Proposition de chaînage des connaissances historiques et patrimoniales Approche multi-échelles et multi-critères de corpus textuels

Matthieu Quantin

UNIVERSITE BRETAGNE SCIENCES LOIRE POUR L'INGENIEUR

CENTRALE NANTES

THESE DE DOCTORAT DE

L'ÉCOLE CENTRALE DE NANTES
COMUE UNIVERSITE BRETAGNE LOIRE

ECOLE DOCTORALE N° 602
Sciences pour l'Ingénieur
Spécialité : Génie mécanique, productique, transport, section CNU 60

Par
Matthieu QUANTIN

Proposition de chaînage des connaissances historiques et patrimoniales
Approche multi-échelles et multi-critères de corpus textuels

Thèse présentée et soutenue à Nantes, le 27/06/2018
Unité de recherche : Laboratoire des sciences du numérique de Nantes (LS2N)

Rapporteurs avant soutenance :

Andreas RIEL	Maître de conférences, INP Grenoble
Livio DE LUCA	Directeur de recherche, CNRS Marseille

Composition du Jury :

Président :	Bruno BACHIMONT	Professeur des universités, UTC
Examinateurs :	Florent LAROCHE	Maître de conférences, Ecole Centrale Nantes
	Jean-Louis KEROUANTON	Maître de conférences, Université de Nantes
	Muriel GUEDJ	Maître de conférences, Université de Montpellier
Dir. de thèse :	Alain BERNARD	Professeur des universités, Ecole Centrale de Nantes

Proposition de chaînage des connaissances historiques et patrimoniales

décembre 2018

Remerciements

Je remercie mon directeur de thèse, Alain Bernard, Professeur des Universités à l'École Centrale de Nantes, ainsi que mes co-encadrants, Florent Laroche, Maître de conférences en génie industriel à l'École Centrale de Nantes et Jean-Louis Kerouanton, Maître de conférences en histoire des techniques à l'Université de Nantes. Parmi eux, je tiens à remercier spécialement Florent Laroche qui m'aura été d'un soutien inconditionnel et présent au quotidien, qui m'aura systématiquement associé à ses projets de recherche, merci pour les riches échanges autant professionnels que personnels.
Je remercie également les membres du jury, présidé par le professeur Bruno Bachimont, je remercie Muriel Guedj, maître de conférence, et le professeur Vincent Cheutet, je remercie tout particulièrement Andreas Riel, Maître de conférence et Livio De Luca Directeur de recherche, pour l'intérêt qu'ils ont manifesté et pour avoir accepté d'évaluer mes travaux, c'est un grand honneur.
Je tiens aussi à mentionner les membres de mon comité de suivi de thèse qui ont élargi les horizons de mes travaux : Florence Hachez Leroy et Benoît Furet.

Dans les coulisses de cette thèse, trois personnes ont œuvré au bon déroulé de ces trois années de part leurs qualités humaines et leur implication de chercheur : Benjamin Hervy, Pierre Teissier et Loïc Jeanson.
Je remercie Loïc Jeanson, doctorant talentueux et à l'écoute, qui a su apporter une belle énergie aux recherches menées.
Je remercie Pierre Teissier, maître de conférence en Histoire des techniques, chercheur exemplaire et collègue idéal. Je le remercie pour la richesse de ses interrogations et conseils, pour la matière scientifique qu'il m'a apporté et le temps qu'il a dédié à notre collaboration.
Je remercie tout particulièrement **Benjamin Hervy**, jeune docteur centralien, qui a été mon plus proche collaborateur de travail, co-developpeur initial de *Haruspex*, chercheur motivant, curieux, généreux et sérieux, impliqué sans réserve dans l'avancé de ces travaux de thèse, dans chacune des expériences et dans chacun des résultats ainsi que dans l'écriture de publications. Je remercie aussi Benjamin pour son soutien, son écoute et sa compréhension des situations de thésard.

Je souhaite ici remercier mes collègues de l'École Centrale et du laboratoire (IRCCyN - LS2N) :

— Les membres de l'équipe IS3P à laquelle j'ai été rattaché
— Émilie et Virginie pour leur gentillesse et leur aide sur les questions administratives
— Les membres du LS2N coté LINA et notamment ceux de l'équipe TALN, Chantal Enguehard tout particulièrement pour nous avoir livré les secrets de *ANA* et pour ses conseils avisés, mais aussi Pascale Kuntz pour les orientations sur le traitement de données.
— Les membres du conseil de laboratoire pour m'avoir accepté parmi eux et intéressé à la gestion d'un laboratoire en mutation. Je remercie tout particulièrement Michel Malabre directeur de recherche et directeur du laboratoire IRCCyN pour sa bienveillance.
— les collègues du 4e étage du laboratoire : Touskee (Vincent), Victor, Clément et Loïc, ainsi que Polo et les deux Marie du laboratoire voisin, acteurs du quotidien de cette thèse. Je leur dois les plus riches heures de coinche au café du Théâtre Universitaire, lueur du quotidien de thésard.

Je souhaite aussi remercier mes collègues issus d'autres cercles de recherches :

— Les membres du consortium 3D de Huma-Num et particulièrement ceux de l'équipe *Archivage* : Bruno Dutailly, Sylvie Eusèbe et tout spécialement Sarah Tournon-Valiente et Valentin Grimaud.
— Les membres du Centre François Viète pour leur accueil et/ou la richesse des questionnements d'historien des techniques qui ont alimenté cette thèse : Jenny Boucard, Stéphane Tirard, Anaïs Durand, Martine Acerra, Scott Walter et tout particulièrement Anaël Marrec.
— Stéphane de la MSH pour sa bonne humeur et pour avoir accueilli nos discussions avec Benjamin dans leur bureau.
— Les littéraires du laboratoire l'AMo, particulièrement Françoise Rubellin pour son énergie et sa curiosité.
— Les membres du projet de recherche ReSeed (ANR) qui m'ont intégré à leurs problématiques et particulièrement Michel Cotte pour son immense savoir et sa passion pour le patrimoine industriel.

Enfin, je termine ces remerciements par ceux qui me côtoient me supportent et me soutiennent au quotidien : avant, pendant et après la thèse. Je pense à mes amis, à Lauren, Ghislain, Simon, Julien et ma famille et surtout à Alice.

Table des matières

Liste des tableaux

Table des figures

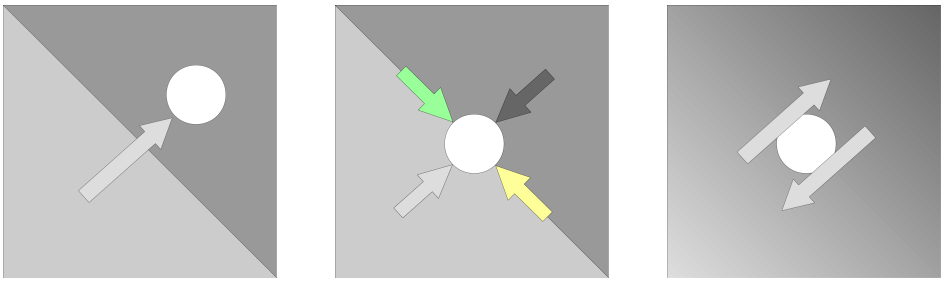
Introduction

Dans la grande famille des *intelligences artificielles* et de la *gestion des connaissances*, cette thèse présente une méthode de chaînage des connaissances historiques et patrimoniales. Elle est basée sur les contenus textuels. Ces contenus peuvent être utilisés dans plusieurs finalités : analyse historique et documentation patrimoniale. Un des moyens consiste à calculer des proximités entre les textes du corpus. Ces proximités peuvent être de nature différente : espace, temps, thématique, etc. (multi-critère). De plus, les proximités thématiques par exemple sont “zoomables” (multi-échelle).

Contenus Wikipedia. Certains contenus de cette thèse sont proches de contenus Wikipédia. Cela concerne figures et textes, essentiellement dans le Chapitre 1 : État de l’art. Ces contenus ont été produits dans le cadre de cette thèse et publiés sur Wikipédia par moi-même. En effet, j’estime qu’un état de l’art est une contribution utile à la communauté et que le format « wiki » est davantage approprié à la diffusion que la thèse.

Contexte de la thèse

Cette thèse a une composante minoritaire en Histoire des sciences et des techniques, épistémologie, patrimoine avec le laboratoire Centre François Viète (CFV) EA_1161 ; et une composante majoritaire en science pour l’ingénieur, gestion des connaissances avec le Laboratoire des Sciences du Numérique de Nantes (LS2N) UMR_6004. L’intention de ce travail est pluridisciplinaire (cf. figure 1b) : le dialogue entre plusieurs disciplines sur un objet d’étude et l’adéquation entre les moyens ; une composante interdisciplinaire (cf. figure 1c) émerge finalement puisque les frontières entre disciplines s’estompent et l’enrichissement est mutuel, au-delà de l’objet d’étude contingent.



(a) Transdisciplinarité (b) Pluridisciplinarité (c) Interdisciplinarité

FIGURE 1 – Schématisation de 3 types d’interactions entre disciplines.

La collaboration entre ces deux unités de recherche a donné lieu à la formation d’un groupe de recherche nommé Évolution des procédés et objets techniques (Évolution des procédés et objets techniques (EPOTEC)). À ce titre nous emmargeons au Consortium 3D de la Très grande infrastructure de recherche (TGIR) Human-Num, un groupe de recherche national sur la 3D en Sciences Humaines.

EPOTEC

Le groupe EPOTEC est fort d’une longue collaboration dont la filiation par génération de chercheurs (encadrement de thèses), est décrite en figure 2a. Les membres en Sciences pour l’ingénieur sont concernés par plusieurs objets de recherche : la muséologie, la géométrie (3D), la sémantique. Chacun exerce dans un domaine de prédilection, et de fortes interactions sont à noter (cf figure 2b) :

— Pôle géométrie : Florent Laroche, thèse en 2007 (Laroche, 2007) : étude du patrimoine en 3D via des outils de Conception Assistée par Ordinateur (CAO) pour heuristique en histoire en intégrant des connaissances mécaniques. Rétro-conception et simulation d’objets techniques anciens. Plus que la représentation, c’est la fonctionnalité de la géométrie qui importe dans ce travail. Développement du Digital Heritage Reference Model (DHRM).

— Pole muséologie : Benjamin Hervy, thèse en 2014 (Hervy, 2014) : intégration de connaissances sur un cycle de vie patrimoine dans le cadre d'un musée. Valorisation en contexte muséal (collection permanente) avec le projet Nantes1900. Encore une fois, plutôt que la géométrie (les éléments en eux-même) de la maquette, c'est l'objet de musée et les connaissances historiques dont il est porteur qui sont mises en valeur.
— Pole sémantique : Matthieu Quantin, thèse actuelle : réseau d'information depuis des corpus de texte. Analyse et documentation historique en lien avec le patrimoine.

Les réalisations autour du patrimoine. Cette thèse approfondit la dimension documentation et sémantique pour le patrimoine. Les travaux majeurs de notre équipe, listés ci-dessous, constituent le cadre d'application de cette thèse : ils participent à l'émergence du besoin ou permettent de réaliser des expériences.

— Pont transbordeur de Nantes (2014) : rétro-conception et valorisation en réalité augmentée et réalité virtuelle (3D)
— Arsenal de la Marine de Lorient (2014) : étude systémique du complexe industriel, portuaire et militaire
— Salons Mauduits (2015) : acquisition 3D, documentation et application en réalité virtuelle
— Forges de Paimpont (2015) : Rétro-conception et application en réalité virtuelle
— Curiosity (2015) : application en réalité augmentée
— Salle à manger tournante de Néron (2016) : rétro-conception et fabrication d'une maquette
— Pic du midi (2017) : Documentation et modélisation du site. Application en réalité virtuelle

Les constats. Les constats suivants sont issus des réalisations de l'équipe.

— Importance de la sémantique et du contexte. L'objet patrimonial est au cœur de données complexes
— Besoin de calcul de distance, de proximité, exploitation de données historiques (contexte) dans le musée, pour le patrimoine. Il existe un cloisonnement fort entre données historiques et patrimoine. Constat issu de l'expérience Nantes1900 (Hervy, 2014).
— Travail de l'historien, consiste à créer des liens non binaires. La majorité des cas de données numériques traitées par les historiens sont des corpus textuels de taille moyenne. Peu d'outils sont adaptés.

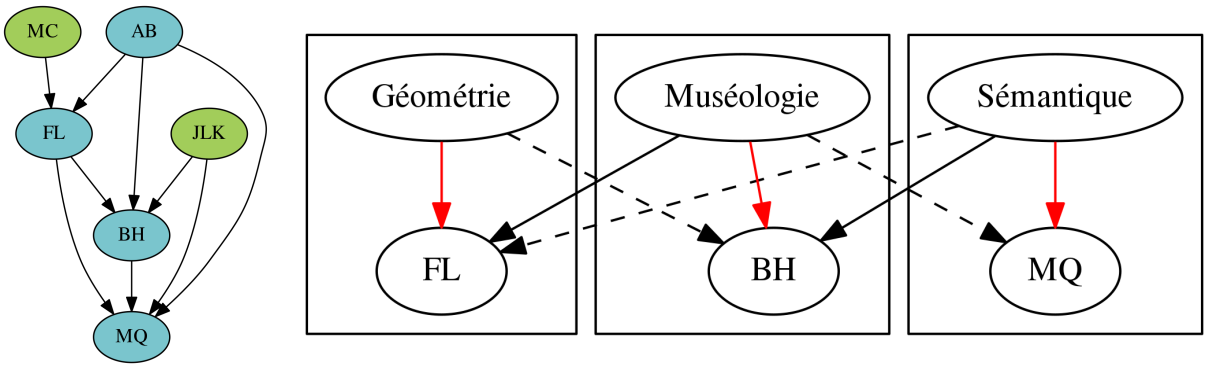
Huma-Num et le Consortium 3D

Le consortium 3D de la TGIR Huma-Num est formé de laboratoires de recherche français [1] utilisant le 3D dans un cadre de sciences humaines. Ce consortium est divisé en plusieurs groupes de travail : *Vocabulaire*, *Cahier des Charges*, *Matériel/Logiciel* et *Archivage*.

Le groupe *archivage* auquel je participe est composé des laboratoires suivants : LS2N, INRAP, LARA, Archéovision, CIREVE. Ce groupe agit comme une extension de l'équipe EPOTEC pour la problématique d'archivage perenne des modèles 3D scientifiques, fortement orienté sur la documentation du modèle et les raisonnements. Le groupe *archivage* a produit un guide de l'archivage, un modèle de données et un logiciel pour générer les archives. Les réflexions ont été menées conjointement avec le Centre Informatique National de l'Enseignement Supérieur (CINES). Ce travail, hors-cadre de cette thèse, n'est pas présenté dans ce manuscrit.

Enfin la TGIR Huma-Num propose de nombreux services pour les laboratoires, mon travail de thèse a ainsi pu bénéficier d'un accès à un supercalculateur (hébergé par l'IN2P3) pour tester le passage à l'échelle de certains développements.

1. Archéovision (Bordeaux), MAP (Marseille), CIREVE (Normandie-Caen), LARA (Nantes), ASM (Montpellier), CEPAM (Nice), EPOTEC (Nantes), MSH Val-de-Loire (Tours), INRAP (Paris), MOM (Lyon).



(a) Filiation

(b) Domaines d'étude

FIGURE 2 – En vert les sciences humaines, en bleu les sciences pour l'ingénieur. MC : Michel Cotte, AB : Alain Bernard, FL : FLorent Laroche, JLK : Jean-Louis Kerouanton, BH : Benjamin Hervy, MQ : Matthieu Quantin. Flèche rouge : domaine principal, flèche pointillée : domaine secondaire

Positionnement de l'équipe et de cette thèse

L'équipe. EPOTEC prend le parti de défendre un approche multimodale : les représentations géométriques (3D), voire fonctionnelles sont des volets d'accès aux connaissances patrimoniales. Y compris pour le patrimoine matériel, les connaissances sont structurantes. Ces connaissances s'organisent et s'analysent dans une approche épistémologique externaliste. C'est-à-dire que les données ne font sens que dans un contexte plus large : on ne peut pas étudier une machine sans s'intéresser aux conditions de travail, au contexte politique, au prix des matériaux, à la destination de la production, etc. Ainsi avant de s'intéresser à la 3D et aux rendus, notre équipe se focalise sur la gestion des connaissances pour le patrimoine 3D.

La thèse. Dans ce cadre, ma thèse est focalisée sur la gestion des connaissances et l'analyse des données historiques, la dimension 3D est une des finalités possibles et reste au second plan de ce travail.

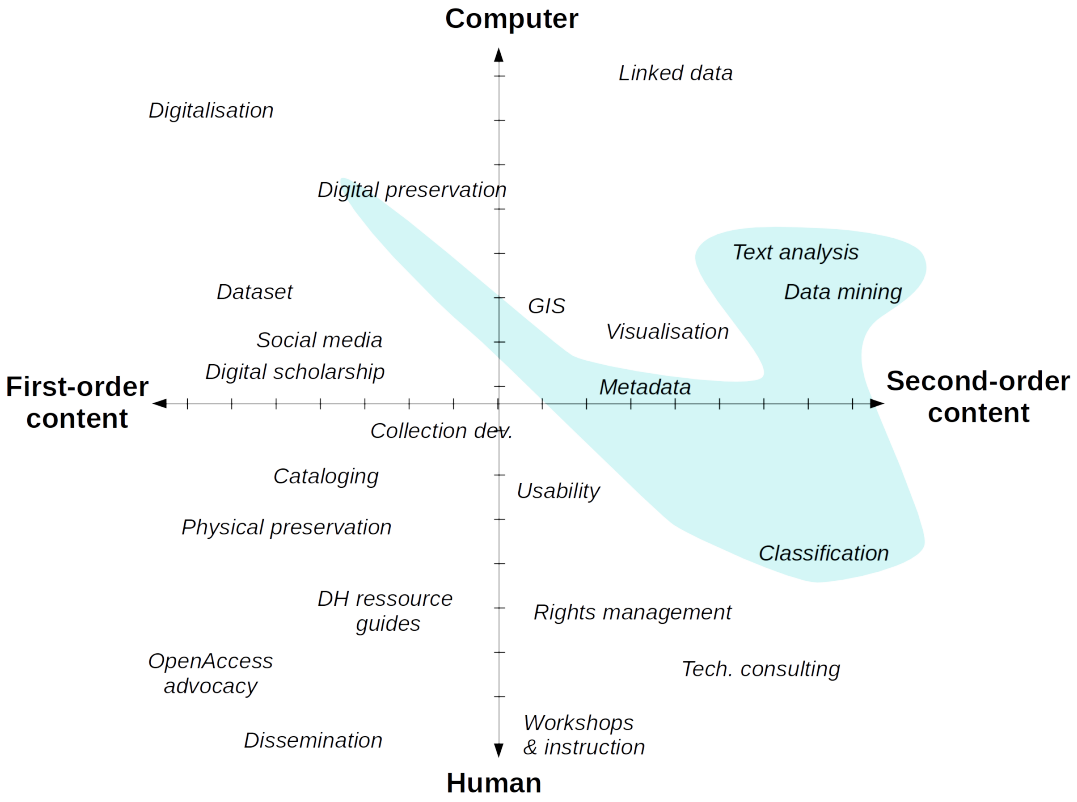


FIGURE 3 – Périmètre de la thèse au sein de la typologie des humanités numériques établie par Sula (2013)

Les données numériques étendent (voire renversent) la perspective classique de la documentation (accumulation des données, constitution de corpus, de collection) en renouvelant le questionnement historique. Les préoccupations se tournent maintenant vers l'accès, les possibilités de navigation et d'exploitation de ces données. En figure 3, le champ du numérique s'étend de la gauche (first order content) vers la droite (second order content) du graphique. À l'extrême de cette extension, ce sont les problématiques big-data, l'humain s'efface, il n'y a plus de constitution de corpus. Comme indiqué sur la figure 3, nous nous concentrerons sur l'exploitation des données, centrée sur la notion de corpus chère à l'histoire.

Nous constatons que les données historiques ne circulent pas facilement vers les données patrimoniales. Les ontologies sont la solution technique la plus souvent évoquée face à ce problème. Dans le cadre des ontologies du domaine culturel, l'humain est utilisé dans une fonction peu intéressante, tandis que de grand espoirs sont posés sur les inférences automatiques. Nous cherchons à renverser les rôles. Nous proposons un process partant des textes de corpus d'historien, dirigés vers l'analyse, la valorisation de cette analyse conjointement avec la documentation du patrimoine numérique concerné, et finissant par la gestion de l'archivage perenne de ce patrimoine numérique en lien avec le CINES. C'est le sens de la figure 3, qui évite soigneusement quelques éléments de la partie supérieure du diagramme, préférant plonger dans la partie inférieure.

La méthode de chaînage des connaissances historiques et patrimoniales proposée est basée sur les contenus textuels. Ces contenus peuvent être utilisés dans 2 finalités : analyse historique et documentation patrimoniale. Un des moyens consiste à calculer des proximités entre les textes du corpus. Ces proximités peuvent être de nature différente : espace, temps, thématique, etc. (multi-critère). De plus, les proximités thématiques par exemple sont multi-échelle ("zoomables").

Les grandes étapes de ce process sont décrites par le schéma SADT en figure 4.

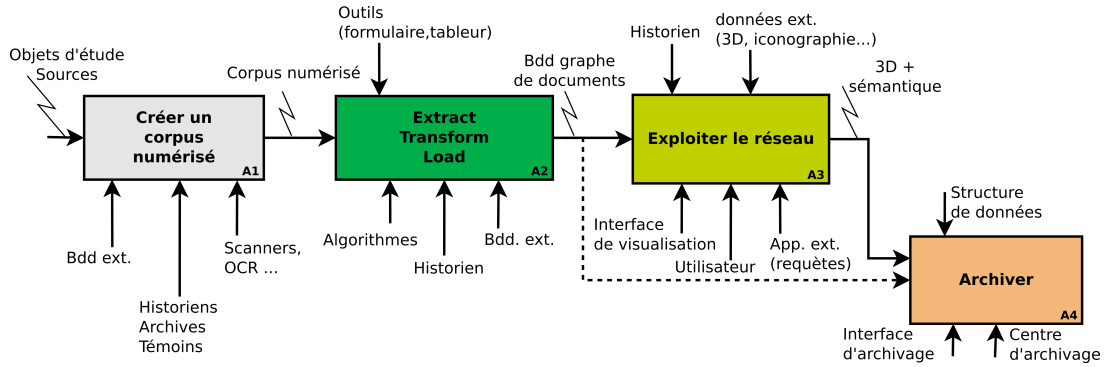


FIGURE 4 – Vue globale du process de chaînage des connaissances historiques et patrimoniales proposé par cette thèse

En négatif. Parmi beaucoup d'autres choses qu'elle n'est pas, cette thèse ne s'inscrit pas dans le domaine du "Big Data". En effet nous correspondons à peu de critères : Kitchin (2014) définit que les big data doivent concerner de gros volumes (Terabytes au moins) de données, traitées rapidement (proche du temps réel), très variées (structurées et non structurées) et exhaustives sur un domaine. Ce n'est pas non plus un modèle à implémenter, ni un guide de bonnes pratiques. Il s'agit par contre d'une méthode expérimentale (avec outil implémenté) répondant à une démarche explicitée et donnant lieu à des résultats d'expériences. Ce n'est pas de la visualisation de données même si ce domaine est utilisé. Enfin ce n'est pas une machine à tout faire automatiquement pour le patrimoine et l'histoire. Cet outil est destiné à l'analyse avec un historien et à la documentation du patrimoine à partir de contenus textuels. Il apporte ainsi de nouveaux potentiels à un expert du domaine et éventuellement à la valeur d'un bien patrimonial.

Objectifs, Hypothèses

Objectifs

L'objectif de cette thèse est de mettre en réseau les textes de corpus issus du domaine de l'histoire afin de :

— permettre des analyses avec l'historien, des interactions avec le spécialiste du domaine. Il s'agit de lui apporter de nouveaux contenus pour produire ensemble une valeur ajoutée historique.
— établir des relations sémantiques entre éléments patrimoniaux documentés par le corpus. Ces relations agissent comme un système de recommandation, et permettent une navigation entre éléments du patrimoine, matériel ou immatériel, affranchie des contraintes d'accès géométriques ou géographique.

Le schéma de la figure 5 illustre l'idée de joindre plusieurs éléments d'un ensemble historique cohérent autour de représentations géographiques, tridimensionnelles à partir de données textuelles.

Hypothèses

Nous formulons plusieurs hypothèses d'ordre général qui guident la construction des objectifs.

L'interaction avec l'historien est le seul moyen de produire des connaissances historiques. En effet plusieurs raisons motivent l'impossibilité d'envisager un outil d'analyse historique autonome (cf section "Critiques des humanités numériques" (1.1.3)).

— Les capacités de formalisation des connaissances qualitatives sont trop faibles.
— Le processus de production des données et de vérification pour l'exigence de qualité serait extrêmement fastidieux.
— L'éthique des humanités évite les processus boites noires.
— La notion de récit est fondamentale en Histoire, les faits explicites ne sont pas directement de l'histoire.
— Les jeux de données étiquetés en histoire des techniques sont rares.
— Une analyse complémentaire humaine nécessaire : l'intérêt de l'analyse numérique réside dans les nouvelles interactions qu'elle rend possibles.
— Les corpus sont à taille humaine.
— L'établissement de pré-catégorie n'est pas souhaitable pour l'analyse historique.

