprétation d'ensembles documentaires*. Thèse de Doctorat, Université de Caen Basse-Normandie.

[Russell, 1918] Russell, R. C. (1918). Index. `http://www.freepatentsonline.com/1261167.pdf`.

[Saracevic, 1970] Saracevic, T. (1970). The concept of "relevance" in information science : A historical review. In T. Saracevic (Ed.), *Introduction to information science* (pp. 111–151). New York : R.R. Bowker.

[Schmid, 1994] Schmid, H. (1994). Probabilistic Part-Of-Speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing* Manchester, UK.

[Schnotz, 2001] Schnotz, W. (2001). Sign systems, technologies, and the acquisition of knowledge. In J. Rouet, J. Levonen, & A. Biardeau (Eds.), *Multimedia learning. Cognitive and instructional issues* (pp. 9–29). Amsterdam : Elsevier.

[Searle, 1969] Searle, J. (1969). *Speech Acts : An Essay in the Philosophy of Language*. New York, NY : Cambridge University Press.

[Sekine & Eriguchi, 2000] Sekine, S. & Eriguchi, Y. (2000). Japanese named entity extraction evaluation - analysis of results. In *Proceedings of Coling'2000 : Computational Linguistics* (pp. 25–30). Saarbrücken, Germany.

[SELF, 2008] SELF (2008). *Définitions*. Société d'Ergonomie de Langue Française. `http://www.ergonomie-self.org/heading/heading27163.html`. [page consultée le 18 janvier 2011].

[Shneiderman, 1996] Shneiderman, B. (1996). The eyes have it : a task by data type taxonomy for information visualization. In *Proceedings of Visual Languages* (pp. 336–343).

[Simon, 1973] Simon, H. A. (1973). The structure of ill-structured problems. *Artificial Intelligence*, (4), 181–202.

[Simon, 1981] Simon, H. A. (1981). *The sciences of the artificial*. Cambridge : The MIT Press.

[Sinclair, 1991] Sinclair, J. M. (1991). *Corpus, Concordance, Collocation*. Oxford : Oxford University Press.

[Sinclair, 2004] Sinclair, J. M. (2004). *Trust the text : language, corpus and discourse*. London : Routledge.

[Sinclair et al., 2004] Sinclair, J. M., Jones, S., & Daley, R. (2004). *English collocation studies : the OSTI report*. Continuum International Publishing Group.

[Smith & Waterman, 1981] Smith, T. F. & Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1), 195–197.

[Sperber & Wilson, 1986] Sperber, D. & Wilson, D. (1986). *Relevance : Communication and Cognition*. Oxford and Cambridge : Blackwell and Harvard University Press.

[Sperber & Wilson, 1995] Sperber, D. & Wilson, D. (1995). *Postface to the second edition of Relevance : Communication and Cognition*. Oxford : Blackwell.

[Strawson, 1970] Strawson, P. (1970). *Meaning and Truth : An Inaugural Lecture delivered before the University of Oxford on 5 November, 1969*. Oxford : Clarendon Press : Oxford University Press.

[Stubbs, 2001] Stubbs, M. (2001). *Words and phrases : corpus studies of lexical semantics*. Oxford : Blackwell.

[Tannier, 2006] Tannier, X. (2006). *Extraction et recherche d'information en langage naturel dans les documents semi-structurés*. informatique, Ecole Nationale Supérieure des Mines Saint-Etienne, Saint-Etienne.

[Tesnière, 1959] Tesnière, L. (1959). *Eléments de syntaxe structurale*. Paris : Klincksieck.

[Treisman, 1986] Treisman, A. (1986). Features and objects in visual processing. *Scientific American*, 254(11), 114–125.

[Tricot, 1993] Tricot, A. (1993). Ergonomie cognitive des systèmes hypermédia. In *Actes du Colloque de prospective Recherches pour l'Ergonomie* (pp. 115–122). Toulouse.

[Tricot, 2001] Tricot, A. (2001). Interpréter les liens entre utilisabilité et utilité des documents électroniques. In M. Mojahid & J. Virbel (Eds.), *Les documents électroniques, méthodes, démarches et techniques cognitives*. Paris : Europia.