Contraintes

Les objectifs précités doivent être remplis en respectant des contraintes propres aux pratiques et exigences de l'Histoire. Sans quoi l'intention de produire *avec* l'historien restera lettre morte. Les points énoncés ci-dessous paraîtront (je l'espère) triviaux à tout historien, il fondent notre approche :

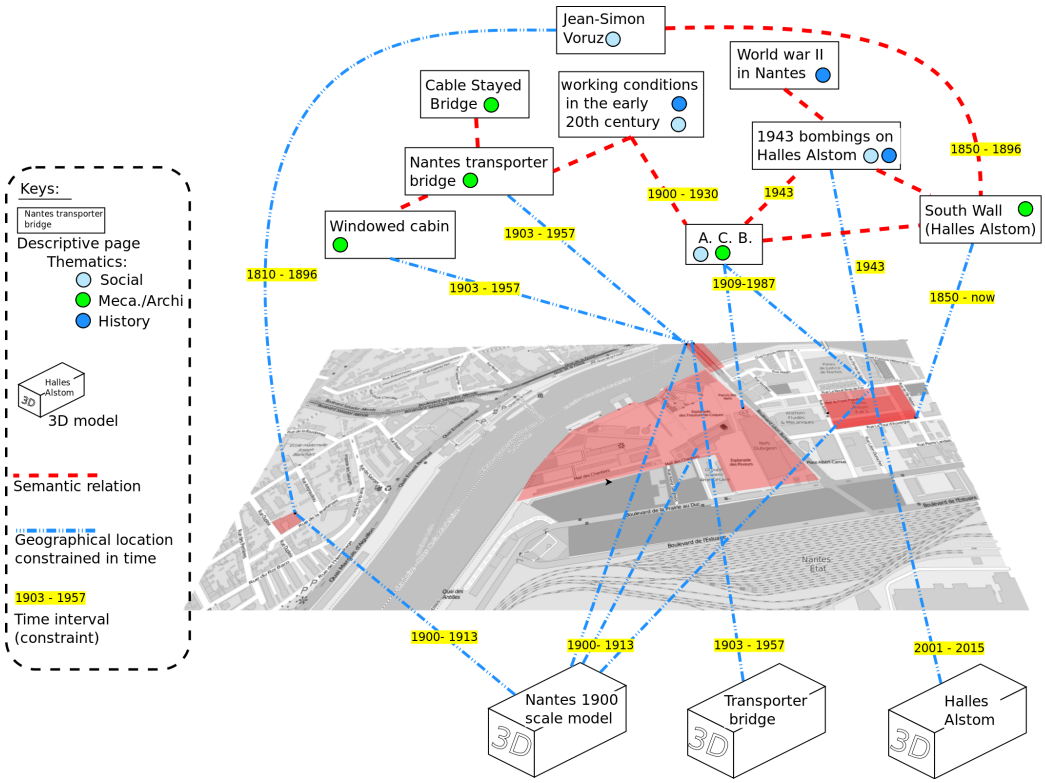


FIGURE 5 – Illustration des objectifs avec l'exemple du port de Nantes : relations sémantiques entre des patrimoines, documentations et informations parfois sans localisation

Contenus non structurés. Le terme "structuré", au sens informatique, fait référence à l'organisation des données. Concrètement, la structure informatique d'un texte en Histoire se limite souvent au découpage en parties, sous-parties, etc. La plupart ne présentent pas explicitement de découpage. Les éditions numériques enrichies (typiquement en *Text Encoding Initiative* (TEI), voir section 2.2) sont encore rares.

Volume du corpus. Les corpus que l'historien peut constituer, analyser et auxquels il est habituellement confronté contiennent entre 30 et 10 000 pages.

Qualité des données. Les données textuelles ne respectent pas nécessairement de dictionnaire (néologisme, jargon technique), certains mots peuvent être mal orthographiés, certains mots peuvent manquer. Ceci est particulièrement vrai lorsque le corpus est issu de *Reconnaissance optique de caractères* (OCR).

Pas d'*a priori*. Toute modélisation du corpus *a priori* doit être évitée. L'analyse serait considérablement biaisée par la création de classes (catégories) à trouver. Ces classes opéreraient une restriction du domaine d'étude et des contenus des textes. L'analyse doit permettre de conserver le maximum de la variabilité des contenus. Par ailleurs ce type de travail doit au préalable faire l'objet de consensus entre chercheurs. À ces titres nous ne pouvons pas utiliser les techniques de type Reconnaissance et classification d'entités nommées (NERC).

Corpus unique. Complémentairement au point précédent, le corpus est considéré comme unique. Chaque partie du corpus peut uniquement compter pour ce qu'elle est et jamais comme représentante d'un ensemble plus vaste. En d'autres termes, nous évitons les généralisations, surtout lorsqu'elles ne sont pas intentionnelles. Pour ces raisons nous éviterons les algorithmes impliquant de l'apprentissage, évitant ainsi les biais provenant d'un autre corpus et ou d'une partie du corpus sur le reste.

Nuances. Les humanités, peu manichéennes, s'intéressent rarement à fixer des variables booléennes. Mis à part les faits simples (untel est le fils de tel autre), la nuance est de rigueur. L'idée même de quantifier des proximités sur un mode précis (untel partage 78% de ses objets de recherches en chimie avec tel autre) peut s'avérer problématique. Ceci nous impose la logique floue plutôt que les associations binaires, et nous incite à toujours vérifier le sens d'une quantité.

Quantifications. Il existe un intérêt à compléter la lecture qualitative d'un corpus textuel, par une analyse quantitative. Si les 2 univers sont incommensurables, il est néanmoins possible de les exploiter conjointement (sans les comparer) dans une analyse combinée.

Présentation du mémoire

La figure figure 6 présente le déroulé simplifié de ce manuscrit.

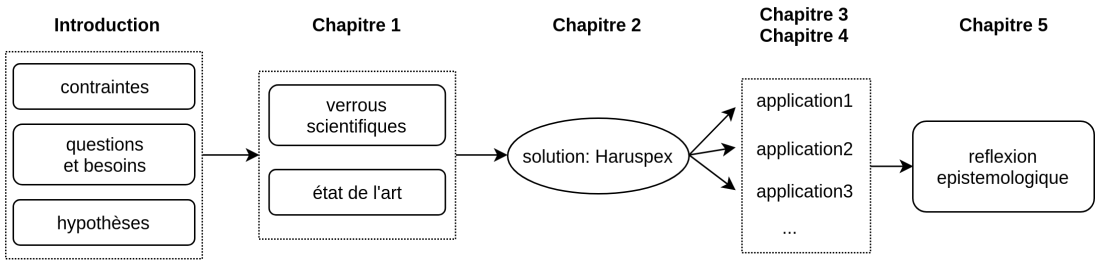


FIGURE 6 – Déroulé simplifié de la thèse

Chapitre 1 : État de l'art. Ce chapitre décrit les enjeux et les problématiques soulevés par cette thèse. La thèse étant pluridisciplinaire, ce chapitre (1) introduit quelques notions et débats en humanité numérique, (2) présente les systèmes de formalisation des connaissances et les technologies associées et (3) approfondit les avancées algorithmiques pour l'analyse des données textuelles brutes. Quelques résultats hors-cadre du patrimoine et de l'histoire, issus du grand domaine de la *gestion de connaissances* sont également examinés. Ce chapitre énonce les verrous nous résolus par l'état de l'art.

Chapitre 2 : Proposition scientifique : Haruspex. Activité A2 sur la figure 4. Ce chapitre décrit la méthode proposée pour analyser les corpus de textes. Cette méthode implémentée est nommée *Haruspex*. Le terme anglophone « *haruspex* » (haruspice en français) désigne une personne pratiquant l'art (divinatoire) de lire dans les entrailles d'un animal sacrifié. *Haruspex* propose de nouvelles vues d'un corpus de textes bruts, sans apprentissage ni pré-traitement. *Haruspex* est en cours de dépôt de logiciel.

Chapitre 3 : Application : Cartographie d'un corpus en Histoire de la Chimie du Solide. Activité A3 sur la figure 4. Ce chapitre propose une analyse historique utilisant *Haruspex*. L'analyse concerne un corpus d'archives orales sur l'émergence de la chimie du solide. Ce chapitre met en application les développements précédents.

Chapitre 4 : Autres applications : corpus de textes et patrimoine. Activité A3 sur la figure 4. Ce chapitre propose différentes exploitations du réseau au travers de cas d'étude qui ont contribué au développement d'*Haruspex*. Les cas d'étude sont les suivants :

— corpus de sous-parties de mémoires de recherche, en lien avec un corpus de fiches de médiation de musée. Il s'agit d'étudier la complémentarité des graphes produits.
— corpus d'articles scientifiques avec une composante temporelle.
— cas de valorisation d'un corpus d'archives en lien avec un patrimoine numérique (3D) : les salons Mauduits de Nantes.

Chapitre 5 : Épistémologie d'une méthode numérique. Ce chapitre établit un retour réflexif sur la démarche suivie et les résultats : c'est une épistémologie pratique. Il traite d'une relation inventée entre Histoire et génie industriel, aux confins des avancées de l'informatique. Ce chapitre apporte les éléments d'explication de la proposition en termes épistémologique et d'historiographique.

Chapitre 6 : Conclusion.

Chapitre 1

État de l'art

"Les hommes font l'histoire du numérique, le numérique peut-il faire l'histoire des hommes ?"

Contents

Introduction

L'Histoire comme discipline manipule en entrée des documents sources, principalement des textes, et en sortie des connaissances, souvent formalisées sous forme de textes également. Le patrimoine matériel de type scientifique et technique est un domaine où les connaissances sur l'objet sont primordiales. Ces connaissances sont, pour partie au moins, issues d'études historiques.

Nous tentons ici d'esquisser les éléments structurant le paysage des méthodes numériques liées à l'histoire et au patrimoine. Ce paysage se complexifie avec l'engouement des chercheurs et des financements depuis une vingtaine d'années. L'immensité de ce champ de recherche buissonnant rend impossible l'exhaustivité de cet état de l'art. Nous focaliserons donc cette étude sur le champ proche de celui de l'analyse des données. Le versant « gestion » de ces données ne sera abordé que dans le cadre restreint de l'archivage pérenne et de la formalisation des connaissances.

Section 1 : Nous nous intéressons d'abord aux humanités numériques : La fascination pour les nouveaux outils et les promesses de progrès technologiques incitent toujours plus de chercheurs à se tourner vers la « terre promise » (ici du numérique). L'heure d'écrire une histoire de cette discipline semble être arrivée, de nombreuses contributions émergent depuis 3 ans. Ce nouvel espace de recherche n'est pas homogène, les critiques sont intéressantes à relever.

Section 2 : Dans un second temps nous étudions le concept de gestion de données dans le monde du patrimoine, les outils associés, les standards dans ce domaine et les modèles permettant la gestion conjointe des informations et de l'objet notamment numérique.

Section 3 : Dans cette section, nous étudions les modes de structuration classiques des connaissances en humanités (voir figure 1.15) pour construire un édifice des données vers les connaissances (voir figure 1.2) présenté en section 2

Section 4 : Cette section clé présente les mécanismes existants pour l'analyse de textes. Nous étudions les propositions des recherches en traitement automatique des langues naturelles (TAL) et en traitement de données applicables aux textes.

1 Patrimoine et humanités numériques

1.1 Les humanités numériques

Les humanités produisent des corpus de textes, hétérogènes par leur forme et leur contenu et spécifiques par leur terminologie et leur signification. Ces corpus sont classiquement analysés à travers des lectures exhaustives et des interprétations constamment remises sur le métier. Depuis plusieurs décennies cependant, chercheurs des sciences informatiques et des sciences humaines développent des méthodes quantitatives d'analyse des corpus textuels.

1.1.1 Historique

Émergence et développement. Un des exemples les plus marquants de la fondation des humanités numériques (*digital humanities*, Humanités Numériques (HN)) a eu lieu dès les années 1950. Voulant établir l'index verborum des œuvres de Thomas d'Aquin, Roberto Busa, un prêtre jésuite, a obtenu le soutien du fondateur d'IBM, Thomas J. Wanted, qui a mis à sa disposition quelques ordinateurs et informaticiens d'IBM pour traiter les onze millions de mots du corpus aquinien (Hockey, 2007).

De telles analyses numériques se sont largement répandues depuis, notamment avec l'essor d'internet et du web à partir des années 1990. Elles constituent aujourd'hui le champ identifiable des « humanités numériques » avec la publication en 2004 d'un impressionnant livre collectif (Schreibman *et al.*, 2004) marquant la reconnaissance internationale de ce terme (*digital humanities* en anglais). Ce champ, à la croisée de multiples cultures épistémiques, est fascinant par de nombreux aspects. D'abord parce qu'il modifie profondément les modalités d'édition et de conservation des textes (changement de supports et de médias). Ensuite parce qu'il introduit de nouveaux médiateurs dans la circulation des informations, (Stiegler, 2009) parle d'économie de la contribution à propos de cette reconfiguration du rôle des éditeurs et des profanes. Enfin parce qu'il façonne de nouvelles pratiques, dont l'analyse des corpus textuels qui nous occupe ici.

Périodisations. L'histoire de ces pratiques numériques est récente (Hockey, 2007). Berry (2011) analyse le tournant computationnel en 2 vagues principales :

1. Les humanités se servent de l'efficacité des machines plutôt qu'elles ne les impliquent dans une critique. Cette première vague concerne les textes, les iconographies et les bases de données. Le *Digital Humanities Manifesto 2.0* fait démarrer cette vague dans les années 1980. On peut parler de transdisciplinarité (cf. figure 1c.
2. Les humanités numériques génèrent et interprètent des données, les rassemblent. Ceci caractérise ce que Berry appelle le *computational turn*. Elles sont capables de critiquer/mesurer la qualité des résultats. Cette seconde vague concerne les données structurées typiquement les ontologies et les données liées (cf. section 2.2).

Cette opposition entre 2 vagues se retrouve également dans l'opposition « *distant reading* » popularisé par Moretti (2005) faisant l'apologie des machines à lire face aux techniques de « *close reading* » qui — selon Kitchin (2014) — s'intéressent au texte en soi (termes employés, styles de discours, rhétorique, non-dit, etc.) pour en comprendre le sens.

Les objets de la première vague restent au centre des préoccupations des humanités : analyses de textes et iconographies. Néanmoins, les HN s'ouvrent à d'autres pratiques : les objets virtuels tridimensionnels (3D), les pratiques d'édition critiques numériques (avec la TEI), le crowdsourcing, l'analyse des réseaux sociaux, etc.

À propos de ces nouveaux éléments, Burnard (2012) propose une autre périodisation, tripartite, dont les motifs du découpage sont davantage orientés par les problématiques d'édition [1] :

1. *Litterary and Linguistic Computing* (LCC) (1960-1980) correspond au premier age, le plus long, le moins structuré. Il débute après la guerre et se caractérise par l'usage de l'ordinateur pour établir des index, des analyses de styles basées sur les occurrences de mots. Cet âge évolue de front avec l'histoire quantitative. Il marque la naissance de la textométrie, discipline encore largement pratiquée aujourd'hui (voir section "Outils d'analyses de textes existants" (4.4)).
2. *Humanities Computing* (HC), le second age, commence dans les années 80. Les modèles quantitatifs dominent toujours. La rupture avec l'age précédent consiste en la naissance d'une prise de recul et d'une réflexion sur les pratiques numériques en humanités. Un grand débat émerge sur la structuration d'une nouvelle discipline. Les préoccupations se tournent alors vers l'archivage et l'uniformisation des pratiques, la standardisation (voir section "Modèles et descripteurs de données pour le patrimoine" (2.2)) : métadonnées, schéma TEI, etc.
3. L'age des *Digital Humanities* (HN) s'ouvre avec la naissance du Web dans les années 90. Il correspond au 3e âge et se focalise sur la notion de *distribution* : stockage des connaissances (fortes des normalisations précédemment élaborées) et calculs distribués (*grid computing*). Les corpus sont en ligne et les analyses portent sur les productions nativement numériques, massivement accessibles, comme les réseaux sociaux. La frontière producteur/utilisateur se brouille, avec l'émergence le crowdsourcing.

La première périodisation est nettement axée sur le rapport de l'informatique aux données (d'abord analyses brutes, puis raisonnement, inférences automatiques…) Elle marque une évolution des attentes. Elle est ancrée dans une idée de *progrès*. La seconde périodisation est axée sur les usages, davantage ancrée dans les faits, mais limitée dans les perspectives qualitatives. Les avancées sont quantitatives : toujours plus de données textuelles accessibles.

1.1.2 Enjeux, efforts liés aux humanités numériques

Enjeux. Ladurie (1973) déclare « L'historien de demain sera programmeur ou ne sera plus ». Aujourd'hui, 45 ans plus tard, des historiographes prudents rappellent l'« impératif philologique » (Vico, 1725), la nécessité de restituer les conditions de production des savoirs dans leurs variations et leur inévitable fragilité.

En effet avant d'être des HN, encore faut-il prétendre au statut d'humanité. Une approche de ce noyau dur (*hard core*) des humanités est proposée par Lakatos *et al.* (1980) : au quotidien les universitaires mènent des raisonnements sur des données floues, finement extraites de leurs corpus de sources. Ces raisonnements sont principalement basés sur des hypothèses implicites, des fondements ontologiques et des connaissances externes (hors-corpus). Les résultats construisent des distinctions finement nuancées, démarquant des problématiques épistémologiques, impliquant la notion de rédaction et de récit.

De nombreux auteurs comme le Deuff (2014); Clavert (2013) pensent que le meilleur des HN est à venir et qu'il faut accompagner l'historien vers son prochain statut de programmeur. le Deuff (2014) fait l'hypothèse d'une évolution de la production de l'historien vers une Interface de Programmation d'une Application (API), qui serait un dépassement de l'habituel PDF en ligne, offrant un cadre de diffusion élargi. Les données produites seraient interrogeables et en perpétuelle évolution. Le mot API est à mon avis mal choisi ici, cette idée fait référence à la formalisation des données (section 2.2). Ce mouvement s'inscrit dans l'idée générale, énoncée au tournant entre les *Humanities Computing* et les *Digital Humanities* par Rosnay (1995) qui se demande si nous assistons au développement d'une humanité cognitivement augmentée.

En écho à la périodisation en 2 vagues de Berry (2011), avec moins d'enthousiasme, Dyens (2008) distingue technologies de l'intelligence et technologies intelligentes. Les premières sont dédiées à la construction de nouvelles idées, nouvelles méthodes, nouvelles approches de la connaissance. Les secondes tendent à se substituer à l'intelligence humaine et ambitionnent de raisonner aussi bien que l'expert du domaine. Le risque réside alors dans la croyance que nous avons en ces résultats. En ce sens les HN sont ce que Stiegler (2007) nomme un *pharmakon* : un remède et toujours aussi un poison.

Efforts. La création de la très grande infrastructure de recherche (TGIR) Huma-Num dédiée aux HN en France en 2013 marque l'intensification des efforts pour la mise en œuvre du numérique dans les sciences sociales. L'association francophone des humanités numériques, *Humanistica*, créé en 2014 vise à rassembler et animer la communauté. Au-delà de la réflexion, quelques projets sont également menés. De grandes structures se développent dans beaucoup d'autres pays occidentaux, DARIAH-DE en Allemagne et CNR ILC en Italie par exemple. Ils montrent l'intérêt de la communauté internationale des chercheurs de toutes disciplines et de toute la société pour ces enjeux.

1.1.3 Critiques des humanités numériques

Face à l'optimisme et l'enthousiasme ambiants, dans une société soumise à la technologie et enrayée dans le « présentisme » (Hartog, 1993; Mounier, 2017), certains penseurs évoquent les HN comme une rencontre ratée entre une technique et une discipline (Genet et Zorzi, 2011) en écho de la phrase de Olsen (1993) : « Computer- aided literature studies have failed to have

1. l'auteur de cette périodisation est un pionnier de l'édition numérique (TEI)

a significant impact on the field as a whole ». Mounier (2017) n'hésite pas à employer les mots « tarte à la crème des discours sur l'innovation à l'Université » et critique principalement l'enrôlement des humanités numériques au profit du simple numérique. Cet enrôlement déplace l'enjeu (central) de connaissances et de critique, vers un jeu d'analyses prédictives. Selon l'auteur, les prédictions mettent l'historien à l'écart.

Drucker (2011) critique la forme graphique que prennent les analyses en HN et l'effet boite noire que ces formes impliquent. L'auteur déclare que la rhétorique de ces représentations, tirée des sciences empiriques, cache leurs biais épistémologiques. Il est impossible d'examiner les sources de ces constructions.

L'échec des HN encore non surmonté (Liu, 2016) est renforcé par les nombreuses critiques (Prendergast, 2005; McCarty, 2016) à l'égard du *Distant reading* promu par Moretti (2005). Le *distant reading* représente ici la seconde vague, il consiste à ne plus lire les textes mais à visualiser des masses de textes sous forme de graphes, arbres, etc. Enfin la critique du « fantasme récurrent de la bibliothèque universelle » de Welger-Barboza (2001) vise le mythe de l'interopérabilité des données et la dématérialisation des institutions culturelles. Ces critiques marquent une tendance à associer la seconde vague des analyses computationnelles avec l'empirisme, le positivisme et d'autres entreprises suspectes. Ce sont effectivement des écueils qu'il faudra éviter dans notre travail.

1.2 Patrimoine et numérique

L'industrie contemporaine produit massivement des données et artefacts numériques (*born digital*, notamment en 3D). Après les premiers logiciels de CAO volumique, invention attribuée à Pierre Bézier avec UNISURF en France dès 1968, la production numérique s'intensifie rapidement. À partir des années 1980 deux phénomènes participent de sa démocratisation dans le monde industriel : le développement des capacités graphiques pour visualiser de la 3D (à l'écran) et la baisse des coûts, qui deviennent compétitifs face au prix de la main d'œuvre (dessinateur industriel). Dans cet élan, en France on observe l'industrialisation du logiciel EUCLID (issu du CNRS) en 1980 et d'un logiciel phare du domaine en 1981 : CATIA, encore leader du marché aujourd'hui. Aujourd'hui, plus de 30 ans après cette explosion du numérique industriel nous réalisons que ces objets virtuels sont candidats à notre patrimoine industriel. Notre décennie (2010+) va potentiellement voir l'arrivée programmée de ces objets au titre de patrimoine. En effet, 30 ans cela correspond à la limite basse recommandée par l'Inventaire Général du patrimoine avant la réalisation de l'enquête (de Massary et Coste, 2007). Des institutions muséales comme le CNAM devront sans doute prochainement gérer ces cas d'objets techniques nativement numériques.

Rosenzweig (2003) estime que ces masses de données nativement numériques mettent le patrimoine en porte-à-faux entre pénurie et abondance : « notre système pour préserver le passé est en crise » : mal documentée et sans système de sélection, ni préservation, ces informations, réduites en données numériques sont *présentistes* à l'extrême.

Accentuant la tendance, le récent essor et la démocratisation des techniques d'acquisition tridimensionnelles et photographiques (couplées à des systèmes de stockage performants), permet aux musées et laboratoires de recherche de numériser leurs collections ou objets de recherche (scans 3D, photographie Haute Définition (HD)). Des campagnes de numérisation massives fleurissent dans tous les musées du monde occidental, avec souvent le soutient de l'état, comme au Canada avec le projet *Digital Canada 150* ou en France avec différents programmes de numérisation nationaux. La littérature concernant l'acquisition 2D et surtout 3D est très nombreuse, de nombreuses études ont déjà été réalisées (Gomes *et al.*, 2014).

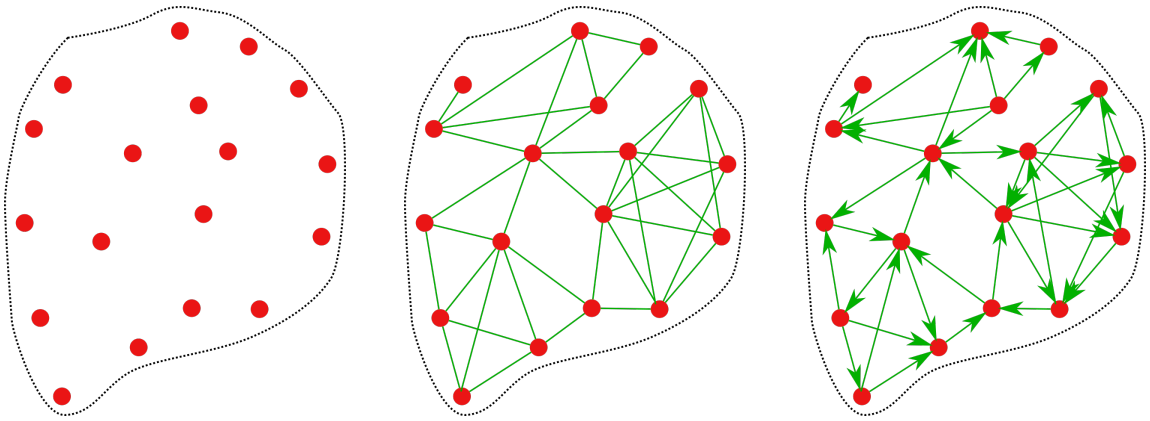
Nous présentons ici notre conception du patrimoine, les cadres principaux du patrimoine en lien avec le numérique, et enfin une série d'approches sur le patrimoine et le numérique : documentation, analyse, valorisation, diffusion.

1.2.1 Considérations sur le patrimoine

> « [...]nous, qui depuis le présent, avons reconnu à cet objet une valeur et considérons que ceux qui l'ont créé feraient, pour nous, de « bons » ancêtres culturels » (Davallon, 2002)

Processus de patrimonialisation Plusieurs acceptations du patrimoine rivalisent. Celle de Ruskin (2008), qui énonce en 1849 que le patrimoine (materiel exclusivement) tire sa valeur d'un temps actif passé : celui de l'investissement moral et de la qualité du travail dont il a fait l'objet (et non pas d'une restauration). Celle de Riegl (1984) propose une grille de valeurs et de sous-valeurs pour analyser le patrimoine matériel ; il insiste sur la valeur historique du patrimoine, dépendante de la lecture qu'on en a au présent, par opposition à une valeur universelle. Choay (2009) perpétue cette pensée en considérant que le patrimoine « est une construction intellectuelle ». Davallon (2002) dans cette même tradition développe le concept de « filiation inversée » qui décrit les étapes suivantes du processus de patrimonialisation :

1. la trouvaille
2. l'authenticité de l'objet vis-à-vis du monde d'où il vient
3. l'authenticité du monde d'où il vient
4. la *typicité* (Heinich, 2009) de l'objet (représentatif du monde d'où il vient)
5. la célébration par l'exposition
6. la transmission aux futures générations (conservation)



(a) Intégrité de composition (b) Intégrité structurelle (c) Intégrité fonctionnelle

FIGURE 1.1 – Les 3 niveaux d'intégrité, valeur fondamentale du patrimoine mondial de l'UNESCO, en lien avec l'authenticité.

Critères pour le patrimoine numérique Les principes ratifiés par l'ICOMOS (1996) s'intéressent à l'enregistrement du patrimoine. La charte de l'UNESCO (2003) décrit un ensemble de « bonnes pratiques » relatives à la conservation du patrimoine numérique. Cette charte sera rapidement suivie par des préconisations plus pratiques : la charte de Londres (AAVV, 2009), qui énonce 6 principes pour la visualisation assistée par ordinateur dédiée à la recherche et la valorisation du patrimoine culturel. Cette dernière sera étendue par les principes de Séville (of Virtual Archaeology, 2013) dans le cas d'applications à l'archéologie virtuelle. Cependant aucun de ces cadres ne décrit de conseils de mise en œuvre, ni de cahier des charges technique afin de respecter les préconisations d'analyses scientifiques et de capitalisation de connaissances. Sans prétendre à l'écriture d'un tel cahier des charges nous ferons une proposition technique pour le couplage de la médiation de musée avec la rigueur d'analyse historique nécessaire pour le patrimoine. Le tableau 1.1 compare ces différents cadres.

	Critères de classification	Préconisations pour l'analyse scientifique	Préconisations méthodologiques	Conseils de mise en œuvre
UNESCO	O	N	N	N
ICOMOS	N	O	∼	N
London Charter	N	∼	O	∼
Seville Principles	N	∼	O	∼

TABLE 1.1 – Tableau récapitulatif des cadres du patrimoine numérique (O pour oui ; N pour Non). Repris de Hervy (2014)

Critères UNESCO Les critères UNESCO sont nombreux, des « Orientations devant guider la mise en œuvre de la Convention du patrimoine mondial » (UNESCO, 2016), nous approfondissons ici 2 critères fondamentaux : l'*authenticité* (points 79 à 86) et l'*intégrité* (points 87 à 95).

L'*authenticité* consiste à estimer la valeur du patrimoine en se basant sur la crédibilité ou véracité des sources d'informations en relation avec les caractéristiques originelles et subséquentes du patrimoine. Il s'agit d'estimer si les caractéristiques suivantes du patrimoine portent encore un sens tels que les sources le décrivent dans son état authentique (non altéré) : forme et conception ; matériaux et substances ; usages et fonctions ; techniques, traditions et systèmes de gestion ; situation et cadre ; langue et autres formes immatérielles.

L'*intégrité* estime l'état du patrimoine en question. Chacune des étapes est soumise à l'examen d'authenticité. L'intégrité peut se déployer en 3 niveaux ou sous-parties :

1. Intégrité de composition (fig. 1.1a) : connaître le bien. Réaliser l'inventaire et la cartographie des composants du bien. En connaitre les manques. Établir le contour du bien.
2. Intégrité structurelle (fig. 1.1b) : comprendre le bien. Comprendre les relations qu'entretiennent les éléments issus de l'intégrité de composition. Par exemple : le fonctionnement mécanique d'une machine ou les fonctions des parties d'un site.
3. Intégrité fonctionnelle (fig. 1.1c) : utiliser le bien. Cette idée se rapproche du patrimoine vivant. Il s'agit d'une valeur d'usage du bien. L'objectif est de statuer si les relations précédemment établies sont encore exploitées, concernent encore un usage.

2 Gestion des données pour le patrimoine

2.1 Outils et concepts de gestion de données

Cette partie s'intéresse aux grands concepts de la gestion des données jusqu'aux connaissances. Ce domaine est immense, les outils décrits ici sont sélectionnés pour ce qu'ils peuvent apporter au domaine du patrimoine et sont approchés sous l'angle patrimonial.

Données, Information, Connaissances Tentant de définir ces notions indépendamment de la discipline, Zins (2007) établi une typologie complexe. Heureusement ces éléments font l'objet de définitions claires dans le paradigme numérique Ackoff (1989); Rowley (2007). Nous les adopterons donc. Nous ne traiterons pas de la partie *wisdom* (sagesse) qui est hors-sujet [2].

Les données sont des signes discrets, et correspondent à la partie la plus élémentaire d'un système d'information, ce qui correspond *a minima* à un train de bits dans notre paradigme numérique.

L'information est composée de signes signifiants, découle des données, et rend les données exploitables. Il s'agit *a minima* d'un train de bits avec le moyen de les décrypter : encodage, schéma de lecture, etc. Cette description permet de répondre à « quoi » correspondent les données dont l'information est constituée, éventuellement *qui*, *quand* et *où*.

Les connaissances sont issues des informations, elles correspondent à la capacité d'associer des informations (notion de contexte) pour en produire des instructions. Ce sont des informations utilisées. La connaissance est donc liée a minima au raisonnement à partir d'informations. Ces raisonnements permettent de répondre à « comment » ou « pourquoi » ces informations ont un sens. Elles peuvent être transmises ou extraites par expérience.

Ces définitions ne concernent pas la nature de l'élément mais plutôt son potentiel fonctionnel. Par exemple cette thèse est élémentairement faite de **données** (suite de signe discrets). Ces signes forment des mots en français écris en alphabet latin encodé utf8. Il s'agit donc de données exploitables en français, ce sont des **informations**, ces mots sont organisés en phrases et paragraphes, parties, chapitres, etc. pour en augmenter l'exploitabilité. L'exploitation du contenu, voire appropriation, ou toute opération dépassant la suite de mots, créant des relations entre eux et éventuellement avec d'autres éléments extérieurs donne lieu à des **connaissances**. Drucker (2011) établit une distinction étymologique et pratique entre *data* et *capta*. Les *capta* sont captés

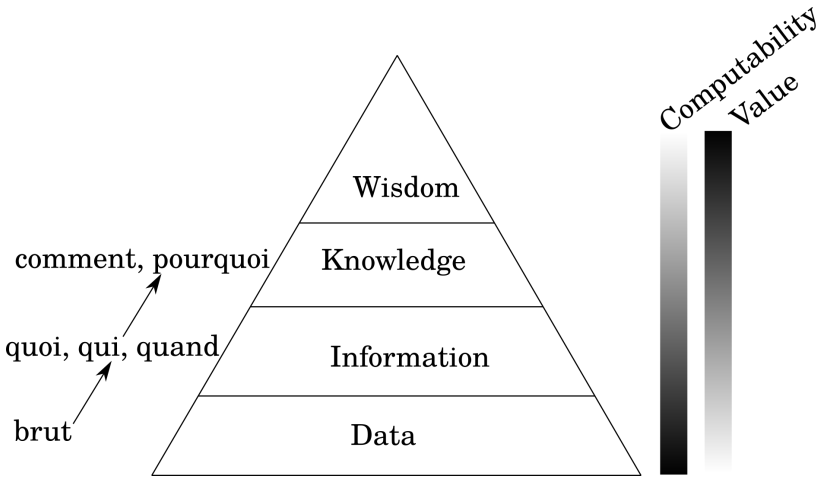


FIGURE 1.2 – Représentation classique des 4 niveaux de la pyramide DIKW, seuls les 3 premiers nous intéressent

activement, les données sont reçues. Il explique que les humanités font état de production situées, partielles et construites et non simplement données établies comme faits pré-existants. Malgré cette intéressante distinction, nous conserverons « donnée » (*data*) comme terme générique.

2.1.1 Bases de données

La gestion de ces données liées au patrimoine (textes, notices et iconographies) existe depuis longtemps au niveau national et international : bases Palissy, Mérimée, Mémoire, Archidoc, Gallica, Joconde, etc. Outre ces bases officielles, de nombreuses autres bases existent, pour de nombreux projets liés au patrimoine numérique, chacune avec son propre modèle de données. Elles sont parfois accessibles à distance. L'interopérabilité et la standardisation ne sont pas nécessairement des atouts, mais toujours des contraintes. L'usage de la base de données dicte les contraintes. Les modèles de données sont souvent formalisés avec la représentation *Unified Modelling Language* (UML), d'autres contraintes sur ces données peuvent être exprimées en *Object Constraint Language* (OCL). Les requêtes sont formulées en *Structured Query Language* (SQL), adapté à la formulation de type « clé-valeur ». La forme de la requête dépend du schéma de données. Les possibilités d'analyse sont définies par ce schéma.

2. Excepté peut-être en philosophie du numérique, ce qui n'est pas le sujet ici.

Certaines de ces bases sont aujourd'hui mappées (alignées) sur des modèles de données plus génériques (souvent des ontologies), ce qui les rend requêtables par différents moyens : soit via leur modèle de données (spécifique), soit via le modèle de l'ontologie (générique). L'exemple de la structure de données du projet Nantes1900 (cf fig. 1.3) est à l'opposé des bases standardisées de

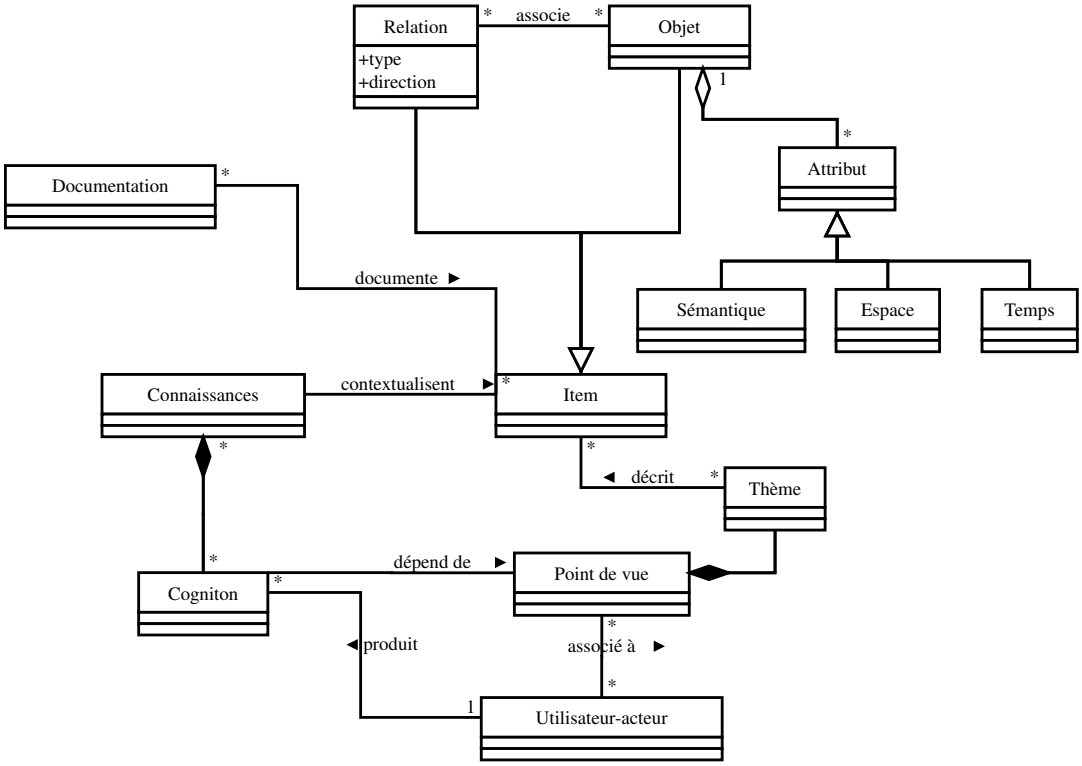


FIGURE 1.3 – Modèles de données de Nantes1900 (Hervy, 2014)

diffusion du patrimoine, elle est destinée à la valorisation in situ. L'organisation des données est orientée utilisateur et permet une grande souplesse coté objets : les relations et les objets sont des `items`. Les objets sont décris par des attributs. Les items (objet ou relation) peuvent être documentés, et sont regroupés en thèmes. Les connaissances de l'utilisateur contextualisent les items. Il sélectionne des thèmes qui filtrent les items.

2.1.2 OLAP : Bases de données multi-dimensionnelles et multi-échelles

Les bases de données OLAP (Traitement Analytique en Ligne) sont un moyen de stocker et d'accéder efficacement à des données multidimensionnelles et multi-échelles (Chaudhuri et Dayal, 1997). Ces données sont conceptuellement représentées dans un hypercube. Chaque dimension de ce cube est une dimension des données. Par exemple : temps, espace, type, etc.

Définition. On parle de cube Traitement Analytique en Ligne (OLAP), même lorsqu'il s'agit d'un hypercube. On considère 3 types d'éléments dans un base OLAP, ils sont ici expliqués dans le paradigme des schémas de métadonnées (voir section 2.2).

— la mesure : l'objet concerné
— la dimension : un élément du schéma de structure
— le label : un élément du schéma de valeur

Les systèmes OLAP fonctionnent par agrégation de contenu et par création de vues (les dimensions du cube). Les agrégations fonctionnent surtout avec des relations *many-to-one* (répétition d'attributs), c'est-à-dire que plusieurs instances de mesure peuvent partager une même valeur de dimension.

Par exemple, sur la figure 1.5, plusieurs instances d'enregistrement UNESCO doivent pouvoir pointer vers un même type de patrimoine UNESCO (culturel, naturel ou paysage culturel). Plusieurs instances de mesure partagent cette même dimension. Les cubes OLAP (figure 1.4) permettent d'explorer les dimensions communes à plusieurs données et leurs croisements. Il faut parfois rendre explicite une hiérarchie dans une dimension, dans le cas de dimension classique (le temps) elles sont créées à partir des données formatées (années, mois, jour, etc.)

Le cube OLAP est toujours créé à partir de structures de données relationnelles particulières :

— en étoile (*star*) : schéma simple avec une classe centrale (mesure) et un seul niveau de profondeur pour les dimensions
— en flocon (*snowflake*) : les dimensions peuvent avoir plusieurs sous composantes
— en constellation de fait (*fact constellation*) : schéma en flocon partageant des tables de label. Par exemple les labels temporels sont utilisés par plusieurs dimensions de la mesure initiale.

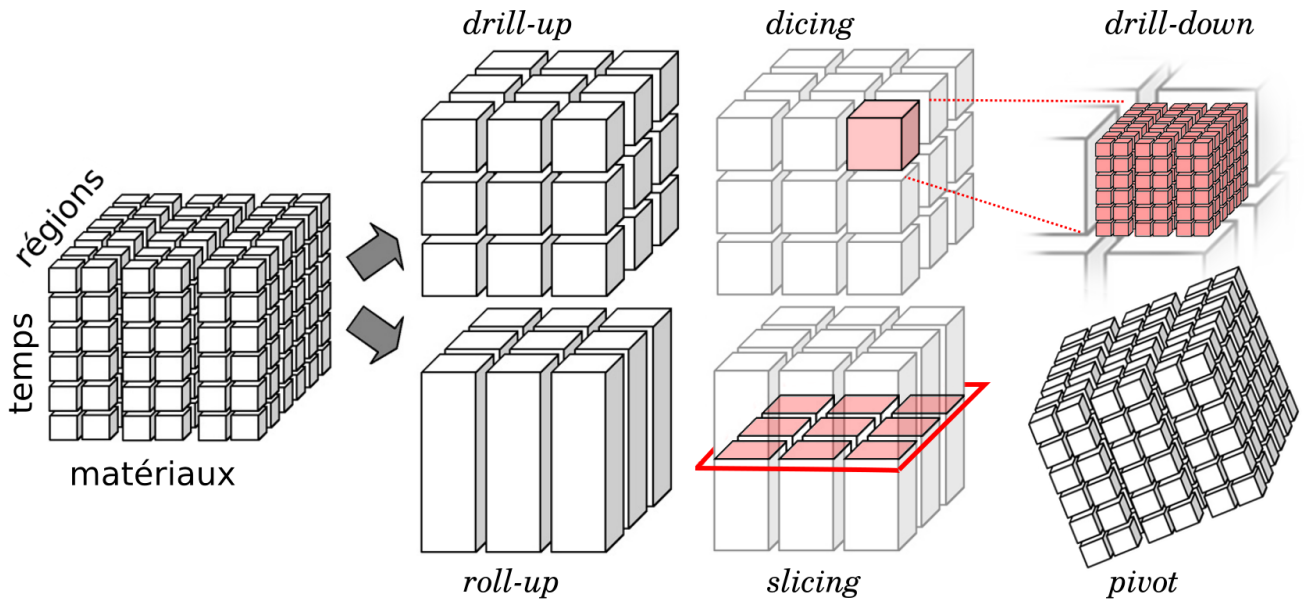


FIGURE 1.4 – Représentation d'un cube OLAP de 3 dimensions : temps, régions et matériaux. Quelques opérations. En pratique des hypercubes de dimensions supérieures à 3 sont souvent utilisés

Les schémas en « constellation de faits » sont typiques de OLAP. L'exemple de la figure 1.5 est une « constellation de faits » car la table *time* contient les labels de plusieurs dimensions. Les structures de données différentes, par exemple celles de Nantes19001.3 ne permettent pas de générer un cube OLAP, cependant il est parfois possible d'envisager un cube « conceptuel » qui recoupe plusieurs dimensions. Les opérations sur les dimensions du cube permettent des analyses à la volée via les requêtes.

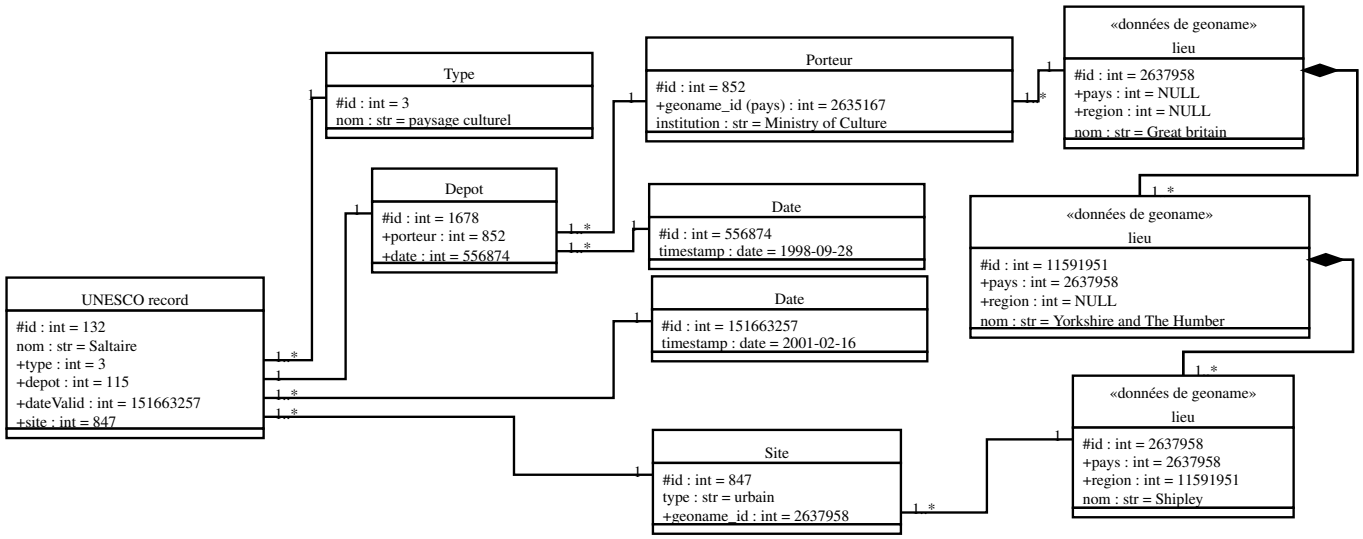


FIGURE 1.5 – Exemple de modèle *fact constellation* : relations entre classes instanciées (enregistrement UNESCO).

Les requêtes sont écrites en expression multidimensionnelles (MDX). Les exemples ci-dessous utilisent la figure 1.4. Les requêtes spéciales peuvent concerner une tranche du cube (*slicing*) un sous-cube (*dicing*), des zooms sur les niveaux de données (*drill-up, drill-down*), par exemple passer de région, à pays ou à département. La projection de données le long d'une dimension (*roll-up*), permet de créer une nouvelle *vue* des données en réalisant une opération, par exemple une somme de toutes les superficies concernées par les matériaux. Enfin des transpositions sont également possibles pour des manipulations avancées, ou des représentations. Les recherches récentes s'intéressent aux possibilités d'obtenir des cubes OLAP de graphes (Graph OLAP)(Zhao *et al.*, 2011; Gómez *et al.*, 2017), c'est à dire des graphes dépendant de dimension ou d'échelle (ex : un réseau de scientifiques à telle époque, ou tel lieu). Dans l'approche inverse, ce sont des graphes localement enrichis par des cubes (Jakawat, 2016). D'autres approches sont fortement orientées vers les *Geographical Information System*, système d'information géographique. Base de données spécialement structurées pour enregistrer des informations spatiales et opérer des transformations dessus. (systèmes d'information géographique).

2.1.3 Modèle RDF et bases de données graphe

Le modèle *Ressource Description Framework* (RDF) organise les données, non plus en couple clé-valeur (structure relationnelle), mais en triplet (voir exemple 1.4). Des descripteurs réticulés utilisent cette notion de triplet (voir section 3). Rien n'empêche les triplets d'être stockés dans des tables. Cependant des formes plus performantes sont disponibles pour stocker ces données, ces formes sont associées à d'autres types de requêtes, avec leurs langages. Concrètement un triplet RDF établit un lien entre 2 entités. Lorsque des entités sont communes à plusieurs triplets, les données forment naturellement des graphes. La figure 1.6 illustre cette opération. Les 2 options principales sont les *triplestore* (*triple* pour triplet) ou les bases de données objets

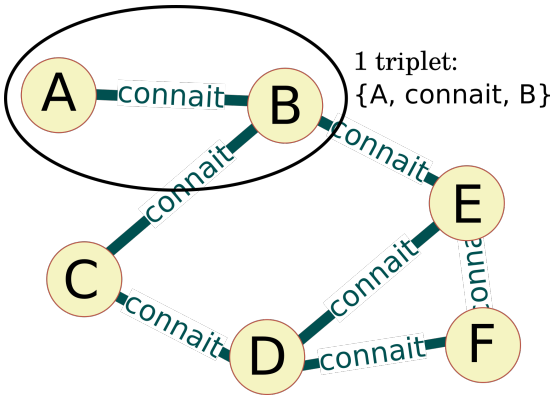


FIGURE 1.6 – Un ensemble de triplets partageant des entités communes forme un graphe

orientées graphe (système clé-valeur). Les bases de données graphe, systèmes en rupture avec les formes classiques de structure de données (SQL) font partie de la grande famille des bases « no-SQL ». Elles permettent de stocker rapidement les données, et surtout de requêtes efficaces et simples, évitant les jointures entre tables (inhérents au langage SQL) et intégrant nativement des éléments de théorie des graphes. Certaines n'ont pas besoin de modélisation de données en amont (*no-schema*). Néanmoins un schéma de données peut guider la mise en œuvre du graphe, et potentiellement la contraindre. Les *triplestore* sont des formes particulières de base de données graphe, ils ne peuvent contenir que des triplets RDF donc des Un graphe binaire est un graphe où chaque entité est lié par un arc non-pondéré à d'autres entités first (graphes binaires).

Les langages de requête dépendent des technologies utilisées. Par exemple *Neo4J* leader des bases graphe utilise *Cypher* son propre langage, *OrientBD* utilise une version étendue du langage SQL pour éviter les jointures en requêtant les arcs du graph directement, *ArangoDB* largement utilisée permet la coexistence de graphes et de clé-valeurs et propose des requêtes par le langage GraphQL (langage developpé par facebook), etc. Les *triplestore* (ex : *Virtuoso*) utilisent le langage de requête pour serveurs de triplets RDF (SPARQL). En fonction des technologies et des applications, les *benchmarks* donnent un avantage mitigé aux bases graphe pour traiter les données (même graphes) par rapport aux bases relationnelles, cependant il est certain que la construction des requêtes est simplifiée.

2.2 Modèles et descripteurs de données pour le patrimoine

Les HN tendent à structurer leurs données en connaissances explicites. Pour cela les systèmes évoqués en section 2.1.1 utilisent des descripteurs. Certains descripteurs sont standards ou très utilisés, d'autres sont développés à des fins uniques. L'usage de descripteurs standards permet une meilleure diffusion des données et une mutualisation des efforts. Certaines initiatives sont extensibles, il est alors possible de prendre tout ou partie des éléments déjà définis et de rajouter des éléments permettant une description plus fine, liée à un contexte d'utilisation précis.