[Tricot, 2003] Tricot, A. (2003). *Apprentissage et recherche d'information avec des documents électroniques*. Mémoire d'Habilitation à diriger des recherches, Université de Toulouse de Mirail, Toulouse.

[Tricot, 2006] Tricot, C. (2006). *Cartographie sémantique - Des connaissances à la carte*. Thèse de Doctorat, Université de Savoie, Chambéry.

[Ukkonen, 1992] Ukkonen, E. (1992). Approximate string matching with q-grams and maximal matches. *Theoretical Computer Science*, 92(1), 191–211.

[Valette & Slodzian, 2008] Valette, M. & Slodzian, M. (2008). Sémantique des textes et recherche d'information. *Extraction d'information : l'apport de la linguistique, Revue Française de Linguistique Appliquée*, 13(1), 119–133.

[Van der Henst, 2002] Van der Henst, J. (2002). La perspective pragmatique dans l'étude du raisonnement et de la rationalité. *L'Année Psychologique*, 102, 65–108.

[van Rijsbergen, 1979] van Rijsbergen, C. J. (1979). *Information Retrieval*. London : Butterworths.

[Vergne, 2003] Vergne, J. (2003). Un outil d'extraction terminologique endogène et multilingue. In *Actes de TALN 2003*.

[Vergne, 2004] Vergne, J. (2004). Découverte locale des mots vides dans des corpus bruts de langues inconnues, sans aucune ressource. In *Actes de JADT 2004 : Journées internationales d'Analyse statistique des Données Textuelles*, volume 2 (pp. 1158–1164).

[Véronis, 2006] Véronis, J. (2006). Etude comparative de six moteurs de recherche. `http://blog.veronis.fr/2006/02/moteurs-et-le-gagnant-est.html` [page consultée le 23 novembre 2010].

[Vickery, 1959a] Vickery, B. C. (1959a). The structure of information retrieval systems. In *Proceedings of the International Conference on Scientific Information*, volume 2 (pp. 1275–1290). Washington.

[Vickery, 1959b] Vickery, B. C. (1959b). Subject analysis for information retrieval. In *Proceedings of the International Conference on Scientific Information*, volume 2 (pp. 855–865). Washington.

[von Glasersfeld, 1994] von Glasersfeld, E. (1994). Pourquoi le constructivisme doit-il être radical ? *Revue des Sciences de l'Education*, 20(1), 21–27.

[Wacholder et al., 1997] Wacholder, N., Ravin, Y., & Choi, M. (1997). Disambiguation of proper names in text. In *Proceedings of the 5th Conference on Applied Natural Language Processing* (pp. 202–208). Washington.

[Waterman et al., 1976] Waterman, M. S., Smith, T. F., & Beyer, W. A. (1976). Some biological sequence metrics. *Advances in Mathematics*, 20(4), 367–387.

[Williams, 2003] Williams, G. (2003). Les collocations et l'école contextualiste britannique. *Les mots et leur(s) combinatoire(s) : Analyse et traitement des collocations*, (1), 33–44.

[Wilson & Sperber, 2004] Wilson, D. & Sperber, D. (2004). Relevance theory. In L. Horn & G. Ward (Eds.), *The Handbook of Pragmatics* (pp. 607–632). Oxford : Blackwell.

[Wineburg, 1991] Wineburg, S. (1991). Historical problem solving : A study of the cognitive processes used in the evaluation of documentary and pictorial evidence. *Journal of Educational Psychology*, (83), 73–87.

[Wineburg, 1994] Wineburg, S. (1994). The cognitive representation of historical texts. In G. Leinhardt, I. Beck, & C. Stainton (Eds.), *Teaching and learning in history* (pp. 85–135). Hillsdale, NJ : Erlbaum.