On peut classifier ces systèmes en 3 types : plats, hiérarchisés, réticulés (Hodge, 2000). Ces systèmes concernent soit la structure, soit les valeurs (notice d'autorité). Le tableau 1.2 donne des exemples de ces possibilités.

	Plat	Hierarchique	Réticulés
Structure	Formulaire	Taxonomie	Ontologie
Valeur	Dictionnaire vocabulaire contrôlé	Classification Catégorisation	Thésaurus

TABLE 1.2 – Formalisation des connaissances : modèles plat, hiérarchique ou réticulé indiquant une structure ou des valeurs.

Outre les valeurs et la structure, on considère généralement un troisième niveau, qui ne rentre pas dans ce tableau. Il s'agit du niveau dit de contenu (*content metadata*). Ce niveau concerne les recommandations, les bonnes pratiques et les normes de description. Ce niveau ne traite pas des normes de représentation des métadonnées (ISO_11179, registre de métadonnées). Nous ne traitons pas en détail de ces éléments ici. Ils sont peu nombreux, bien suivis et suffisants. Ils n'indiquent pas nécessairement

le « comment » (choix technologique), mais systématiquement le « quoi » (ce qu'il faut décrire). Parmi les plus suivis on trouve *Cataloging Cultural Objects* (CCO) de la Bibliothèque du Congrès (EUA) qui explique comment décrire un objet du patrimoine culturel, *General International Standard Archival Description* (ISAD(G)) du Conseil international des archives (ICA) qui définit les éléments à inclure pour la description d'une archive.

Le tableau en Annexe A reprend la majorité des schémas de métadonnées utilisés dans le monde du patrimoine et de l'histoire.

2.2.1 Descripteurs plats

Marquant les territoires de la connaissance, délimitant clairement les formes, en rupture avec les habituelles esquisses floues (connaissances décrites en phrases), les connaissances doivent être représentées en conformité avec des modèles. Soit manuellement extraites, soit issues de NERC, les données sont inscrites dans des schémas de *structure de métadonnées* (les formulaires), souvent contraintes par des schémas de valeurs (*values*, les réponses possibles) : par exemple « titre » (structure) : « MonaLisa » (valeur) ; « type » (structure) : « peinture » (valeur).

Valeur. Les notices d'autorité plates concernent principalement les noms propres : personnes, lieux, œuvres. Il est possible d'utiliser un schéma de valeurs (notice d'autorité) plat pour renseigner les valeurs d'une structure de données réticulée. Cette pratique est habituelle dans le cas du web sémantique. Toutes les institutions majeures produisent des notices d'autorité souvent nationales pour les noms propres : *Library of Congress*, *BnF*, *Deutsche Nationalbibliothek*, etc. Le projet *Virtual International Authority File* (VIAF) (Bourdon et Boulet, 2015) débuté en 2003 est une méta-notice d'autorité agrégeant 61 notices (issues de 50 pays) alignées. Ce projet réduit environ 130M de notices d'autorités à 20M de notices distinctes en alignant les notices désignant le même objet et en proposant une *Uniform Resource Identifier* (URI).

Structure. Dès les années 1960, MARC pour les bibliothèques, puis à partir de 1990, DublinCore (ISO_158366, 15 champs de métadonnées), pour les objets culturels, s'imposent comme schéma de structure de référence. Le développement de descripteurs plus complets (davantage de champs) est abandonné (ex : abandon en 2008 de DublinCore Qualifié) avec le passage massif aux descripteurs réticulés (section 2.2.3).

```
titre: La draisienne
date: 1817
créateur: Karl Drais von Sauerbronn
Sujet: mécanique
```

Exemple 1.1 – Exemple d'enregistrement DublinCore d'une draisienne

Récemment des outils performants de moissonnage (comme ISIDORE en France) parcourent le web et centralisent les contenus décris par des descripteurs simples (Protocole *Open Archives Initiative Protocol for Metadata Harvesting* (OAI-PMH) basé sur les 15 métadonnées DublinCore de base). On retrouve alors la même idée, évoquée plus loin au sujet du "Web sémantique" (2.2.3) : internet sert de base de données distribuée. Ces descripteurs plats sont toujours largement utilisés (surtout DublinCore) et ont inspiré de nombreux autres schémas.

2.2.2 Hiérarchique

Les taxonomies concernent moins les données historiques qui se tournent aujourd'hui vers les ontologies. Les archives et bibliothèques ont développé 2 standards qui renseignent la structure et les valeurs : La classification Décimale de Dewey (CDD) est toujours en utilisation dans le monde entier (bibliothèques) complétée par la classification décimale universelle (CDU) en 1905 (complétion peu utilisée) ; La classification de la Bibliothèque du Congrès, surtout en vigueur aux USA. Le fonctionnement est établi en classe (10 dans Dewey), sous classes (100 dans Dewey) et sous-sous-classes (1000 dans Dewey davantage en CDU) avec des possibilités de modifier les classes par des indices de tables auxiliaires comme sur l'exemple 1.2.

En pratique les institutions n'emploient pas de catégories si fines. Il a été prouvé que ce niveau de détails était contre-productif dans les bibliothèques : des items similaires sont classés très différemment. La pratique actuelle consiste à classer simplement (pour les documentalistes) et à associer l'item avec une liste de mots-clés (appelée *indexation*) destinée à l'utilisateur (qui ne connaît pas la classification).

2.2.3 Réticulés

Ces modèles transforment le fonctionnement des bases de connaissances. Les modèles de données réticulés permettent de stocker des instances d'entités, leurs classes, les liens qu'elles tissent avec d'autres entités et des classes de liens. Les bases de connaissances du monde entier ont abondamment utilisé ces modèles depuis les années 2010 pour sémantiser les contenus et augmenter la densité d'information (les liens entre entités, la structuration en classes). Le développement de bases orientées graphe (*graph databases*) a permis de prendre en charge ce type de données à des échelles très larges, avec une modélisation de

```
# sujet: Chimie du solide
500: Chimie
 541: Chimie Physique
  541.4: Sujets particuliers de la chimie physique
   541.42: Chimie des états particuliers de la matière
    541.042 1 :Chimie de l'état solide

# aspect1: histoire au XXe siècle
-09: Table auxiliaire des études historiques
 --0904: XXe siècle

#Indice final
541.0421 904
```

Exemple 1.2 – Classification Dewey d'une thèse sur l'histoire de la chimie du solide au XX^{e} siècle

données plus souple en amont. Les projets majeurs de bases de connaissances sont DBpedia et Wikidata ; mais aussi *Knowledge Graph* de Google, qui a récemment phagocyté de nombreux projets majeurs comme *Freebase* et contient aujourd'hui près de 100 milliards de faits.

Thésaurus. Un thésaurus est un vocabulaire contrôlé organisé par différents liens entre les termes : hiérarchie (spécification/-généralisation de termes), synonymie, association, domaine connexe, etc. Ils permettent d'organiser les termes d'un domaine de connaissance. Dans le domaine du patrimoine et des bibliothèques en France, le thésaurus RAMEAU fait autorité. Il complète les notices d'autorités (souvent plates) de noms propres. Les noms communs sont plus propices à la création de thésaurus que les noms propres, mais l'alignement des notices d'une langue à l'autre est plus compliqué (voire impossible). Les termes associés permettent de rompre certaines barrières d'indexation, la hiérarchie (skos : broader et skos : narrower) permet une précision de l'indexation élevée. Les termes sont souvent associés à une URI pour leur utilisation.

```
<http://data.bnf.fr/ark:/12148/cb122349910> a skos:Concept ;
skos:prefLabel "Métiers à filer"@fr ;#terme en question
skos:altLabel #autres dénominations
    "Filage -- Machines"@fr,
    "Machines à filer"@fr,
    "Matériel de filature"@fr ;
skos:broader <http://data.bnf.fr/ark:/12148/cb119469118> ;
    #Machines textiles
skos:closeMatch <http://dewey.info/class/600/> ; #Technologie
skos:narrower <http://data.bnf.fr/ark:/12148/cb155800549> ; #Rouets
skos:related
    <http://data.bnf.fr/ark:/12148/cb11942815d>, #Fileuse
    <http://data.bnf.fr/ark:/12148/cb16186325f> . #Filature
```

Exemple 1.3 – Extrait commenté d'une notice rameau écrite en n3

Ontologies. Définies dès les années 1970, les ontologies sont des schémas de structure plus complets que les schémas plats et les taxonomies. Les schémas (et implémentations) se multiplient à l'aube des années 2000 pour décrire les données et les liens qu'elles tissent entre elles. Prétendant à formaliser toute connaissance, favorisant alors le partage (lecture des données par les ordinateurs) et permettant de raisonner à partir de règles, les *ontologies*(graphes RDF, modèles *Web Ontology Language* (OWL)_2) s'imposent pour décrire les données des Sciences Humaines et Sociales (SHS) (Illien *et al.*, 2013; Sinclair *et al.*, 2006). La rupture avec les schémas hiérarchiques (et/ou plats) consiste à introduire les liens que l'objet tisse avec d'autres objets, en plus de ses propriétés intrinsèques. Les données sont décrites en triplets, selon le modèle RDF (Ressource Description Framework) : sujet, prédicat objet. L'exemple 1.4 illustre ce fonctionnement.

La figure 1.7 illustre l'enjeu de la structuration des données en ontologie. Dans ce cadre il est possible d'effectuer de requête sur le lien entre les 2 éléments. Par ailleurs, d'autres liens implicites existent entre ces éléments, ils sont également modélisés : la figure 1.8 montre les liens entre 3 instances d'entités (classes) et la hiérarchie de ces entités, cette complexité est gérée par les inférences de l'ontologie (arbre d'héritage). Les propriétés également sont organisées selon un arbre d'héritage.
Une partie importante du travail de l'historien est de lier/recouper des sources d'information. On observe une similarité avec les opérations de structuration en ontologie.

Matthieu Quantin	est auteur de	cette thèse			
(sujet)	(prédicat)	(objet)			
		cette thèse	est écrite en	français	
		(sujet)	(prédicat)	(objet)	

Exemple 1.4 – exemple RDF : découpage d'une phrase en triplets

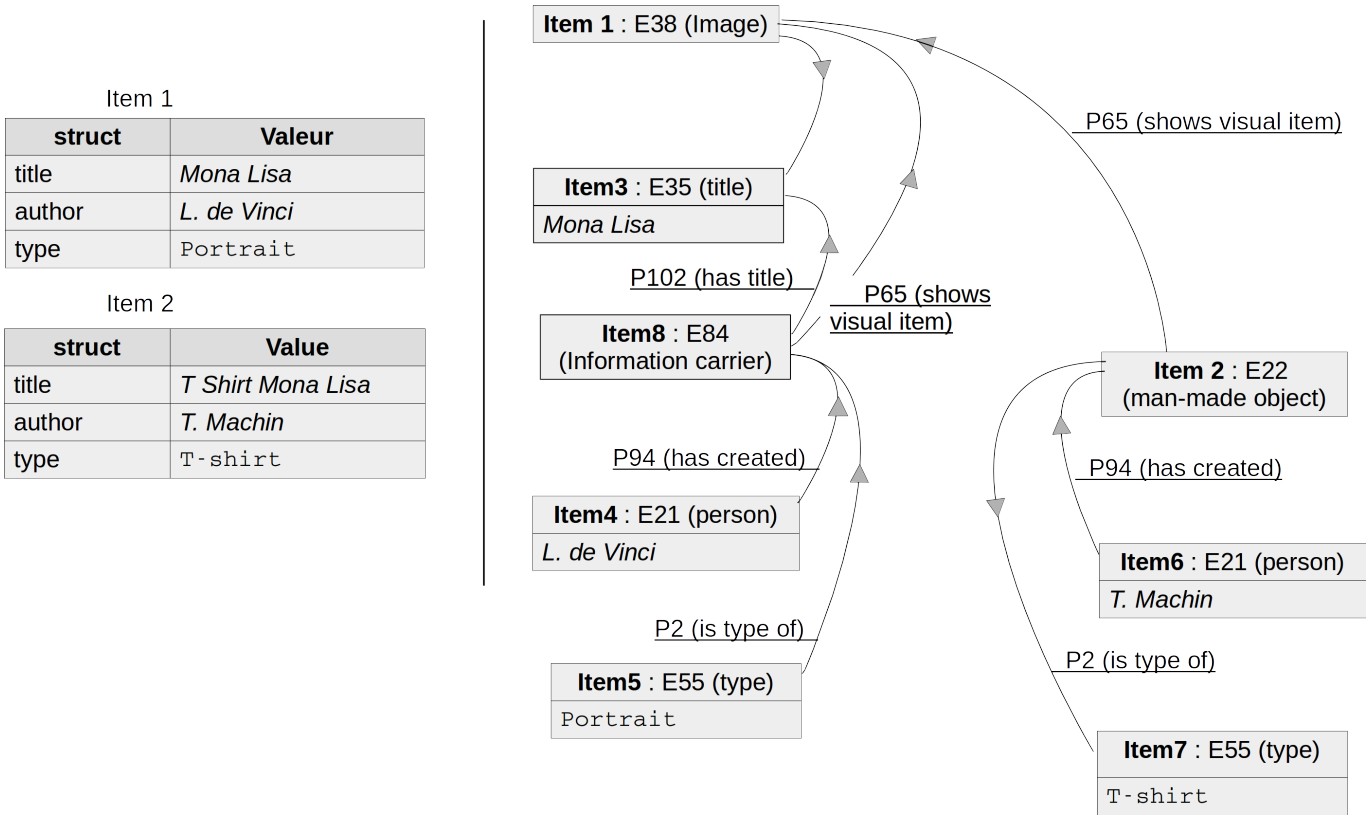


FIGURE 1.7 – 2 items : la Joconde et un T-shirt de la Joconde. À gauche avec Dublin Core (plat), à droite avec CIDOC-CRM (ontologie). Les valeurs en police à chasse fixe sont issues de schéma de valeurs (ex : thésaurus). Les valeurs en italiques sont des chaînes de caractères.

Les ontologies floues. Les ontologies floues (Bobillo et Straccia, 2011) résolvent une limitation des ontologies dites binaires en permettant d'attribuer une pondération à une relation entre entités ou propriété d'une entité. Cette approche réconcilie les ontologies avec la notion de nuance d'une relation en science humaine et améliore la précision en augmentant les possibles (flous). Par exemple il est possible de décrire une propriété non plus comme une valeur unique mais comme une quantité, voire une pondération (exemple 1.6). Tout l'enjeu étant de définir les opérations possibles sur ces valeurs et de créer les raisonneurs (Bobillo et Straccia, 2016). Ce fonctionnement est développé pour répondre à des contraintes de quantification (ex : prix d'un objet), ce

```
:className rdf:type owl:Class ;
:annotationProperty <annotationValue>
```

Exemple 1.5 – Exemple de propriété floue d'une instance, exprimé en Turtle

fonctionnement ne remet pas en cause les triplets énoncés précédemment.

La quantification de relation entre entités est plus compliqué et peut remettre en cause la notion de triplet RDF pour optimiser les raisonneurs. Par exemple, renseigner le degré de proximité entre 2 personnes, au lieu de simplement signaler qu'elles se connaissent ou non. Il faut déclarer une nouvelle classe ayant plusieurs propriétés pour conserver la réification en triplets RDF, cette expression est lourde en syntaxe *Extensible Markup Language* (XML), mais abordable en Turtle (voir exemple 1.7 illustré en figure 1.9). Une autre possibilité, lorsque toutes les relations sont pondérées, est de construire des quadruplets Lopes *et al.* (2010), pour une application locale, détournés de leur objectif officiel (W3C recommendation) de sourcer le graphe dont est issu le triplet Carothers (2014). Il faut ensuite déterminer si cette quantification est réciproque (ce n'est pas le cas en figure 1.9), d'éventuelles relations d'inclusion (ex : `worksWithLevel` $\subset$ `worksWith`), des moyens de comparaison (dans le cas de valeurs plus abstraites comme « un peu », « beaucoup », etc.).

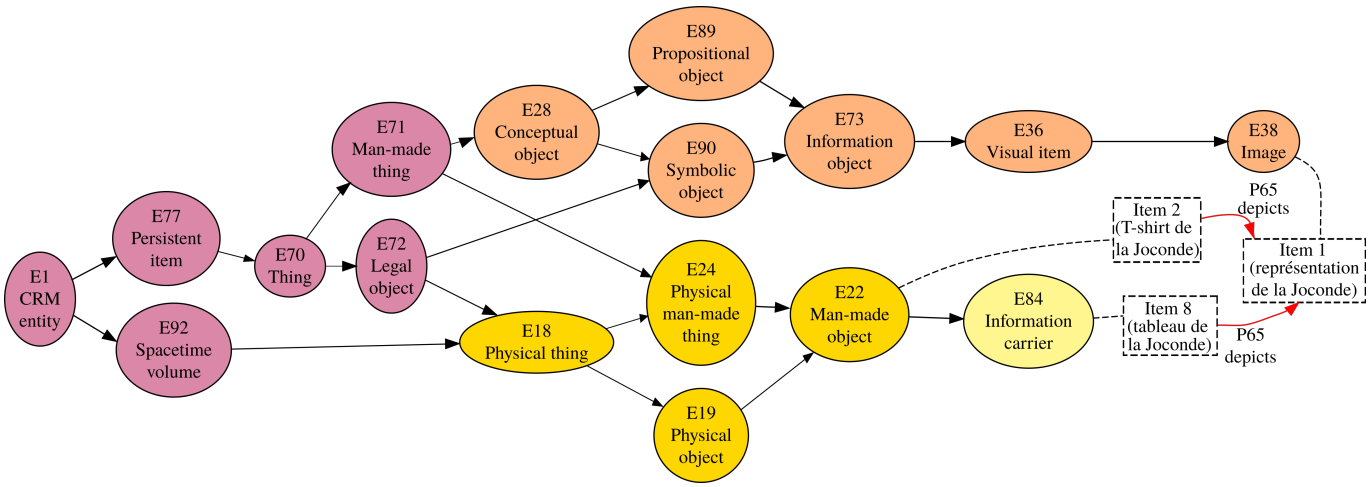


FIGURE 1.8 – Représentation graphique de l'arbre d'héritage pour les classes des instances Joconde (tableau), T-shirt de la Joconde et Joconde (représentation).

```
myOnto:Athlete rdf:type owl:Class;
myOnto:estFort xsd:integer.

<#John>
a myOnto:Athlete;
myOnto:estFort 80.
```

Exemple 1.6 – Exemple précédent instancié pour désigner la force d'un athlète

Web sémantique Des standards du monde de l'histoire et du patrimoine émergent. Ils sont interopérables avec les grands standards internationaux pour les données liées sur le web : CIDOC-CRM (Crofts *et al.*, 1999) avec *DublinCore Metadata Initiative (DCMI) Terms* (DCterms), *Friend of a frend* (FoaF), etc (Doerr, 2003). Les données peuvent alors être décrites par des schémas interopérables. L'idée du web sémantique est de stocker ces données en ligne et d'en permettre l'accès de partout, éventuellement à des machines pour formater des contenus. Les instances sont alors identifiées URL pérennes : les URI[3]. Les instances ne sont alors plus « dupliquées » ou « recopiées » mais pointées via le web, on parle de références (nœuds communs). Les instances de référence sont gérées par les autorités : par exemple la BNF. Ainsi liées, les données donnent naissance à des graphes distribués sur plusieurs systèmes (serveurs). Ce chaînage de données utilisant le même formalisme et les mêmes références permet la diffusion et le partage.
La figure 1.10 représente cette idée : les items sont des URI, ce sont nécessairement des entités, dont les types sont potentiellement issus de plusieurs vocabulaires (CIDOC-CRM, FoaF, DCterms, etc.). Ces items sont liés par des propriétés également issues de plusieurs vocabulaires. Les URI sont potentiellement hébergées dans des entrepôts différents (dataBNF, DBpedia, à la maison, etc.).

Le web est alors utilisé comme base de données distribuée contenant des connaissances explicites, dit web sémantique. Ces graphes de données et les propriétés associées (Blue *et al.*, 2002) sont largement étudiés par différentes disciplines (biologie,

3. une url perenne décrivant une ressource, par exemple sur la figure 1.7, pour désigner `item1`, on utilise `http://data.bnf.fr/ark:/12148/cb11944800v`

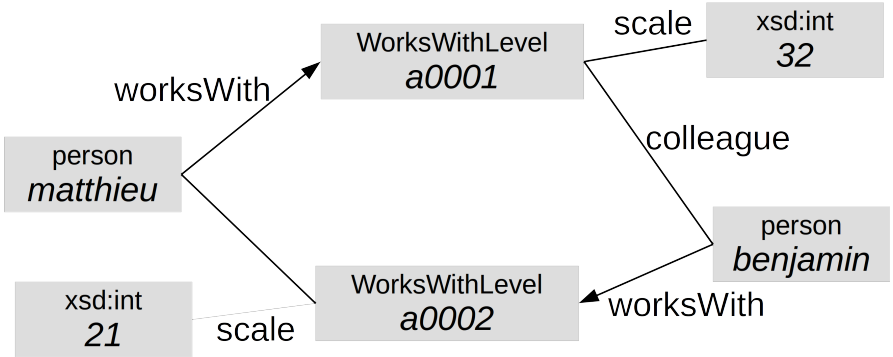


FIGURE 1.9 – Représentation d'une relation floue non symétrique entre 2 entités en réification RDF

```
myInst:matthieu myOnto:worksWith myInst:a0001 .
myInst:a0001 a myOnto:worksWithLevel;
   myOnto:colleague myInst:benjamin;
   myOnto:scale 32.
```

Exemple 1.7 – Exemple d'une relation floue entre instances, exprimée en Turtle

mathématiques, informatique). Ces études mènent à un second grand principe des ontologies (et donc du web sémantique) : la capacité à raisonner sur les données à partir de règles pré-établies.

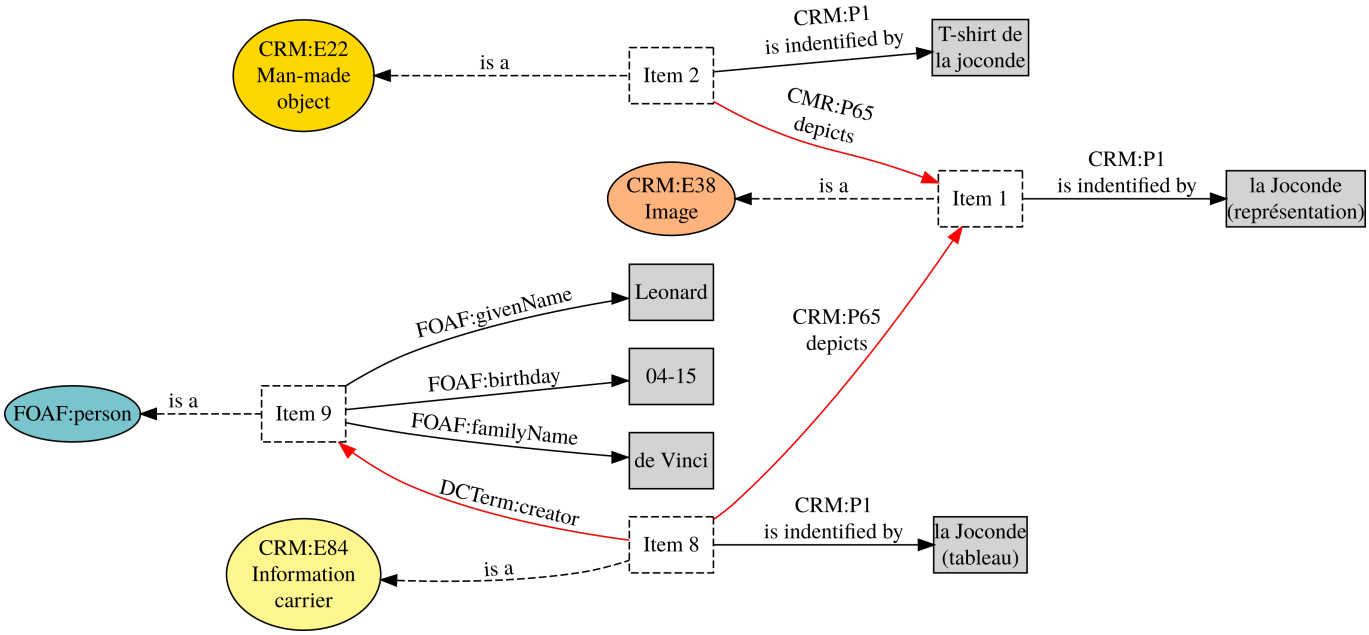


FIGURE 1.10 – Représentation de l'idée du web sémantique : URI liées par des vocabulaires standards. En carré pointillé les URI, en carré gris les chaînes de caractères, en rond les entités, les flèches noires sont des propriétés avec des chaînes de caractères, en rouge avec d'autre URI.

La plupart des territoires des HN sont touchés par ce mouvement de structuration en ontologies (Rapti *et al.*, 2015) : des grands projets cités précédemment jusqu'aux petites initiatives locales (Srinivasan et Huang, 2005; Damova et Dannells, 2011). Ces dernières sont souvent attirées par l'idée d'inférences et la volonté de s'afficher détenteur/contributeur d'un savoir formalisé, d'apporter leur brique à la tour de Babel contemporaine : le *Giant Global Graph* (Berners-Lee, 2007). Ces techniques articulées autour des ontologies constituent la clé d'une typologie empirique des humanités numériques identifiée par Heimburger et Ruiz (2012) : (1) l'émergence de pratiques documentaires originales, (2) l'apparition de nouveaux modes de diffusion de la recherche et (3) la naissance de formes inédites d'échanges scientifiques et pédagogiques. Je fais remarquer que l'analyse n'est pas une application citée. Comme précédemment écrit, nous nous nous focalisons l'analyse de corpus avec un historien.

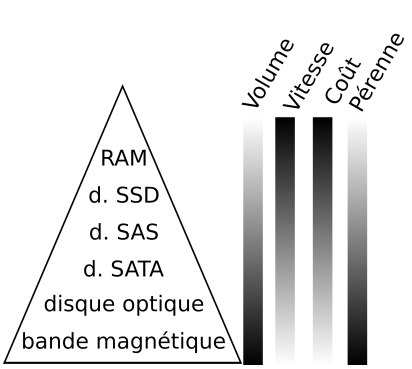
2.3 Outils et méthodes de gestion des connaissances pour le patrimoine

2.3.1 Archivage perenne

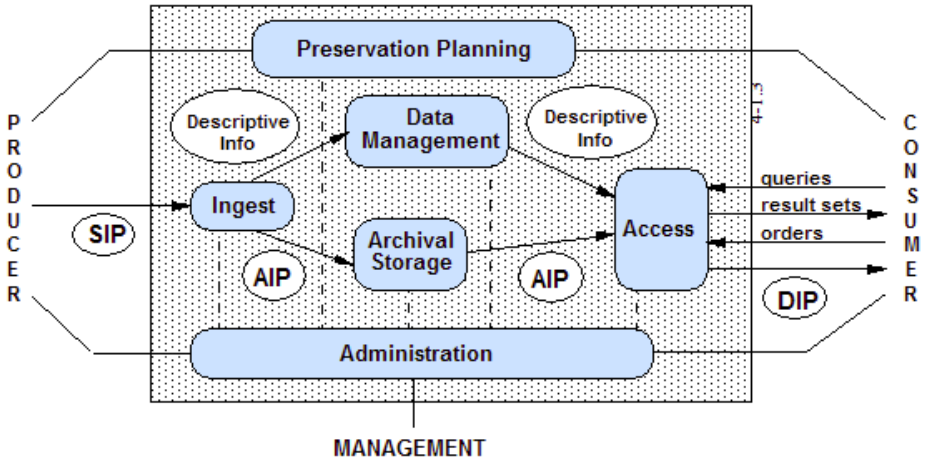
L'archivage pérenne n'assure pas uniquement la conservation de données au sens défini en section 2.1 : sauvegarde d'un train de bit. En effet, il faut assurer les possibilités de lecture c'est-à-dire assurer la pérennité des informations. Les contraintes sont donc :

— Conservation physique de l'information, Gestionnaire de stockage hiérarchique (HSM) avec une contrainte de pérennité (voir 1.11a).
— Conservation des accès (moyens) à l'information : imposer l'usage de formats standards publiés (ouverts) et faire évoluer les informations archivées avec l'évolution du standard. Une autre option plus permissive (en test) consiste à conserver les logiciels et environnement d'exécution. Il s'agit d'un projet UNESCO et INRIA (Zacchiroli, 2017).
— Conserver le sens de l'information : qui, quoi, comment, etc. (voir section 3)

Le dépôt d'une archive doit être réalisé selon le modèle Open Archival Information System (OAIS) (ISO 14721 :2012) . Dans sa dernière version (2012), le processus de dépôt implique l'élaboration de plusieurs éléments de description (*package information*) à différents stades de l'archive.



(a) Pyramide des supports, le temps d'accès est inversement proportionnel à la pérennité.



(b) Modèle OAIS d'un dépôt d'archive

FIGURE 1.11 – Archivage pérenne : support matériel et description des données

Software	Développeurs	Fonctionnement
Actimuseo	A&A partners	Monoposte
Webmuseo	A&A partners	Version entièrement web des services de Actimusée
EMu	Axiell	Fonctionnement local client-serveur
Flora-Musee	Decalog	Outil entièrement Web. Mutualisation de musées
Gcoll	Videomuseum	Développé pour Joconde
Micromusee V6	Mobydoc	Monoposte + module de publication web
S-Museum	SKINsoft	Outil entièrement Web

TABLE 1.3 – Principaux outils de gestion des collections en musée

— *submission information package* (SIP) lors de la soumission. Une partie commune à tout dépôt concerne la gestion administrative (droits, responsable, durée, etc.). Des parties variables en fonction des contenus et des gestionnaires de dépôt peuvent être exigées.
— *archival information package* (AIP) pour les informations archivées, généré en interne. C'est le contenu et la description du contenu du dépôt : les données, et leurs descripteurs (SIP et autres métadonnées générées en interne).
— *dissemination information package* (DIP) lors d'une requête (coté utilisateur). Il contient les informations en fonction des droits de diffusion et droits d'accès de l'utilisateur qui fait la requête.

Certains schémas de métadonnées (voir section 3) sont standards pour les descriptions, compatibles avec le protocole OAIS :

— L'organisation : *metadata encoding and transmission standard* (METS) ou description archivistique encodée (EAD). Ces schémas gèrent les dépendances structurelles (hierarchie des éléments).
— Les contenus : DublinCore décrit minimalement chaque objet de l'archive
— les relations entre objets de l'archive sont décrites par le schéma nommé stratégies d'implémentation de métadonnées de présevation (PREMIS), qui gère les dépendances sémantiques (référence, rôle de l'agent, etc.)

2.3.2 Outils de gestion du patrimoine 2D

Nous nous intéressons maintenant aux outils existants destinés à la gestion des données patrimoniales. Les institutions sont massivement équipées. Ce sujet est déjà largement étudié par ailleurs (Yakel *et al.*, 2011). Trois éléments me semblent primordiaux : les outils de gestion officiels pour les collections de musées, les outils de gestion de collection hors institution (ou en parallèle aux outils officiels), les outils de gestion spécifiques à des besoins de musées. De très nombreuses applications pour la gestion des collections existent, la plupart intègrent les vocabulaires standards du monde patrimonial (cf section 2.2 et annexe A).

Outils officiels À ces outils de projets et valorisation, il faut ajouter les outils officiels de gestion des collections. Ces outils sont complémentaires aux outils précédents. Le service des musées de France établit une liste d'outils de gestion qualifiés (validés) pour la gestion des collections. Ces outils présentés dans le tableau 1.3 sont compatibles avec les bases classiques comme *Joconde*, le Récolement décennal et l'inventaire réglementaire (technique délimité par l'Arrêté du 25 mai 2004).

Outils spécifiques. À ces 2 catégories précédentes (outil de gestion hors institution et outils officiels) il faut ajouter tous les outils spécifiques développés pour certains projets, parfois au sein d'un musée. Certains fonctionnent avec les vocabulaires standardisés, d'autres sont encore plus spécifiques. C'est le cas du projet Nantes1900 (Hervy *et al.*, 2012) pour l'analyse et la valorisation d'un objet important (maquette) du musée d'Histoire de la ville de Nantes, ainsi que d'autres collections comme *Patstec* (Patrimoine Scientifique et Technique Contemporain) ou la base *Recital* (REgistres de la Comédie-ITALienne) qui utilise des outils de transcription en ligne à des fin d'analyse d'un corpus initialement détenu par la Bibliothèque Nationale de France (BNF). Certains de ces outils très orientés vers l'analyse intègrent des fonctionnalités OLAP (*On Line Analytical Processing*) permettant d'analyser les données sous plusieurs dimensions et d'effectuer des opérations sur celles-ci (Chaudhuri et Dayal, 1997).

Les CMS, outil du foisonnement hors-institutions. Un récent foisonnement de données patrimoniales est (en partie) dû à l'existence de nombreux frameworks web (Content Manager Systems, CMS) open-source, au développement très actif, à l'initiative (et soutenu par) des instituts de recherche pour la gestion des données patrimoniales. SyMoGIH (Beretta et Vernus, 2012) basé sur Drupal s'apparente à un CMS dédié à l'histoire. Ces outils sont très orientés diffusion, ils nous intéressent ici

CMS	Institution	Version	Licence	Date	Activité	Note
Omeka	R. Rosenzweig Center ; G. Mason Univ.	2.5 + S	GPLv3	2007 – now	++	Nombreux plug-in, actif en France
Arches project	The Getti Conservation Instit.	4	AGPLv3	2011 – now	++	Basé Django (Python). Orienté espace-temps (SIG intégré)
AtoM (Access to Memory)	International Council on Archives	?	AGPLv3	2007 – now	+	
CollectiveAccess	5 universités (EU, USA)	1.7 + 2	GPLv3	2012 – now	++	Front-end / Back-end séparés, compatible avec omeka front-end
Archon	Univ. of Illinois	2	Illinois Open Source License	2004 – 2014	+	abandon de l'institution, projet open-source
Mukurtu	Washington State Univ.	2	GPLv3	2007 – now	$\sim$	
SyMoGIH	LARHRA (Lyon)	?	?	2012 – now	+	SaaS, pas de téléchargement

TABLE 1.4 – Présentation de frameworks CMS permettant de gérer les collections patrimoniales, facilement déployables. Plusieurs milliers d'instances de ces CMS sont actuellement en ligne. Sans être spécifiques aux humanités, d'autres CMS sont courrament utilisés (ex : Drupal).

pour leurs qualités en gestion des données. Les CMS listés dans le tableau 1.4 offrent de nombreuses fonctionnalités présentées précédemment : interface pour base de données, structure de métadonnées standards, thésaurus, dépôt OAI-PMH, triple-store, édition de TEI, etc. Le protocole OAI-PMH permet le moissonnage des données produites à l'échelle locale (via métadonnées DublinCore au minimum). Ce moissonnage permet de mutualiser l'ensemble des productions à une échelle plus globale. Par exemple : le projet ISIDORE de la TGIR Huma-Num rassemble essentiellement des documents historiques, ou encore le projet « videomuseum » qui rassemble les collections d'art contemporain. Enfin pour des raisons techniques (volume, débit, etc.) la 3D est peu présente sur internet.

2.3.3 La recherche dédiée au patrimoine numérique en 3D

Le sujet de la 3D pour le patrimoine a déjà été abordé par de nombreux auteurs, sous de nombreux angles. Les problématiques abordées sont celles des enjeux pour la recherche dans le domaine du patrimoine (Koller *et al.*, 2009), des méthodes et technologies (Pavlidis *et al.*, 2007; Remondino, 2011; Akca, 2017), des usages : l'analyse et la reconstitution associées à des connaissances (Stanco *et al.*, 2011), des bonnes pratiques d'enregistrement (Remondino et Campana, 2014), de la rétro-conception mécanique (Laroche *et al.*, 2008), de la reconstitution (Gomes *et al.*, 2014), de l'annotation sémantique via les sources (Manuel *et al.*, 2016). Au-delà, de nombreux états de l'art s'intéressent à la captation 3D, domaine qui dépasse largement celui du patrimoine.

Cette thèse ne traite pas de 3D directement, néanmoins, une des finalités est la gestion et l'analyse de la documentation du patrimoine. Nous étudions ici les approches qui n'ont pas été abordées par la thèse de Hervy *et al.* (2014).

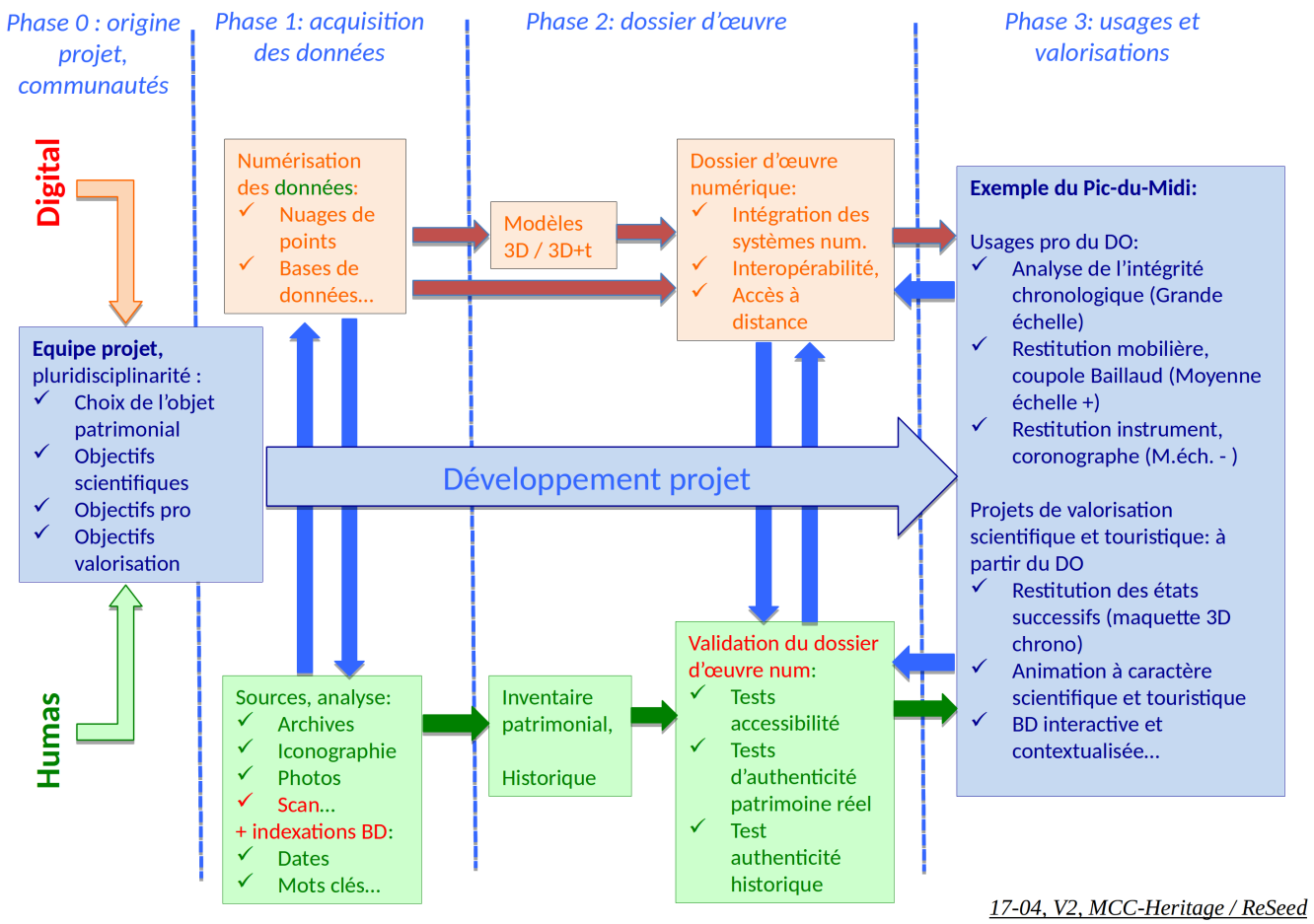


FIGURE 1.12 – Schéma des fonctionnalités nécessaires pour un modèle 3D utilisable pour l'analyse de la chaîne de valeur du patrimoine, Michel Cotte (2017)

Enjeux. L'analyse du patrimoine par les outils informatiques de manipulation de données 3D se retrouve dans plusieurs disciplines, l'avantage principal de cette technologie réside dans la manipulation des objets enregistrés (Nicolas *et al.*, 2016).
Mais selon Cotte (2009); Scopigno *et al.* (2011) la gestion du patrimoine 3D implique surtout une gestion des connaissances associées (documentation, méthodes) et des possibilités d'analyse. En effet cela permet de le justifier dans sa valeur ou dans sa reconstitution. Les sources utilisées sont souvent 2D : iconographies, analyses historiques, archives, descriptifs, etc. (cf sections précédentes pour leur gestion et analyse).

Intégration des acteurs dans un projet (interdisciplinaire). Le projet *Nantes1900* a contribué à une formalisation implémentée de ce process dont témoigne le schéma SADT de la figure 1.13

2.3.4 Frameworks pour le patrimoine 3D

Les modèles d'information géographique (SIG) dit « 4D » intègrent une dimension temporelle ((ISO, 2015; De Roo *et al.*, 2013) et concernent plutôt l'archéologie. D'autre framework sont intéressants pour le patrimoine industriel et technique.

HBIM. Pour la gestion du patrimoine bâti, une option plutôt orientée rétro-conception (temps court) réside dans le récent (depuis 2008) standard 3D de l'architecture et du génie civil *Building Information Model* (BIM). Le modèle est étendu pour permettre la documentation du patrimoine sur toutes ses phases de vie : Heritage BIM (*Building Information Model for Heritage* (HBIM)) et prendre en compte les ontologies du patrimoine (voir section "Modèles et descripteurs de données pour le patrimoine" (2.2)) Quattrini *et al.* (2017). Limité au patrimoine bâti, BIM est développé avec le modèle de données *industry foundation class* (IFC). Avec les mêmes défauts que la CAO, il peut difficilement prendre en compte l'imperfection de la réalité. Une dynamique forte se développe autour de l'abstraction des numérisations (nuages de points) vers les modèles volumiques augmentés d'informations sémantiques. La figure 1.14 montre les 2 versants de HBIM : d'un côté les données liées et de l'autre la gestion des volumes (3D).

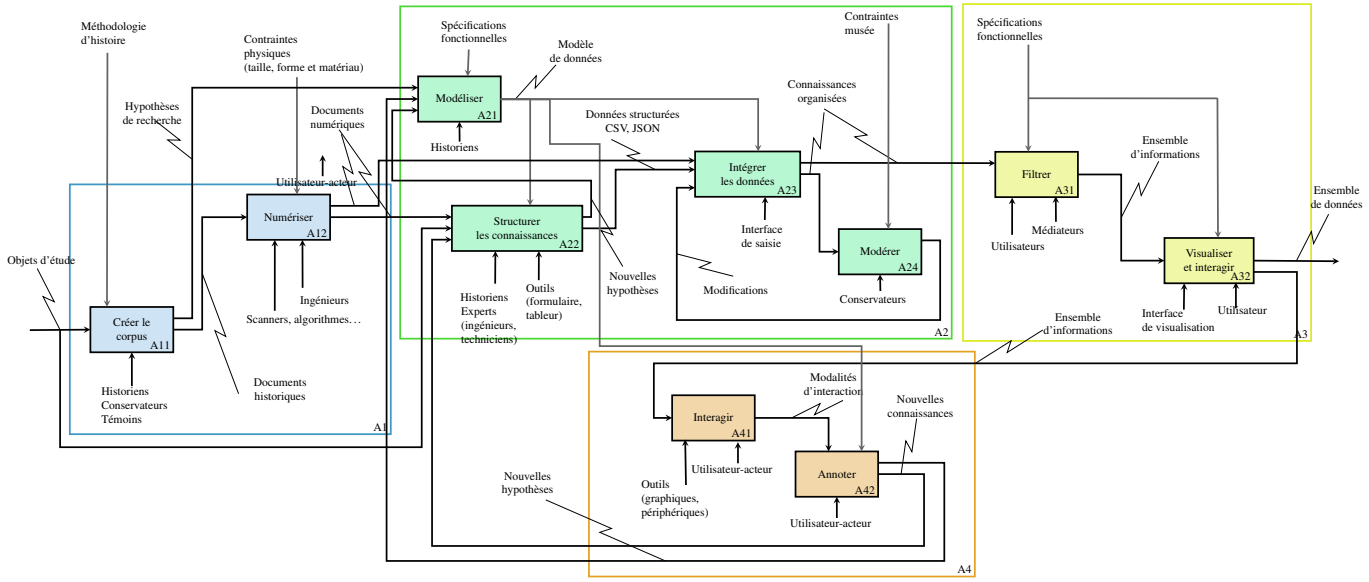
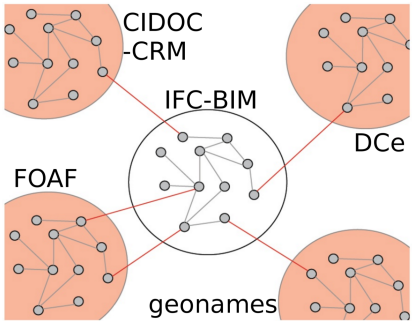
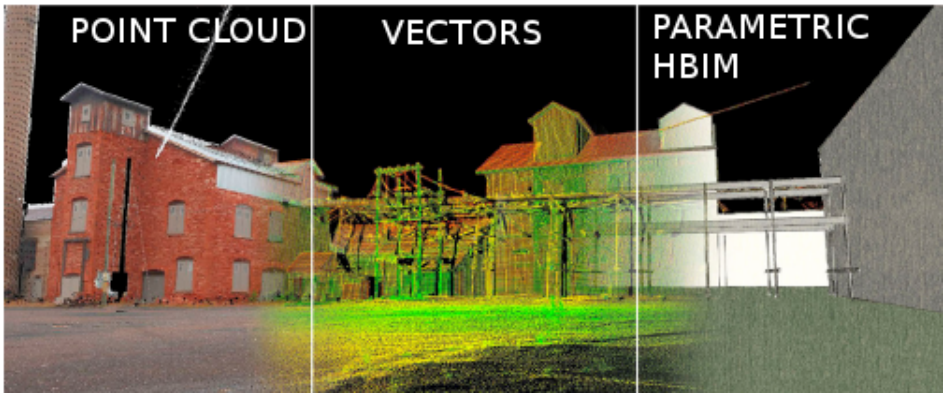


FIGURE 1.13 – Processus de gestion du patrimoine numérique 3D et des données associées (Hervy, 2014)



(a) Liens entre classes IFC et autres ontologies : données liées



(b) Abstraction du nuage de points vers un modèle 3D (Garcia-Fernandez *et al.*, 2015)

FIGURE 1.14 – HBIM : données du patrimoine et gestion des enregistrements 3D

CityGML. CityGML est un schéma d'extension du *Geography Markup Language* (GML), destiné à décrire les villes en 3D à plusieurs échelles. Ce schéma indique des notions de géographie (terrain), de géométrie (forme), d'apparence (texture) et de sémantique via divers attributs. Il est composé de plusieurs classes : `Building_model`, `Appearance_model` par exemple et permet de développer des *Application Domain Extension* (ADE). De nombreux projets récents étendent cityGML, surtout les classes de `building_model` pour décrire du patrimoine bâti en série ou avec un contexte urbain, par exemple pour les toits chinois (Li *et al.*, 2017), pour le suivi de conservation (Zalamea *et al.*, 2013), pour quelques sites UNESCO en Italie, pour détecter des formes (fenêtres, portes, etc.) (Slade *et al.*, 2017) etc.

Extended Matrix. Le framework *Extended Matrix* (Demetrescu et Fanini, 2017) est conçu pour documenter les étapes de reconstitution du patrimoine 3D (très orienté archéologie du bâtiment) et compte aussi s'aligner sur les descripteurs standards du patrimoine (ontologies). Il permet de créer des suites logiques d'opérations d'assemblages d'éléments architecturaux et de modification du bâti (étapes de construction, extensions, destruction partielles, etc.). Les suites d'opération sont décrites schématiquement. Ce type de projet est complémentaire avec les SIG 4D et rationalise l'approche HBIM aux éléments qui intéressent les reconstitutions du patrimoine : assemblages et chronologie.

DHRM - Nantes1900 (Laroche, 2007) dans sa thèse propose un processus d'analyse du patrimoine technique : le *Digital Heritage Reference Model* (DHRM). Il s'agit d'un modèle conceptuel dont la finalité est de gérer l'objet technique numérisé (3D) et rétro-conçu (CAO) avec les informations qui s'y rapportent (usages dans le temps) aussi appelées « contexte ».
Une caractéristique des données patrimoniales réside dans leur relation au temps. L'intégration de données temporelles peut se faire selon 2 modes, le temps long et le temps court :

— Le temps court correspond aux maquettes CAO animées, ou simulations de fonctionnement.

— Le temps long correspond au temps historique, à l'inscription de l'élément dans un contexte, à son développement ou sa construction. On parle alors de rétro-conception.

Cotte (2009) présente un modèle d'analyse basé sur les maquettes virtuelles animées (CAO) pour le patrimoine industriel et technique (temps court). Aujourd'hui il développe un processus intégrant les fonctionnalités nécessaires à chaque étape pour produire un modèle 3D destiné à la recherche dans le monde du patrimoine (figure 1.12). Ce processus intègre les valeurs du patrimoine mondial de l'UNESCO. La 3D permet en effet de saisir les contours, de montrer les manques d'information et de simuler le fonctionnement d'un bien. Cela correspond aux besoins d'évaluation de l'intégrité (section 1.2.1).

2.3.5 Outils pour le patrimoine 3D

Des outils numériques pour manipuler les frameworks sont nombreux. D'autres outils, orientés sur l'analyse et la documentation des modèles 3D pour le patrimoine permettent une meilleure gestion de la documentation, en dehors des grands cadres précédemment cités.

Outils d'analyse. Des outils comme 3D-Systek (Axaridou *et al.*, 2014) nativement conçus pour le patrimoine permettent d'annoter et de documenter des modèles 3D virtuels, particulièrement les objets hors-bâtiments. Seule une gestion du temps long est intégrée. D'autres comme la suite *Nubes* (Vallet *et al.*, 2012) du programme *3D Monument* (destiné au patrimoine architectural) piloté par le MAP GEMSAU, tentent de répondre à ces problématiques de relations entre la représentation de l'édifice (forme, dimensions, état de conservation, restitution hypothétique) et des informations hétérogènes concernant différents domaines (technique, documentaire, historique).

Enfin le projet GISSAR (Desjardin *et al.*, 2012) nous intéresse par son positionnement sur la chaîne d'informations : il prend en compte les incertitudes et établit une typologie des sources d'imperfections des données de la saisie à la 3D. Ce type de travaux établit un pont entre sources historiques et monde patrimonial (avec valorisation) en conservant la qualité de la source (logique floue).