[Wüster, 1968] Wüster, E. (1968). *Dictionnaire multilingue de la machine-outil. Notions fondamentales, définies et illustrées, présentées dans l'ordre systématique et l'ordre alphabétique. Volume de base anglais-français = The Machine Tool. An Interlingual Dictionary of Basic Concepts comprising an Alphabetical Dictionary and a Classified Vocabulary with Definitions and Illustrations. English-Franch Master Volume.* London : Technical Press.

[Zaghouani, 2009] Zaghouani, W. (2009). *Le repérage automatique des entités nommées dans la langue arabe : vers la création d'un système à base de règles.* Mémoire de master, Université de Montréal, Montréal.

[Zhou et al., 2007] Zhou, W., Yu, C., Smalheiser, N., Torvik, V., & Hong, J. (2007). Knowledge-intensive conceptual retrieval and passage extraction of biomedical literature. In *Proceedings of SIGIR'07 : Special Interest Group on Information Retrieval* (pp. 655–662).

[Zipf, 1949] Zipf, G. K. (1949). *Human behaviour and the principle of least effort - An introduction to human ecology.* Addison-Wesley Press.

Annexe A

Liste des amorces classées par type et par niveau

Sommaire

A.1 Amorces d'organismes

Amorce	Niveau
UNIV	1
CEA	1
IFREMER	1
ENS	1
INRA	1
INSERM	1
CNRS	1
INRIA	1
AFSSA	1
CLINIC	1
INFIRMARY	1
AGENCY	1
ASSO	1
SOCIETY	1
INSTITUTION	1
ACAD	1
NIH	1
AMMS	1
CAS	1
RAS	1
CAAS	1
MIT	1
KAIST	1
NRC	1
CNR	1
OFFICE	2
UMR	2
FAC	2
FDN	

Amorce	Niveau
HOSPITAL	3
SCHOOL	3
COLL	3
CTR	3
ECOLE	3
INST	3
SERVICE	4
DEPT	5
DIV	6
PROG	6
BRANCH	7
TEAM	8
PROJECT	8
GROUP	8
UNIT	8
SECTION	8
LAB	9

TABLE A.1 – Liste hiérarchisée des amorces d'organismes utilisées pour la normalisation des entités nommées

A.2 Amorces d'entreprises

Amorce	Niveau
AG	1
CO	1
CIE	1
CORP	1
HOLDING	1
INC	1
LT	1
LTD	1
LLC	1
PLC	1
PTY	1
SA	1
SEC	1
SL	1
SAS	1
KK	1
NV	1
BV	1
OYJ	1
OY	1
GMBH	1
KG	1

Amorce	Niveau
AS	1
LL	1
AOOT	1
OOO	1
SPA	1
AB	1
SRL	1
LTDA	1
SARL	1
SNC	1
SCA	1
SCS	1
SOCIETY	1
FDN	2
SERVICE	4
DEPT	5
LAB	9

TABLE A.2 – Liste hiérarchisée des amorces d'entreprises utilisées pour la normalisation des entités nommées

A.3 Amorces d'adresses

Amorce	Niveau
PARK	sans objet
PARKWAY	sans objet
WALK	sans objet
AVENUE	sans objet
STREET	sans objet
HIGHWAY	sans objet
ROAD	sans objet
ALLEY	sans objet
DRIVE	sans objet
FAIRWAY	sans objet
GATE	sans objet
GROVE	sans objet
LANE	sans objet
PATHWAY	sans objet
PLACE	sans objet
TERRACE	sans objet
TRAIL	sans objet
COVE	sans objet
SQUARE	sans objet

Amorce	Niveau
RUE	sans objet
ALLEE	sans objet
IMPASSE	sans objet
BOULEVARD	sans objet
VIALE	sans objet
CITTADELLA	sans objet
BUILDING	sans objet
APARTMENT	sans objet
ROUTE	sans objet
CEDEX	sans objet
CHEMIN	sans objet
POBOX	sans objet
CIUDAD	sans objet
CTVILLE	sans objet
BP	sans objet
STRASSE	sans objet
BERGSTRASSE	sans objet

TABLE A.3 – Liste hiérarchisée des amorces d'adresses utilisées pour la normalisation des entités nommées