Projets en cours. Le projet Semapolis (Aubry *et al.*, 2015) de l'Agence Nationale de la Recherche (ANR) s'intéresse à la sémantisation de photos urbaines et la construction de modèles 3D sémantisés avec des techniques avancées d'analyse d'images et d'apprentissage à grande échelle.

Le projet H2020 *Inception* (Maietti *et al.*, 2017) (2015-2019) est aussi focalisé sur la modélisation sémantique 3D : faciliter le flux de numérisation 3D et intégrer des connaissances extérieures, gérer l'ensemble suivant différentes dimensions (temps et espace à minima), permettre de diffuser ces données via le web sémantique. La démarche utilise le HBIM. Le projet devra aussi respecter les descripteurs standards avec une visée de diffusion via le web sémantique.

Dans cette même optique, l'ANR ReSeed, davantage indépendant des techniques existantes et des standards, se focalise sur la sémantisation du patrimoine numérique scientifique ou technique. Cette ANR devra prendre en compte la démarche de patrimonialisation et les critères associés (cf. figure 1.12).

2.4 Récapitulatif

Le tableau 1.5 compare les principaux outils disponibles pour la gestion du patrimoine numérique dans des projets hors institution.

Nous identifions que la documentation du patrimoine constitue le nœud de sa valeur tout autant que l'objet lui-même. Nous avons évoqué la gestion des sources du patrimoine, sa documentation figée. C'est à dire une documentation manuellement associée au bien.

Verrou scientifique (1)

Le verrou que nous abordons ici est celui d'une **vision dynamique du patrimoine**. Pour l'instant, si le patrimoine est un objet du présent, un objet contingent, sa documentation est figée. Ceci nous amène à la captation de données capable de documenter le patrimoine dans cette évolution. Nous devons alors traiter de l'extraction de nouvelles données sources, leur structuration pour la documentation du patrimoine, mais aussi la production de nouvelles connaissances (historiques) contextuelles par l'analyse de sources déjà identifiées.

3 Des données textuelles aux connaissances explicites

Riel *et al.* (2008a) montre que la majorité de l'information explicite d'une organisation est disponible sous forme de texte, et que cette situation est stable dans le temps, aucune modification de cette organisation n'est prévue. Le texte est donc une source de prédilection.

D'un côté (sur la gauche de la figure 1.15) une voie de structuration explicite est décrite par cette section (3). Cette voie se décompose en plusieurs étages : d'abord l'extraction pour capter des données et obtenir des informations sur celles-ci ; puis la

		semaPolis	*Symogih*	*3D-SYSTEK*	*3D Icons (pipeline)*	*CMS*	*HBIM*	*Nantes1900 DHRM*	*Extended Matrix*	*City GML*
		outils					frameworks			
Stdard	Dev. actif	🔴	🟠	•	🟠	🟢	🟢	🔴	🟠	🟢
	Utilisable	🔴	🟠	🟠	🟢	🟢	🟢	🔴	🟢	🟢
	Interopérable	•	🟢	•	🟢	🟢	🟠	•	🟠	🟢
Analyse	fonctions type OLAP	🔴	🔴	🟠	🔴	🔴	•	🔴	•	🟠
	Inférences	•	🟠	•	🔴	🔴	🟠	•	🟠	🔴
	Logique floue	🟠	•	•	•	•	•	🔴	•	•
KM	Doc. (sources)	•	🟢	•	🟠	🟢	🔴	🟢	•	🟠
	Multi-echelle	•	•	🔴	🔴	🔴	🟢	🟠	🟠	🟢
	Temps long	•	🟠	🔴	🔴	🟠	🟢	🟠	🟢	🟠
	Géo (GIS)	•	🟢	•	🟠	🟠	🟢	🟢	🔴	🟢
	Collection / série	•	🟢	🟠	🟢	🟢	🔴	🟠	•	🟢
3D	Géométrie (3D)	🟢	•	🟢	🟢	•	🟢	🔴	🔴	🟠
	Simu. (tps court)	•	•	•	•	•	🔴	•	🔴	🟠
	Zonage 3D	🟢	•	🔴	•	•	🟠	🟠	🟠	🟢
	Assemblage	🟠	•	🟠	🟠	•	🔴	🟢	🟢	🟠
	Reco. de forme	🟢	•	•	•	•	•	•	•	🟠
	3D sémantique	🟠	•	🟠	🔴	•	🟠	🟢	🔴	🟠

TABLE 1.5 – Comparatif des fonctionnalités d'outils développés pour la gestion du patrimoine numérique.
• : mauvais ou absent, 🔴 : prise en charge minimale, 🟠 : pris en charge, 🟢 : excellent

structuration pour représenter ces données et raisonner dessus. Cette première voie est une réponse aux espoirs de la « seconde vague » des humanités numériques (voir section 1.1.1). Cette voie considère que la structuration et l'interopérabilité des données permet leur mise en contexte et la production de nouvelles données par raisonnements automatiques. Cette approche s'intéresse à résoudre le problème soulevé par la gestion du patrimoine : Il s'agit surtout d'une gestion de la documentation et des informations associées.

De l'autre côté, un large pan, moins organisé est présenté par la section suivante (section "Analyses des données textuelles et connaissances explicites" (4)). Cette voie concerne essentiellement les méthodes sans apprentissage. Conformément à la figure 1.15, elle peut aussi concerner certaines issues de la voie précédente. Ce pan est un reliquat coriace de la première vague (voir section 4). Son objectif principal est d'outiller l'historien pour produire des analyses approfondies. Les données sont ici vues comme un tout suffisant. L'historien joue alors le rôle principal, assisté par l'ordinateur : interpréter et parfois relier ces données à des connaissances extérieures.

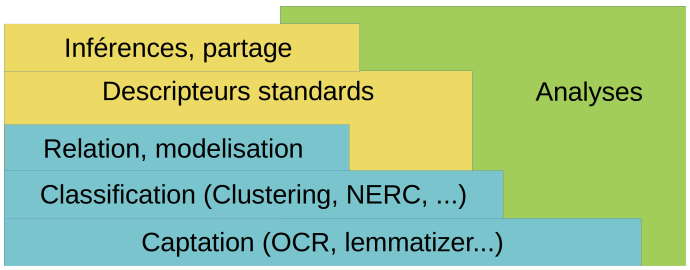


FIGURE 1.15 – Représentation schématique du paysage du NLP en Histoire.

3.1 Captation

Au premier plan, inévitables, les techniques à base d'apprentissage supervisé utilisent des petites quantités de données complètes (ex : caractéristiques d'un patient et maladie(s)) d'abord, pour ensuite déterminer des données incomplètes (par exemple en analyse prédictive : « étant donné les caractéristiques d'un patient, quel risque présente telle maladie »).

Ce premier plan ouvre l'horizon des HN : elles permettent de capter les données brutes. Les techniques classiques utilisées (modèles de Markov cachés, réseaux de neurones, arbres de décisions, plus proches voisins (kNN)...) servent à l'extraction de contenus formatés : discours oral, registres, tableaux, caractères, parties de textes. Elles servent à identifier les supports, qui, comme les caractères d'un alphabet constituent des ensembles immuables, finis et pré-établis. Ces techniques d'extraction sont

surtout mises à l'épreuve dans le monde de l'histoire sur de gros projets et/ou très en amont : par exemple OCR pour alimenter des bases telles Gallica (BNF), GoogleBooks, ISTEX etc ; racinisation, étiqueteur morpho-syntaxique (PoS tagger), en pré-traitement de projets.

Le tableau 1.6 montre des exemples de lemmatisation et racinisation (stem), qui sont 2 techniques visant à limiter la diversité des mots composant les textes. La racinisation augmente la précision (moins de faux-positifs) mais diminue le rappel (davantage de faux-négatifs).

Mot	naviguer	navigation	naviguait	navigue	navigations
Lemme	naviguer	navigation	naviguer	naviguer	navigation
Stem	navig	navig	navig	navig	navig

Mot	universités	université	univers	universel	universelles
Lemme	université	université	univers	universel	universel
Stem	univers	univers	univers	univers	univers

TABLE 1.6 – Exemples de lemmatisation et de racinisation (stem)

En français principalement deux algorithmes de lemmatisation/PoS tagger existent : TreeTagger (Schmid, 1994) et MATE (Bohnet, 2010). Du côté de la racinisation, *SnowBall*, qui fonctionne par un ensemble de règles syntaxiques (regex) sans apprentissage a été testé avec beaucoup d'échecs.

3.2 Catégories latentes

Dans un second plan, basées sur les textes captés par le premier plan et sur les mêmes algorithmes d'apprentissage [4], plusieurs méthodes d'extraction de catégories latentes existent.

3.2.1 Niveau sémiotique

La NERC se concentre sur un plus haut niveau d'abstraction que l'extraction de contenus formatés. Elle permet de repérer certains types d'entités dans un corpus : lieux, dates, personnes, institutions, monnaie, animaux, etc. Le nombre d'entités est fini et souvent faible en fonction des « défis » : 7 (MUC), 8 (IREX) ou 5 (ACE). Surtout destiné à indexer et désambiguïser les contenus textuels, la NERC opère suivant les critères définis par un système de descripteurs pré-établi (Goerz et Scholz, 2010). L'étude de Nadeau (2007) montre l'état de la recherche dans cette discipline. Éprouvées dans le domaine spatio-temporel, ces techniques se fiabilisent avec les efforts de recherche. En effet le problème des NERC n'est pas une tâche solutionnée, notamment dans d'autres langues que l'anglais (Marrero *et al.*, 2013). Des travaux récents ont obtenu de bons résultats (F-measure de 88%) pour la détection d'entité spatio-temporelles en français (Azpeitia *et al.*, 2014; Bornet et Kaplan, 2017). Des travaux spécifiques à certains types de données tendent à créer des classes très fines pour un champ spécialisé ; par exemple le projet ANIMITEX (Alatrista-Salas *et al.*, 2014) extrait des données spatiales (géographie, aménagement du territoire). D'autres travaux visent un typage fin des entités (*fine-grained*) en construisant une hiérarchie de 150 entités sans restreindre le domaine d'étude à un champ d'application (Satoshi Sekine et Nobata, 2002). Plus proche de nos problématiques, car s'adaptant à des domaines inconnus, l'Extension d'ensembles d'entités (ESE) consiste à trouver les entités d'une classe à partir d'un set donné. Par exemple étendre le set « assemblages mécaniques : {écrou, vis, boulon, soudure, rivet} » à partir de textes de mécanique (Gupta et Manning, 2014). Des travaux complémentaires permettent d'associer une instance à plusieurs entités (Zhou et Zhang, 2007) : *multi-instance multi-label*. Enfin, les travaux les plus avancés, indépendants du domaine d'étude, consistent à créer des catégories à partir des mentions des instances au niveau du corpus (Yaghoobzadeh et Schütze, 2015). Ces techniques utilisent des outils de *word-embedding* (voir section 4.2.4).

3.2.2 Niveau fichier

Le *clustering* permet de classifier les textes en catégories (mathématiques, mécanique, médecine, par exemple si l'algorithme est supervisé). Cela permet aussi de classifier des images par forme ou couleurs (cf. OCR). Dans le cas d'apprentissage, des algorithmes supervisés ou semi-supervisés (Grira *et al.*, 2004) peuvent être utilisés, en fonction de la représentativité du corpus d'apprentissage (initial). L'algorithme supervisé requiert que l'on définisse initialement tous les clusters possibles (toutes les disciplines dans l'exemple précédent) et que chacune soit bien représentée dans les données destinées à l'apprentissage. La section "Clustering" (4.2) étudie plus largement ces méthodes, ainsi que celles sans apprentissage.

3.2.3 Contributions d'acteurs extérieurs

Des initiatives ouvertes et récentes destinées à collecter les données auprès des acteurs du terrain émergent. Ces acteurs peuvent parfois être simplement un utilisateur prêt à relire un texte, ou identifier un numéro sur une partie d'image. On parle de

4. la tâche est la même d'un point de vue logique, qu'il s'agisse de caractères, de mots ou de textes

crowdsourcing. L'idée réside dans la décentralisation des contributeurs, capables de remplacer ou de vérifier le travail d'un algorithme. Les tâches sous-traitées à l'utilisateur-contributeur sont souvent la captation et l'assignation de catégories, par exemple sur un texte. Parfois la contribution peut concerner un témoignage voire un dépôt d'archives personnel. Cette approche est parti-

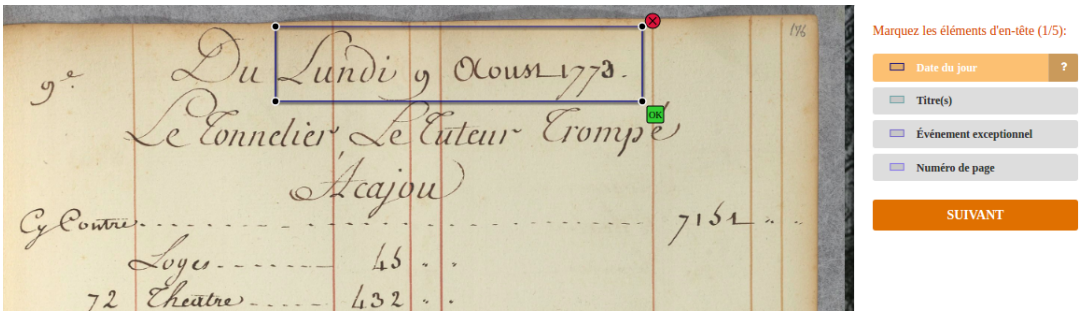


FIGURE 1.16 – Exemple de système de crowdsourcing : marquer et identifier les entités d'une page manuscrite (date, personnages, comptes, etc.). projet Recital

culièrement adaptée à la constitution de séries sous-entendues par la notion de *typicité* de Heinich (2009), voir section 1.2.1.

3.3 Relations

Un usage direct des NERC est l'assignation de métadonnées à des textes avec des méthodes de classification binaire. Binaire ici signifie que la classification n'est pas quantifiée. Au-delà de la classification, des travaux s'intéressent à établir des relations entre entités, entre textes. Plusieurs familles d'algorithmes opèrent, par exemple Han *et al.* (2003) utilise les algorithmes Machine à Vecteur de Support (SVM) (voir section 4.2.4) pour une tâche de *relationship extraction* [5].

Plusieurs projets phare marquent le fantasme du « tout sémantique », le projet NELL (Mitchell *et al.*, 2015; Carlson *et al.*, 2010) (*never ending language learner*) vise à extraire des informations structurées de pages web non-structurées pour peupler une ontologie contenant des millions de faits et s'améliorer chaque jour. Le projet FrameNet, pionnier de cette problématique, conçoit le texte comme des cadres (*frames*) à différentes échelles : entités, phrase, paragraphe. Les cadres entretiennent des relations entre eux. Chaque `cadre` est centré sur une action, définie par une `unité lexicale` (LU) : un lemme et ses synonymes. Les `éléments de cadre` (FE) sont propres aux LU, certains éléments sont centraux (*core*), d'autres contextuels (*non-core*). Par exemple l'unité lexicale « survivre » demande 2 éléments : le « survivant » et la « situation dangereuse » (voir exemple 1.8).

```
À <FE type="noncore" name="place">Nantes</FE>
<FE type="core" name="survivor">Matthieu Quantin</FE>
<LU name="survivre">a survécu</LU> à
la <FE type="core" name="dangerousSituation">rédaction
de sa thèse en été</FE>
```

Exemple 1.8 – Exemple de cadre et éléments en FrameNet

Les relations et entités sont recherchées en référence à un modèle pré-existant. Aujourd'hui, ce modèle est souvent une ontologie, que l'on cherche à instancier.

D'autres formes de création de liens peuvent être produites sans *Relationship extraction*, à l'aide de modélisation, par exemple à partir des catégories de documents, en organisant une hiérarchie (héritage) entre elles. Dans ce cas on ne cherche pas à instancier une ontologie, mais un modèle d'entités comme une taxonomie. Cela permet de décrire les relations entre classes donc entre entités de classes différentes.

3.4 Structuration des données extraites des textes

Les données ainsi extraites peuvent servir dans le cadre des modèles de structuration de connaissances présentés en section "Des données textuelles aux connaissances explicites" (3).

3.4.1 Niveau fichier.

La NERC peut indexer des textes, en obtenant des lieux ou des personnes mentionnés. Cette indexation ne doit pas être confondue avec les métadonnées d'un texte. En effet, les NERC ne visent pas à trouver l'auteur du texte ou les droits de diffusion (qui sont rarement explicitement écrits dans le texte). Des techniques de captation (section 3.1 avec apprentissage supervisé

5. *Relationship extraction* consiste à établir des liens entre entités nommées (par ex. « X est marié avec Y »)

pourrait repérer des champs où certaines métadonnées sont explicitement renseignées. Par exemple les auteurs d'un article scientifique, la date, le titre.

Potentiellement des méthodes de clustering avec apprentissage supervisé pourraient permettre de structurer un ensemble de fichiers. Cette structuration consisterait à associer automatiquement des métadonnées à chaque fichier. Dans les faits, cette pratique n'est pas courante, les champs de métadonnées concernés serait « auteur » et éventuellement « date ». En effet les métadonnées d'un fichier sont difficilement prédictibles à partir de son contenu.

3.4.2 Niveau sémiotique.

Le projet *FrameNet* (voir section 3.3) a plutôt vocation à recenser les contenus et proposer une méthode d'analyse. Le projet TEI est devenu le standard de description des contenus textuel. Il propose un ensemble de balises (markups représentées en XML) complémentaires aux métadonnées du document. Ces balises permettent notamment d'étiqueter le contenu texte avec des entités nommées, comme dans l'exemple 1.9. Une grande partie du projet TEI, orientée sur la mise en forme de textes, ne nous intéresse pas ici. Le projet a la particularité d'être extensible par ADE, ainsi avec un base commune, il est possible de développer de nouvelles balises pour décrire les éléments spécifiques à un domaine. Par exemple, la définition de balises pour les textes d'opéra est en cours. Ce type de structuration s'inscrit dans la continuité de l'extraction d'entités nommées (section 3.2.1).

```
<name type="person">Matthieu Quantin</name>
a réalisé sa thèse à
<name type="place" ref="http://sws.geonames.org/2990969/">Nantes</name>
dans l'<name type="org">École Centrale</name>
avec <name type="person">Florent Laroche</name>.
```

Exemple 1.9 – Entités nommées et balises TEI (version P5)

À l'inverse des projets de type *FrameNet*, ce schéma ne permet pas de marquer les liens entre les entités, résultant de l'extraction de relations (section 3.3).

3.5 Graphes binaires et données structurées

Des arcs entre les instances de classes (entités) permettent la création de graphes binaires. Les liens peuvent provenir d'instances de propriétés d'ontologies par exemple. Dans le cas de triplet RDF chaque triplet est un arc. Les arcs sont souvent typés : ils portent des attributs, par exemple une classe de propriété. Dans le paradigme RDF, d'autres triplets peuvent préciser l'arc. Ces graphes de données sont bien connus, le web sémantique est souvent représenté en graphe binaire. Les propriétés des graphes sont étudiées en section 4.

La dimension analytique avec les ontologies fonctionne par inférences : ce sont des raisonnements à partir de règles sur les graphes. Le côté *formalisation* des connaissances est ici exploité. Si les données sont suffisantes et les règles complètes, on peut requêter des réseaux complexes, par exemple de relations industrielles : « quels sont les types de produits fabriqués aux XIX^{e} siècle par les usines dont le propriétaire est M. Voruz ? ».

La gestion de la documentation du patrimoine par une voie de structuration des données vers des éléments (informations, connaissances) explicites est une étape fondamentale pour la captation (nouveaux textes), et une voie prometteuse pour les inférences sur les informations explicites (abstraction de haut niveau : entités, relations). Cette approche est compatible avec certains frameworks de gestion des données 3D (section 2.3.4).

Verrou scientifique (2)

Le verrou que nous abordons ici est celui d'une vision analytique du patrimoine, et non seulement documentaire. La voie des NERC implique de créer des classes *a priori*. D'un point de vue analytique, cela permet de vérifier des pistes de recherches, mais difficilement de faire émerger de nouvelles connaissances. En effet les inférences automatiques sont encore loin de celles que l'historien produit. Les outils analytiques de la documentation du patrimoine devrait permettre une analyse des corpus de textes en préservant la diversité des contenus, non bridés par des classes *a priori*. L'objectif est de fournir à l'historien de nouvelles pistes de recherches, sans modélisation préalable. Il s'agit d'une condition de **non-restriction du domaine d'étude**. Ce fonctionnement produirait des analyses poussées des textes avec de bas niveaux d'abstraction et inclurait l'historien dans le processus là où le raisonnement (inférences) est nécessaire.

4 Analyses des données textuelles et connaissances explicites

La formalisation des données (section 3) est riche en potentialités : standardisation des descripteurs, partage de données. Ce flux de données, depuis sa captation jusqu'à formalisation débouche principalement sur un horizon de valorisation et de diffusion.

Comme indiqué en vert sur la figure 1.15, certaines méthodes d'analyses découlent de hauts niveaux de la "Des données textuelles aux connaissances explicites" (3), d'autres se contentent d'un niveau très bas (texte brut). Ce sont des étapes de transformation de données où le texte n'est qu'une liste de mots vers le statut d'information : des proximités entre des textes, des regroupements, des graphes, etc.

L'objectif ici est de penser un espace pour l'analyse, activité centrale en Histoire. Deux grands types d'analyses existent : les analyses prédictives (*Predictive Mining*) et les analyses descriptives (*Descriptive Mining*). Nous limiterons notre étude aux analyses descriptives. Les analyses prédictives étant vivement critiqués et peu utilisées en SHS. Parmi les analyses descriptives qui peuvent intéresser l'historien, nous trouvons les sémantique vectorielle (VSM) avec les calculs de proximités (section 4.1), le clustering (section4.2), les analyses spatio-temporelles (section 4.3) et la lexicométrie de manière générale (section 4.4).

Verrou scientifique (3)

Le verrou que nous abordons ici est celui des biais liés aux techniques d'apprentissage supervisé. En effet, nous considérons l'**unicité et l'unité des corpus** de textes que nous étudions. L'apprentissage supervisé avec des données extérieures au corpus apporte nécessairement un biais à l'analyse du corpus (remet en cause l'unicité), tandis que l'apprentissage supervisé sur une partie du corpus remet en cause l'unité : nous ne pourrons considérer qu'une partie du corpus est représentative de l'ensemble. Nous éviterons ces techniques sans exclure l'apprentissage non-supervisé.

Une constante de la première partie de cette étude est l'étude des mots dans les documents. En effet, nos seules données certaines sont donc des mots, répartis dans les documents. Nous pouvons alors utiliser la notation suivante $A(n \times h)$ est une matrice terme-document, l_i un lemme ou un mot, et p_j un document avec $i \in [0, h]$ et $j \in [O, n]$; où h est le nombre total de lemmes ou mots et n le nombre de documents. Des listes de mots interdits, car non porteurs de sens (*stop-lists*), sont habituellement utilisés pour filtrer les sacs de mots. Ces listes contiennent ponctuation, adverbes, pronoms, conjonctions, etc.

Un dilemme s'établit entre 2 options :

d'une part le « *distant reading* » construit des descripteurs haut-niveau (lieux, époques, thématiques, etc.), abstractions des contenus des textes, avec un nombre réduit de dimensions (nombre d'attributs) pré-établies. Cette approche permet des inférences automatiques. Par exemple : assimiler les documents à des entités nommées par NERC.

d'autre part le « *close reading* » conserve la complexité des textes au-delà du contenu factuel, approche sans a priori les données disponibles. Cette approche limite les inférences automatiques. Par exemple : extraire la terminologie.

4.1 Sémantique vectorielle et proximités

La VSM représente les documents (p_j) et/ou les termes (l_i) par des vecteurs afin de pouvoir calculer des proximités (parfois des distances) entre eux. La figure 1.23 montre des vecteurs documents-mots (par exemple Lascaux : $[8, 5, 0, 0, 10]$). Ce type de description construit des vecteurs de grande dimension (1 dimension par mot dans le corpus).

Remarque : une mesure doit répondre à 4 critères pour être une *distance*, autrement on parle de dissimilarité ou de proximité (qui évoluent en opposées) :

— Positivité : $d(a, b) \geq 0$
— Symétrie : $d(a, b) = d(b, a)$
— Identité : $d(a, b) = 0 \Leftrightarrow a = b$
— Inégalité triangulaire : $d(a, c) \leq d(a, b) + d(b, c)$

4.1.1 Extraction de terminologie

Approche. L'extraction de termes-clé produit des mots qui représentent le contenu de chaque document. L'idée est qu'une combinaison de termes « représente le contenu essentiel du document en une forme condensée » (Rose *et al.*, 2010, traduit par moi-même). Cette forme condensée sera par la suite plus facilement manipulable et interprétable. Conformément au schéma DIKW en fig. 1.2, nous restons au niveau « I », *information*, mais cherchons à produire des informations de plus haut niveau, facilement manipulables et interprétables en *connaissances* (niveau « K »). L'extraction de terminologie se distingue des NERC par le fait qu'aucune catégorie pré-existante n'est recherchée. En général cette discipline cherche à extraire des phrases-clés (composées de plusieurs mots) notamment pour l'alignement de texte (traductions), mais aussi pour extraire des combinaisons de termes caractéristiques d'un texte. Cette dernière fonctionnalité nous intéresse.

Turney (2000) valide cette approche en montrant que parmi les phrases-clés définies par un auteur pour son texte, sans aucun vocabulaire contrôlé, entre 70 et 80% des phrases-clés apparaissent dans le corps du texte. Par ailleurs, Finlayson et Kulkarni (2011) prouvent que les phrases-clés présentent une plus grande stabilité sémantique lorsque leur nombre de composants augmente (complexité). Ce résultat corrobore l'intuition que les mots simples sont plus ambigus que les mots composés, par exemple « article de presse » et « article scientifique » désambiguïsent partiellement le terme « article ».

Nous focalisons notre étude sur l'extraction de phrases-clés ou séquences de n mots contigus (n-grams). Plus particulièrement nous nous intéressons aux n-grams de longueur (ordre n) variable. Davantage capable de maintenir l'information séquentielle

contenue dans un texte que le n-gram d'ordre fixe (Wang et Mccallum, 2005). Allant plus loin sur cette idée de maintenir l'information séquentielle, nous introduisons la notion de résilience à variabilité. Les lemmes (ou stem, voir section 3.1) constituent une réponse à l'échelle du mot, les skip-gram (Mikolov *et al.*, 2013a) une réponse à l'échelle du n-gram. Permettant l'inclusion d'éléments étrangers dans le n-gram, cette technique offre une plus grande flexibilité. Par exemple : dans la chaîne de caractères « pour nos travaux nous utilisons des microscopes électroniques plutôt à balayage », on peut trouver le 1-skip-3-gram *microscope électronique à balayage*.

Les algorithmes existants. Les problèmes d'extraction de terminologie en français, indépendants de tout domaine, ont été traités par des algorithmes depuis le début du TAL. On peut les catégoriser en 2 types :

- Les règles statistiques :
  - la fréquence d'occurrence : le décompte des occurrences est utilisé dans le plupart des algorithmes comme un seuil simple de sélection. Souvent ce seuil opère après d'autres règles.
  - le pouvoir discriminant du terme au sein du corpus : on calcule souvent cette valeur avec IDF.
  - le pouvoir discriminant au sein d'un corpus général (*termhood*) : cette mesure est similaire à la précédente, mais s'intéresse au corpus par rapport à la langue en général. Elle mesure à quel point le mot est relatif à un domaine.
  - la rareté : basée sur les fréquences d'occurrence, cette mesure simple permet d'estimer la sur-représentation d'un terme dans le corpus. C'est une autre approche de *termhood* par rapport à un corpus général.
  - la distribution au sein du corpus, l'information contenue par le mot, souvent mesurée comme entropie ou entropie relative (divergence de Kullback-Leibler (KL) : déviation par rapport à la distribution de poisson, etc.
  - la morphologie lexicale de n-grams (*unithood*) : la cohérence lexicale mesure la cohérence d'un ensemble de mots (dans un corpus donné), la complexité suppose que les termes plus complexes sont plus intéressants.
- règles linguistiques : motifs grammaticaux séquentiels, par exemple des suites d'adjectifs ou noms autant de fois que nécessaire. Ce type de règle implique l'utilisation de PoS tagger pour obtenir la nature grammaticale des mots. Dans le cas d'extraction de n-grams, les filtres linguistiques sont souvent utilisés en primo-sélection.

D'autres types de règles très rudimentaires viennent compléter celles-ci. La méthode des *zones* suppose que certaines zones du texte présentent une meilleure probabilité de contenir des termes caractéristiques (par exemple la première ligne du texte, ou des zones de titre), la méthode des *majuscules* suppose que les noms propres sont intéressants, la méthode des *acronymes expliqués* suppose que les termes désignés par acronyme constituent une caractéristique intéressante, etc.

Certaines méthodes se concentrent sur les règles statistiques. Les travaux plus anciens (années 1990) sont principalement basés sur le décompte des occurrences (fréquence). Nous noterons ici deux ruptures. Les travaux de Daille *et al.* (1994) pour l'alignement de textes bilingues montrent un intérêt pour les règles linguistiques. *ANA* (Enguehard et Pantera, 1995) s'intéresse aux collocations. Les travaux récents s'intéressent aux *collocations* et distributions. Par exemple *RAKE* (Rose *et al.*, 2010) étudie les co-occurrences de termes (*collocation*) en comparant les distributions de paires de termes. *Word2phase* (Mikolov *et al.*, 2013b) (voir section 4.2.4) consiste à repérer itérativement des collocations dans des skip-grams pour apprendre des phrases. Les phrases peuvent potentiellement être rallongées à chaque itération. Par exemple, la collocation « microscope » et « électronique » est repérée en 1re itération, puis la collocation « microscope électronique » et « balayage » est repérée dans la seconde itération, etc. D'autres comme *Likey* (Paukkeri *et al.*, 2008) utilisent un corpus de référence. Certaines approches, focalisées sur les n-grams, obtiennent de résultats efficaces (temps de calcul faible) en mesurant la cohérence (*unithood*) de toutes les paires de termes : *Termight* (Dagan et Church, 1994), *MED* (Bu *et al.*, 2010), *DRUID* (Riedl et Biemann, 2015) par exemple. Certaines d'entre elles sont expliquées en section "Term ranking" (4.1.3)).

Certaines méthodes sont exclusivement basées sur des règles linguistiques (Krauthammer et Nenadic, 2004; Kim *et al.*, 2009) : des séquences de mots dont l'importance est définie par le motif des natures de mots. Ce type d'algorithme présente l'avantage d'une meilleure capacité au passage à l'échelle (*scalable*) que les algorithmes basés sur les règles statistiques.

En pratique, aujourd'hui la plupart des algorithmes existant utilisent une combinaison de ces deux types de règles. Les résultats sont une liste de n-grams, plus ou moins filtrée par des règles en série. Parmi ces méthodes, celle développée par Frantzi *et al.* (2000) intitulée *C-Value/NC-Value* est couramment utilisée. Des algorithmes comme *KIP* (Brook Wu *et al.*, 2005), une méthode mutli-filtres(van Eck *et al.*, 2010), ou *TBX* (Oliver et Vazquez, 2015) sont davantage tournés vers le n-grams. Certaines combinaisons avec des règles linguistiques avec des mesures de rareté (*termhood*) comme *TermExtractor* (Sclano et Velardi, 2007) et *TTC-Termsuite* (Cram et Daille, 2016) présentent de bons résultats.

Enfin d'autres algorithmes utilisent des méthodes par apprentissage supervisé pour ordonner les résultats issus de filtres linguistiques (Jiang *et al.*, 2009; Anette, 2004), éventuellement combinés avec des filtres statistiques simples (Jones et Paynter, 2002; Fatima *et al.*, 2011). Cette approche a été initiée par Frank *et al.* (1999) avec l'algorithme *KEA* qui utilisait une méthode de ranking (SVM ranking) derrière un filtre linguistique (collocations nominales seulement) et une pondération initiale avec *term frequency* (TF) et IDF.

Parmi toutes ces méthodes, peu d'entre elles (Termsuite (Cram et Daille, 2016), FASTR (Jacquemin, 2001), Word2phrase (Mikolov *et al.*, 2013b)) sont tolérantes aux variantes de termes basés sur les mêmes racines, ou combinaison de racines. Il s'agit par exemple de réunir sous une même unité les mots (anglais) *windfarm* et *windmill farm*. Le concept de skip-grams permet également d'expliquer la tolérance à l'identification de variantes comme *windmill offshore farm* (dans la continuité de l'exemple précédent).

4.1.2 Sélection de caractéristiques

La sélection de caractéristiques (*feature selection* en anglais) est une utilisation potentielle des terminologies extraites. Il s'agit d'une application totalement dédiée au clustering : l'objectif est d'éviter le sur-apprentissage et/ou réduire le temps de calcul. Dans certains cas, la diminution du bruit permet d'améliorer la précision. L'idée consiste à réduire les dimensions d'un texte en le représentant par quelques mots importants (les caractéristiques ou *features*).

Aujourd'hui la recherche dans le domaine de la sélection de caractéristiques s'intensifie, les algorithmes sont nombreux. L'enjeu habituel est de pouvoir traiter de grandes quantités de données en les simplifiant. Une récente analyse compare les approches classiques (Liu et Motoda, 2007). Nous les évoquons ici pour leur proximité avec des techniques d'extraction de terminologie discriminantes (trouver les mots importants). Une étude originale s'intéresse à la distribution de Poisson : la déviation de la distribution des mots dans les documents par rapport à une distribution de Poisson (Ogura *et al.*, 2009) fournit des mots qui permettent un clustering binaire (SVM ou k-NN, voir section 4.2) efficace à complexité égale.

En ce sens de nombreux algorithmes basés sur la décomposition en valeur singulière permettent des réductions de dimension en sélectionnant des composantes importantes d'un texte (*word embedding*) : analyse sémantique latente (LSA) et PCA sont parmi les plus connus. *Word2Vec* est un cas un peu différent (voir section 4.2.4).

4.1.3 Term ranking.

Une possibilité d'amélioration des vecteurs est de compter, non pas les occurrences des mots dans les textes mais d'attribuer à chaque mot une pondération. Dans ce cas il s'agit de préparer le (*co-*)*clustering* de matrices. On parle de critère de ranking lorsque cette pondération est utilisée pour ordonner les résultats d'une extraction de terminologie ; certains s'en servent directement pour extraire la terminologie.

IDF et variantes. L'usage de IDF (Salton, 1983) est courant. Cela consiste à mesurer l'importance d'un terme en fonction de sa fréquence d'apparition dans le corpus : plus il est générique moins il est discriminant, donc moins il est important. $\mathrm{IDF(l_i)} = \log \frac{n}{|\{p_j : l_i \in p_j\}|}$, avec n le nombre de documents dans le corpus, et en dénominateur le nombre de documents (p_j) contenant le terme (l_i). Cela donne une réponse à la loi de Zipf, qui indique que la fréquence d'apparition d'un mot est inversement proportionnelle à son rang. Ainsi pour les n mots les plus utilisés, la probabilité d'apparition d'un mot de rang k est environ $1/k$. Cette mesure d'importance est complétée par le décompte de ses occurrences : TF. Ainsi, si un terme présente beaucoup d'occurrences dans peu de documents il est important.

De nombreuses variantes de la formule existent (cf. figure 1.17). Des variantes utilisant la formule existent également. *inverse document frequency* pondéré (WIDF) (Tokunaga et Makoto, 1994) améliore sensiblement la précision, incluant parfois TF et IDF ensemble :

$$\mathrm{WIDF_{ij}} = \log \frac{|l_i : l_i \in p_j)|}{\sum_{j=1}^{n} |l_i : l_i \in p_j|}$$

. *inverse document frequency* modifié (MIDF) (Deisy *et al.*, 2010) est indépendant du nombre de documents, spécialement conçu pour entraîner un classificateur SVM (à kernel non linéaire, voir section 4.2.4) :

$$\mathrm{MIDF_{ij}} = [1 + \log_2 \mathrm{TF}(i,j)] \frac{\mathrm{DFr}(i)}{\mathrm{DF}(i)}$$

avec $\mathrm{DF}(i)$ le nombre de documents dans lequel i apparaît, et $\mathrm{DFr}(i)$ sa valeur normalisée.

WRlog. Le logarithme de l'étrangeté d'un mot (WRlog) mesure la sur-représentation d'un mot dans le corpus par rapport à un dictionnaire de fréquences de mots dans la langue en général (Cram et Daille, 2016). Cette mesure de sur-représentation permet de trouver les mots spécifiques au corpus. Cette mesure peut être vue comme le calcul de WIDF à un niveau supérieur (corpus-langue).

$$WR(l_i) = \begin{cases} 1 + log \dfrac{f_C(l_i)}{f_G(l_i)} & \text{si } f_C(l_i) > f_G(l_i) \\ 1 & \text{sinon} \end{cases}$$

avec $f_C(l_i)$ la fréquence normalisée d'un lemme (l_i) dans le corpus (C), c'est-à-dire le nombre d'occurrences moyen pour 1000 lemmes ; et G un corpus général (la langue). On prend arbitrairement $f_G = 0.1$ (une valeur de rareté) pour les lemmes absents dans le dictionnaire général. Pour les mots-clés composés de plusieurs termes (phrase-clé t_i), le calcul est réalisé pour chaque lemme et la valeur maximale est conservée : $WR(t_i) = \max_{l_i} WR : l_i \in t_i$.

Cohérence lexicale. Le calcul de *multiwords expression distance* (MED)(Bu *et al.*, 2010) est une distance pour les expressions composées de plusieurs termes. Cette distance indique la cohésion des termes entre eux (*unithood*). Il s'agit d'une *information mutuelle ponctuelle* (PMI) entre les termes. On mesure donc les dépendances relatives entre les distributions de termes de la phrase-clé. Cette mesure est inspirée de nombreuses autres mesures dont la *Distance Normalisée de Google* (Cilibrasi et Vitányi,

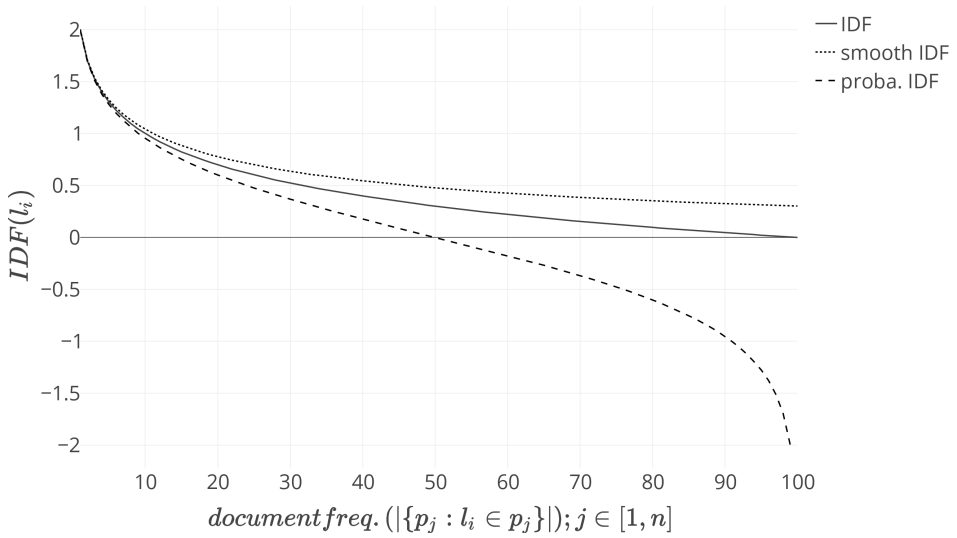


FIGURE 1.17 – Différentes fonctions IDF pour un corpus de 100 documents

2007). Ces mesures sont basées sur la mesure de complexité de Kolmogorov (Kolmogorov, 1963). Elle se place mathématiquement par rapport à 2 autres PMI :

— Fano et Hawkins (1961) mesure la dépendance entre la probabilité de former l'expression clé et la probabilité d'un ou plusieurs de ses composants dans un document ($P(C)|P(A \cup B)$ sur la figure 1.18). Par exemple la probabilité d'occurrence dans un document de « microscope électronique » (C sur la figure 1.18) par rapport à celles de « microscope » (A sur la figure 1.18) ou « électronique » (B sur la figure 1.18). Cette mesure s'intéresse à la dépendance des composants vis-à-vis d'une expression.

— Church et Hank (1990) mesure la dépendance entre la probabilité d'avoir tous les composants et celle d'avoir un des composants dans un même document ($P(A \cap B)|P(A \cup B)$ en figure 1.18). Par exemple la probabilité jointe de « microscope » et « électronique » par rapport à celle de « microscope » ou « électronique ». Cette mesure s'intéresse à la dépendance entre les composants.

Bu *et al.* (2010) propose MED pour mesurer la probabilité de former une expression lorsque les composants sont réunis dans un même document ($P(C)|P(A \cap B)$ en figure 1.18). Cette mesure de cohésion est similaire à celle décrie par Park *et al.* (2002). Par exemple la probabilité d'occurrence dans un document de « microscope électronique » (C sur la figure 1.18) par rapport à celles de « microscope » (A sur la figure 1.18) et « électronique » (B sur la figure 1.18).

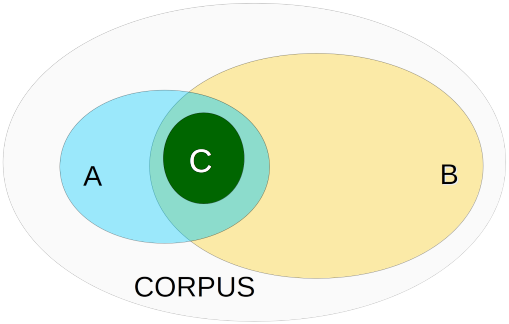


FIGURE 1.18 – Représentation ensembliste des documents où une expression composée apparaît (C) et où ses composants apparaissent (A et B). Par exemple : A= « microscope », B= « électronique », C= « microscope électronique ».

4.1.4 Proximités et graphes

Représentation de documents. Il est possible de créer des graphes à partir des vecteurs représentant les textes. Les vecteurs sont choisis en fonction de ce que l'on veut étudier. Ils peuvent contenir les occurrences (ou pondérations) de tous les mots, les résultats d'extraction de terminologie, de *feature selection* ou d'autres types de dimensions comme les entités nommées.

D'autres formalismes permettent la création de graphes. Hervy (2014) propose un algorithme pour créer un graphe de document avec des arcs non-typés non-pondérés basé sur des listes de mots-clés liées à des documents (voir algorithme 1)

Hervy (2014) démontre l'intérêt de se libérer des schémas inutilement contraignants pour la description du patrimoine (section 3) dans une application interne au musée. Dans les perspectives de sa thèse, il avance l'idée d'élaborer des règles

Algorithme 1 Nantes1900 (graphe)

Entrée: Document et mot-clés associés
Sortie: $V = \{\}$ l'ensemble des nœuds
$E = \{\}$ l'ensemble des arcs

pour chaque document j **faire**
  $V \leftarrow j$
  pour chaque mot-clé $i \in j$ **faire**
    si $i \notin V$ **alors**
      $V \leftarrow i$
    fin si
    créer l'arc $E \leftarrow (i, j)$
  fin pour
fin pour

de connexion entre items. En effet si les approches étudiées jusqu'alors sont centrées sur les graphes binaires (ontologies par exemple), la pondération des liens permettrait la création de nuances, principe important dans l'analyse en histoire et patrimoine.

Verrou scientifique (4)

Le verrou que nous abordons ici est celui d'une approche du patrimoine par la **logique floue** : il semble important de préserver la notion de nuance. Les graphes construits devront proposer une quantification des liens et éventuellement des indicateurs de confiance du calcul de pondération. D'autres techniques que les graphes pourraient intervenir dans la mesure où elles permettent une approche quantifiée des relations créées.

Calcul de proximité. Pour créer le graphe on peut calculer des proximités entre des représentations mathématiques des documents.

Pour calculer la proximité entre 2 documents dans un corpus de N mots ou lemmes dont n sont différents, les proximités classiques (Huang, 2008) utilisent une représentation de chaque document par un vecteur de mots à n dimensions. Dans nos exemples la paire de documents est représentée par une paire de vecteurs (A et B).

— La distance cosinus calcule l'angle entre des vecteurs (voir figure 1.20a). Elle est comprise dans [0,1], 1 étant pour des documents identiques. La L_2 norme est utilisée ici :

$$\cos(A, B) = \frac{A \cdot B}{||A||_2 \cdot ||B||_2} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

— La distance euclidienne utilise une représentation de chaque document par des coordonnées dans un espace à n dimensions. Elle est comprise dans [0,+∞], 0 étant pour des documents identiques :

$$D(A, B) = \sqrt{\sum_{i=1}^{n} |A_i - B_i|^2}$$

— La distance de Manhattan est la simple différence absolue des composantes. Elle est comprise dans [0,+∞], 0 étant pour des documents identiques :

$$MannH(A, B) = \sum_{i=1}^{n} |A_i - B_i|$$

— L'index de Jaccard calcule le ratio entre l'intersection et l'union. Il est compris dans [0,1], 1 étant pour des documents identiques. Il peut utiliser une représentation de chaque document comme un multiensemble (noté $\{\!\{X\}\!\}$) de N éléments ou comme vecteur :

$$simJ(A, B) = \frac{\{\!\{A\}\!\} \cap \{\!\{B\}\!\}}{\{\!\{A\}\!\} \cup \{\!\{B\}\!\}} = \frac{A \cdot B}{A^2 + B^2 - A \cdot B} = \frac{\sum_{i=1}^{n} \min(A_i, B_i)}{\sum_{i=1}^{n} \max(A_i, B_i)}$$

— Le score F1 (coefficient de Sørensen–Dice) est semblable à l'index de Jaccard mais s'intéresse uniquement à une occurrence binaire (pas au nombre d'occurrence, notion multiensembliste). Il est compris dans [0,1], 1 étant pour des documents identiques.

$$F1(A,B) = \frac{2|\{A\} \cap \{B\}|}{|\{A\}| + |\{B\}|} = \frac{2|A \cdot B|}{||A||_2 + ||B||_2}$$

— Le coefficient de corrélation de Pearson, mesure la corrélation et son sens. Habituellement la valeur absolue du coefficient est utilisée pour comparer des vecteurs textes. La L_1 norme est utilisée ici.

$$simP(A,B) = \frac{n\sum_{i=1}^{n} A_i B_i - ||A||_1 \cdot ||B||_1}{\sqrt{(n\sum_{i=1}^{n} A_i^2 - ||A||_2)(n\sum_{i=1}^{n} B_i^2 - ||B||_2)}}$$

— La divergence de KL utilise une représentation de chaque document comme une distribution probabiliste de n mots. Cette divergence mesure l'entropie relative d'une distribution par rapport à une autre. Elle est comprise dans [0,1], 0 étant pour des documents identiques. Cette mesure n'étant pas symétrique ($divKL(A|B) \neq divKL(B|A)$), ce n'est donc pas une distance.

$$divKL(A||B) = \sum_{i=1}^{n} A_i \log(\frac{A_i}{B_i})$$

— La divergence de Jensen-Shannon est une version symétrique et moyennée de la divergence de KL :

$$DivJS(A,B) = \frac{1}{2} divKL(A||C) + \frac{1}{2} divKL(B||C)$$

avec : $C = \frac{1}{2}(A+B)$

Les valeurs peuvent être : le nombre d'occurrences ($|l_i|$), le poids (TF-IDF) ou la probabilité d'apparition par exemple. Cette probabilité peut être calculée de différentes manières, une approche simple est de considérer qu'elle est proportionnelle au nombre d'occurrences : $P(l_i) = \frac{|l_i|}{N}$. La figure 1.20a montre un exemple de distance cosinus entre 2 vecteurs [grotte :2, fresque :1] et [grotte :1, fresque :2].

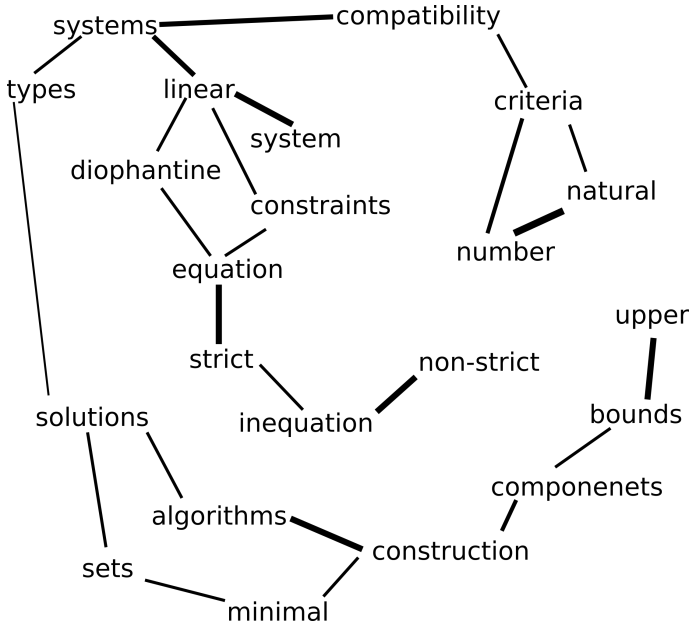


FIGURE 1.19 – Exemple de graphe de co-occurrence simple : chaque document contenant 2 mots différents établit un lien entre eux. L'épaisseur du lien est proportionnelle au nombre de document liants

Ces distances peuvent être calculées sur les données qui nous intéressent : par exemple des tags, des dates, une terminologie ou plus directement l'ensemble des mots contenus dans les documents. Ayant calculé les distances entre tous les documents, on obtient un graphe flou : les documents sont les nœuds (*vertices*) et les proximités sont les arcs (*edges*) entre documents. Ce type de graphe permet par exemple d'étudier la centralité d'un document.

Parfois le graphe inverse est créé : un graphe de co-occurrences d'attributs de document (*term map*). Par exemple, en prenant les mots contenus dans un document comme attribut de celui-ci : les termes sont des nœuds et les documents des arcs entre les

termes. La figure 1.19 illustre ce principe classique en étude de terminologie. Ce type de graphe permet d'étudier la centralité d'un attribut. Les mêmes mesures de graphe que précédemment sont utilisables.

L'utilisation d'algorithmes de *word-embedding* (voir section 4.2.4) permet de simplifier ces graphes en regroupant les mots (nœuds) équivalents, ou de créer d'autres graphes de co-occurrences de termes : les liens entre les arcs sont des co-occurrences dans le vecteur mot. Ce dernier type de graphe peut être orienté, c'est-à-dire que le mot A peut être en tête dans le contexte du mot B sans que l'inverse soit vrai. Par exemple le contexte principal du mot « microscope » peut être « électronique » mais que le contexte principal de ce dernier peut être « configuration ».

De nombreuses analyses utilisent ces représentations en graphes d'un corpus : détection d'événement (section 4.3), extraction de cliques, clustering hiérarchique, etc.

Les graphes Les mesures de base en analyse de graphes sont :

- Centralités, les principales mesures sont :
  - Centralité de proximité : somme des distances d'un nœud à tous les autres nœuds.
  - Centralité intermédiaire : nombre de fois que ce sommet/arc est sur le chemin le plus court entre deux autres nœuds quelconques du graphe.
  - Centralité de degré : centralité basée sur le degré
  - Centralité spectrale : la centralité d'un nœud est déterminée par la centralité des nœuds auxquels il est connecté. C'est une extension de la centralité de degré, avec des poids différents pour chacun des nœuds voisins. Bonacich (2007) montre qu'il s'agit des valeurs propres de la matrice d'adjacence des nœuds. Ce calcul est similaire au *pageRank* de *Google*.
  - Centralité aléatoire : probabilité moyenne d'atteindre le nœud en partant de n'importe quel autre nœud et en suivant les chemins possibles avec une probabilité proportionnelle au poids de l'arc.
- Degré : nombre de connexions que le nœud présente
- Diamètre : distance (chemin le plus court) maximale entre 2 nœuds quelconques du même graphe
- Connectivité : Nombre d'éléments minimum (nœud ou arc) à retirer pour diviser le graphe en 2 parties ou plus.

4.2 Clustering

Nous nous intéressons ici aux algorithmes de *clustering* non-supervisés, ou sans apprentissages. Cette contrainte permet d'éviter toute restriction du domaine d'étude et consensus d'experts. Ces méthodes présentent l'intérêt d'être légères à mettre en place, moins exigeantes en termes de structure de données, mais elles restent grossières, ne permettent pas une analyse fine, à l'échelle de la relation de l'historien à son corpus.

Définitions : Le *clustering* consiste à regrouper les entités — sur la base de leurs propriétés — similaires dans un même groupe et les entités différentes dans des groupes différents. Ce regroupement peut être binaire (1 cluster par entité) ou flou (une entité peut appartenir à plusieurs clusters dans des proportions différentes). Le *co-clustering* consiste en un *clustering* simultané des entités et de leurs propriétés. Dans le cas de texte, on parle de *topic-modeling*.

L'analyse numérique est entièrement basée sur le contenu du corpus et ne dépend pas d'un modèle défini a priori. Ces algorithmes avec apprentissage automatique (non-supervisé) ou sans apprentissage sont principalement basés sur l'algèbre ou les statistiques. Dans ce cas on trouve souvent des calculs matriciels ou des problèmes de graphes.

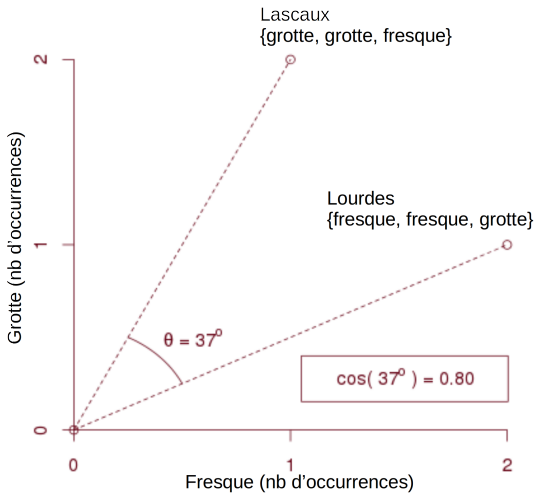
Il existe de nombreux types d'algorithmes de clustering, une typologie est donnée par Berkhin (2006) :

- Hierarchical Methods : Agglomerative Algorithms ; Divisive Algorithms
- Partitioning Methods : Relocation Algorithms ; Probabilistic Clustering ; K-medoids Methods ; K-means Methods ; Density-Based Algorithms
- Grid-Based Methods
- Methods Based on Co-Occurrence of Categorical Data
- Constraint-Based Clustering
- Clustering Algorithms Used in Machine Learning : Gradient Descent ; Artificial Neural Networks ; Evolutionary Methods
- Scalable Clustering Algorithms
- Algorithms For High Dimensional Data : Subspace Clustering ; Projection Techniques ; Co-Clustering Techniques

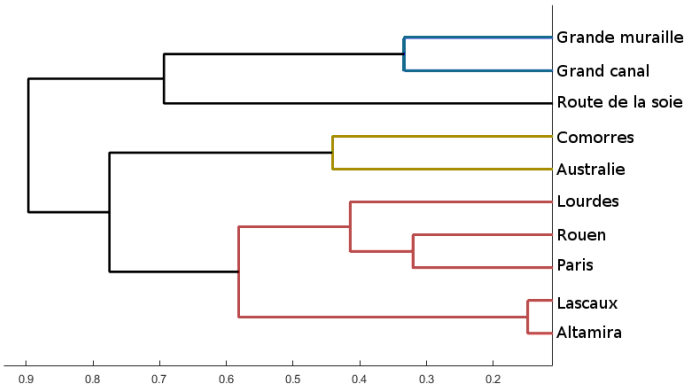
Nous nous appuyons sur cette typologie pour bâtir cette section.

4.2.1 Méthodes hiérarchiques

Les algorithmes de clustering hiérarchique fonctionnent soit par agglomération, soit par division. Ils réalisent des partitionnements de données (un item n'appartient qu'à un seul cluster) et sont basés sur une mesure de distance entre éléments à clusteriser. Dans le cas de texte, cette distance implique de décrire les documents par des vecteurs de mots (voir section "Sémantique vectorielle et proximités" (4.1)). L'optimisation de ces algorithmes date des années 70 (Sibson, 1973).



(a) Distance cosinus entre 2 documents



(b) Exemple de dendrogramme : clustering hiérarchique de description de sites patrimoniaux. La hauteur des « étages » correspond à la distance entre clusters.

FIGURE 1.20 – Le hierarchical clustering

L'option agglomérative regroupe les textes en fonction de proximités. Cette proximité est calculée suivant différentes mesures : cosinus, distance euclidienne, etc. (voir section 4.1.4). L'algorithme naïf est simple : à l'initialisation chaque document est un cluster de 1 document. Une mesure de distance entre clusters est définie, par exemple la distance minimale entre 2 points $\forall$ paire de points issus de 2 clusters (on parle alors de *Single-linkage clustering*). La procédure itérative démarre : on calcule toutes les distances entre clusters, puis on fusionne les deux clusters les plus proches. On recalcule les distances en tenant compte de cette fusion. On recommence les fusions, etc.

De nombreuses mesures de distance entre 2 clusters existent : distance maximale entre 2 points $\forall$ paire de points issus de 2 clusters, distance moyenne, variance maximale, distance minimale après contraction des extrêmes (*CURE*) etc. Les versions optimisées de l'algorithme naïf sont nombreuses.

L'option divisive découpe l'ensemble des documents étape par étape, une simplification du problème étant de supposer qu'on peut le couper en 2 à chaque étape, réduisant le problème à un algorithme simple de *k-means* avec $k = 2$ (voir section 4.2.2). Les résultats de clustering hiérarchiques sont habituellement présentés en dendrogramme (voir figure 1.20b).

4.2.2 Algorithmes de partitionnement génératifs

Cette partie traite exclusivement de k-moyenne et ouvre sur les méthodes similaires (k-médoïdes, etc.) Les méthodes par densité sont présentées dans la section 4.2.5 qui traite du *clustering* sur graphes.

La méthode des k-moyennes divise des données en k clusters minimisant la variance au sein d'une classe. On fixe d'abord le nombre de clusters cible (k). L'algorithme est initialisé en prenant k points (p_k) qui représentent la position moyenne des k clusters. Puis l'algorithme entre en phase d'apprentissage : il assigne à chaque item (donnée) le cluster représenté par le point (p_k) le plus proche (carré de la distance euclidienne) et recalcule la position du point p_k (position moyenne du cluster). La position est calculée en minimisant la distance moyenne des points de données au point p_k. C'est la fonction objectif. La figure 1.21 retrace les étapes de l'algorithme sur un exemple simple en 2 dimensions. En pratique un texte est positionné par les coordonnées (valeurs) son vecteur mot en VSM, il s'agit donc d'un hyper-espace (de plusieurs centaines de dimensions). Lorsque les assignations sont stables, on a un minimum local. L'initialisation (*seeds*) est un facteur déterminant dans la qualité des résultats (minimum local). De nombreux travaux traitent ce problème. Il existe deux méthodes d'initialisation habituelles : Forgy d'une part et le Partitionnement Aléatoire d'autre part. Forgy assigne les k points des moyennes initiales à k données d'entrée choisies aléatoirement. Le partitionnement aléatoire assigne aléatoirement un cluster à chaque donnée et procède ensuite au (avant-premier) calcul des points de moyennes initiales.

K-moyennes++(Arthur et Vassilvitskii, 2007) est un algorithme d'initialisation des k points qui propose d'améliorer la probabilité d'obtenir la solution optimale (minimum global). L'intuition derrière cette approche consiste à répartir les k points des moyennes initiales. Le point de moyenne initiale du premier cluster est choisi aléatoirement parmi les données. Puis chaque point de moyenne initiale est choisi parmi les points restants, avec une probabilité proportionnelle au carré de la distance entre le point et le cluster le plus proche.

La distance et la moyenne peuvent être calculées selon différents modes, créant autant d'alternatives à l'algorithme initial. On peut noter par exemple : k-medoïdes, k-moyennes harmoniques, k-moyennes floues, maximisation d'espérance (Hamerly et Elkan, 2002).

4.2.3 Algorithme pour les données de grandes dimensions

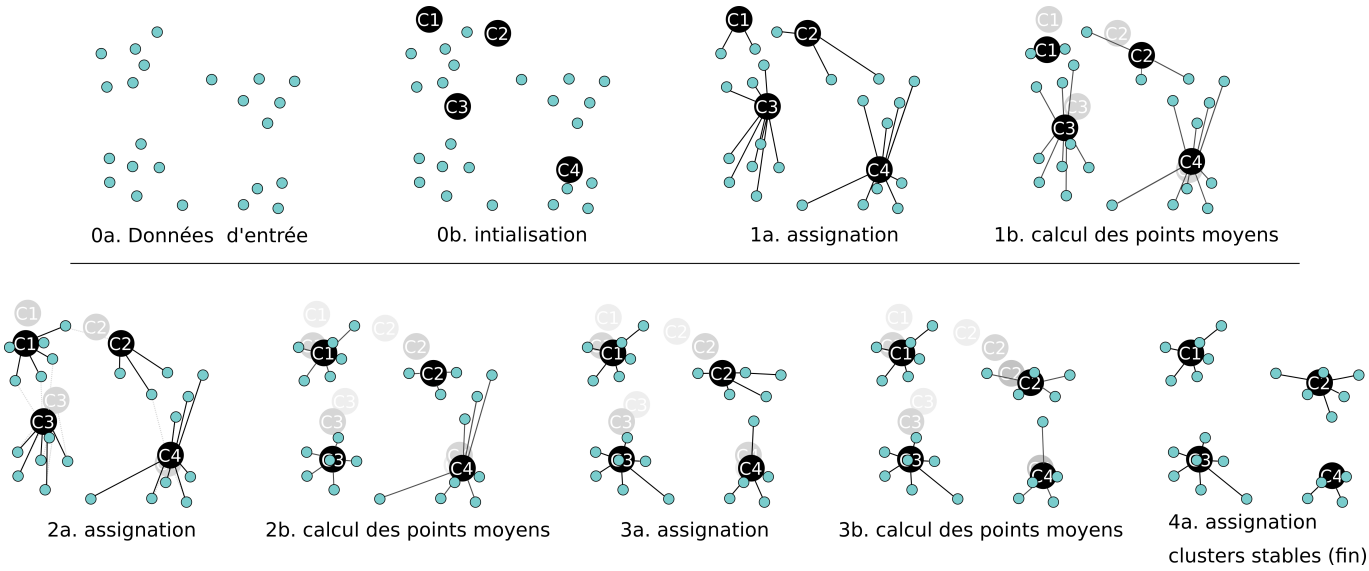


FIGURE 1.21 – L'algorithme des k-moyennes, étape par étape.

Approche du co-clustering. L'hypothèse *sac de mots* suppose qu'on peut faire l'approximation que les textes sont des ensembles de mots (ou lemmes, ou stem), sans ordre. D'après Berkhin (2006, p.33), *High dimension* commence à 16 et devient critique à partir de 20. Au-delà les performances se dégradent et tendent vers la recherche séquentielle. Dans notre cas *sac de mots*, la dimensionnalité des attributs (les mots ou lemmes) est beaucoup plus grande que 20 ; elle se compte en milliers (de mots ou lemmes différents dans un corpus). Nous nous intéressons donc particulièrement à ces algorithmes, en particulier le co-clustering offre de bons résultats sans apprentissage.

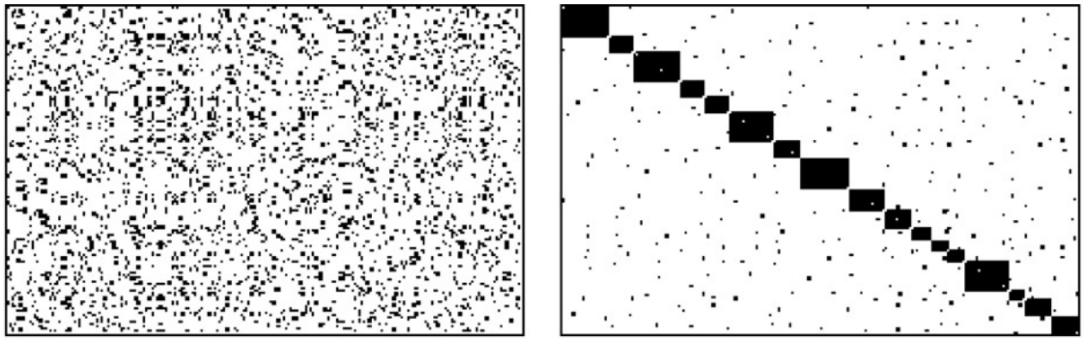


FIGURE 1.22 – Bi-clustering d'une matrice 210x1505. Les deux matrices contiennent les mêmes données. À droite les lignes (rangs) de la matrice sont ré-organisées de sorte à faire ressortir 17 clusters latents de la matrice initiale.

Les algorithmes de co-clustering (cf. figure 1.22) consistent à réduire une matrice A en un « topic-models » de 2 matrices de basses dimensionnalités : $A \approx H.W$, avec $W(n \times k$ avec $k \ll h)$ est une matrice document-topic et H $(k \times h)$ une matrice topic-lemmes (matrice des coefficients). De nombreuses méthodes existent. L'usage de la NMF proposé par Paatero et Tapper (1994), de la LSA proposé par Landauer *et al.* (1998) et de Analyse Latente de Dirichlet (LDA) proposé par Blei *et al.* (2003) sont récurrentes dans le cas des analyses de textes.

On parle de *topic* comme une extension floue du cluster. En effet soit on considère qu'un document est assignable à un seul cluster, soit on considère qu'il est un mélange de topics à définir dans des proportions à définir. Le co-clustering est dans ce dernier cas.

NMF. La NMF est une opération d'algèbre linéaire qui consiste à réaliser directement l'opération de réduction précédemment présentée : $A_{ij} \approx H_{ik}.Wkj$ en minimisant les résidus de l'approximation. Le nombre cible de topics est un paramètre à fournir (hyper-paramètre). Une initialisation aléatoire des matrices cibles et une convergence avec une fonction coût à minimiser (*error function*) est une heuristique naïve. Cette fonction coût peut être tout simplement $||A - H.W||_F$ (moindre carrés) ou par exemple une divergence de KL après avoir normalisé les valeurs des matrices ($\sum_{ij} A_{ij} = B_{ij} = 1$ avec $B_{ij} = H_{ik}.Wkj$). Pour réaliser cette opération l'algorithme classique (Lee et Seung, 2001) est la mise à jour par multiplication (*multiplicative update*) en initialisant H non-negative, puis par itérations des multiplications $H \leftarrow H \frac{W^T A}{W^T W H}$ puis en réalisant l'opération symétrique $W \leftarrow W \frac{A H^T}{W H H^T}$. On note que la $H = \mathbb{I}$ ou $V = \mathbb{I}$ quand $A = H.W$. De nombreuses améliorations ont été proposées en modifiant la fonction coût ou en modifiant l'algorithme sous certaines conditions (Tjoa et Liu, 2010). De récents travaux d'optimisation

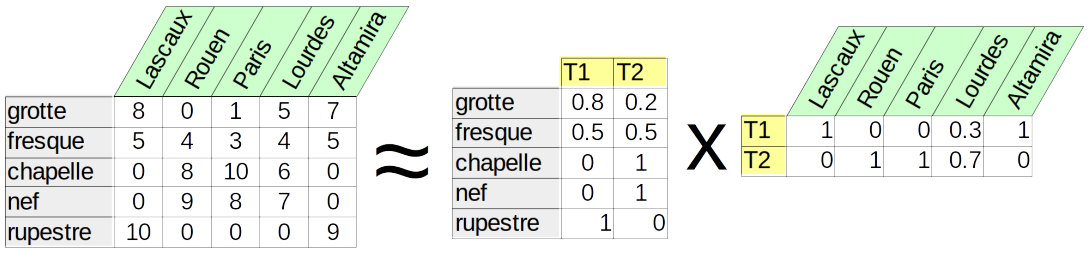


FIGURE 1.23 – Exemple minimaliste illustrant le co-clustering. Les topics (T1 et T2) résultants ne sont pas prévus, ils correspondent au meilleur découpage du corpus initial.

concernent ce domaine en faisant l'hypothèse que la matrice A est presque séparable (cas convexe) (Kapralov *et al.*, 2016; Tepper et Sapiro, 2016).

LSA. LSA étudie les corrélations entre mots (les rangs de la matrice A : apparaissent-ils dans les mêmes documents) en étudiant leurs produits scalaires. La décomposition en valeurs singulières de la matrice des corrélations entre mots permet de faire apparaître les dimensions les plus importantes. Il suffit alors de « tronquer » les matrices singulières après les k valeurs singulières les plus importantes pour obtenir un premier topic-model de vecteurs singuliers (voir la section 4.2.4 qui traite de la même idée).

LDA. LDA suppose que chaque document est un mélange d'un petit nombre de topics et que chaque topic utilise fréquemment un nombre restreint de mots. On fixe un nombre de topics (hyper-paramètre). Les mots de chaque document sont distribués en topics aléatoirement. La phase d'apprentissage commence alors : pour chaque mot de chaque document, on réassigne le topic du mot. Le topic réassigné est celui ayant la plus forte probabilité de générer ce mot. Cette probabilité est calculée sur la base de toutes les autres assignations alors considérées comme valables. Le processus est répété jusqu'à obtention d'une situation stable.

De nombreuses variantes de LDA existent, dont hierarchical Dirichlet process (Teh *et al.*, 2006), technique fournissant un nombre de topics cible inféré automatiquement, appris via le processus stochastique du restaurant chinois. Cette variante décrit des relations entre les topics : des topics de topics (1 seul niveau de hiérarchie).

Verrou scientifique (5)

Le verrou que nous abordons ici est celui de la **représentation des mots**. Les mots sont jusqu'alors considérés comme fondamentalement différents les uns des autres, or, nous connaissons le phénomène des synonymes et du champ lexical : les mots peuvent être proches sémantiquement. Nous devons être en mesure de calculer des distances entre les mots afin d'affiner notre modèle de représentation de documents par des mots (VSM), sans pour autant s'imposer le cadre contraignant de modélisation *a priori*. Le champ lexical d'un mot est toujours dépendant de son contexte d'utilisation ; par exemple le mot « chapiteau » peut appartenir au champ lexical de {« décoration » , « fronton », etc. } en architecture ou de { « scène », « spectacle », etc.} dans un contexte circassien.

4.2.4 Co-occurrences de termes

En TAL, co-occurrence réfère à des représentations de mots (*word-embedding*). Il s'agit de définir les mots par des vecteurs dont les valeurs dépendent des contextes d'apparition des mots. Par exemple les mots « abscisses » et « ordonnées » auront probablement des vecteurs similaires dans un corpus qui explique des graphiques, le mot « peinture » aura un vecteur totalement différent.

Projection PCA est une méthode de projection très répandue. Elle est basée sur la décomposition en valeurs singulières de la matrice termes-documents ou de la matrice de co-occurrence des termes (centrée et éventuellement réduite en fonction des variances). L'objectif est de transformer un espace de h dimensions comportant des variables corrélées, en un espace de dimensions équivalentes où les variables corrélées sont projetées sur des dimensions composites décorrélées (les plus orthogonales possible). Les dimensions dans l'espace d'arrivée sont ordonnées suivant la quantité d'information qu'elles portent (variance). Une matrice de co-occurrences est une matrice symétrique où les associations de termes sont comptées. Le tableau 1.7 montre un exemple, les associations plus lointaines (1 mot d'écart) comptent pour moitié. Dans le cas de matrices document- termes, les distributions de mots fortement entropiques seront mises en valeur, les mots présentant des distributions semblables seront associés. Dans le cas de matrices de co-occurrences, les termes ayant des co-occurrences similaires seront associées. Cette idée signifie que les termes similaires ont des contextes d'apparition similaire. Dans l'exemple du tableau on associera « pomme » avec « pêche » et « fruité » avec « délicieux ».

	pêche	fruit	pomme	délicieux	sucré
pêche		2	0	1.5	1
fruit	2		1	1.5	1
pomme	0	1		1	0
délicieux	1.5	1.5	1		1
sucré	1	1	0	1	

TABLE 1.7 – Exemple de matrice de co-occurrences pour les phrases suivantes : La pêche est un fruit ; La pomme est un fruit ; Les pêches sont sucrées ; Les fruits sucrés sont délicieux ; Les pêches c'est délicieux, les pommes aussi ; La pêche est un fruit délicieux ;

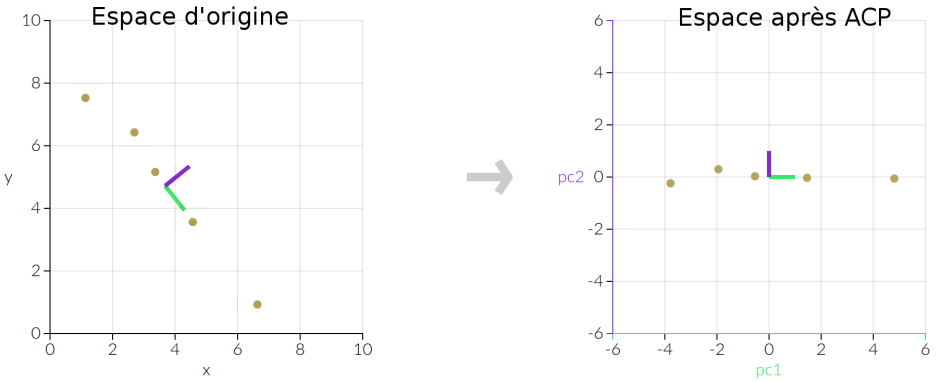


FIGURE 1.24 – PCA de 2 dimensions, après PCA, la dimension `pc2` est quasiment inutile : sa variance est quasi nulle. L'espace a été (presque) réduit à 1 seule dimension

Approche des problèmes non-linéaires. La technique classique de PCA est une projection linéaire. Dans le cas de variables non séparables linéairement, la projection est mauvaise.

t-SNE propose une approche probabiliste de la proximité entre des points. Cette méthode non-linéaire permet une réduction de dimensions. L'idée est de calculer une probabilité pour chaque paire de points de l'espace de grande dimension, la probabilité augmente avec la proximité entre les points. Puis les points sont projetés dans un espace de plus petite dimension (par exemple 2 ou 3 pour la visualisation) avec la même règle. La proximité entre les points i et j parmi N points est la probabilité que i et j soient voisins en utilisant un noyau gaussien sur la distance euclidienne, elle est normalisée et symétrisée :

$$p_{ij} = \frac{p_{i|j} + p_{j|i}}{2mathrmN}$$

avec

$$p_{i|j} = \frac{G(i,j)}{\sum_{k \neq i}^{N} G(i,j)}$$

$G(x,y) = \mathrm{e}^{\frac{-||i-j||^2}{2c^2}}$ L'objectif est de minimiser la divergence des distributions de probabilités (KL) entre l'espace de haute et de basse dimension.

Il peut alors être intéressant de transformer l'espace d'entrée en un espace de plus grande dimension en combinant les dimensions d'origine de manière non linéaires. La figure 1.25 est un exemple simple qui décrit explicitement la transformation. Malheureusement, ce type de transformations « naïves » vers des espaces de grandes dimensions (comme dans notre cas de vecteurs documents-mots) implique des calculs compliqués (produits scalaires) sur les vecteurs. Le produit scalaire entre 2 vecteurs peut être remplacé par une *fonction noyau* qui opère dans un espace de plus grande dimension. Il n'est alors pas nécessaire de définir la transformation explicitement, mais uniquement le type de noyau et ses paramètres (non trivial). Identifier la fonction noyau adéquate implique normalement une phase d'apprentissage automatique. Dans ce cadre les noyaux habituellement utilisés sont : polynomiaux, à base radiale (RBF) et sigmoïdes.

Apprentissage supervisé. Machine à Vecteur de Support (SVM) est une technique de clustering qui utilise ce type de transformation pour travailler dans des espaces de grandes dimensions sans définir de transformation explicite et obtenir les résultats de produits scalaires à moindre coût. L'objectif de cette technique est de trouver un hyperplan qui maximise la distance moyenne aux données. SVM fonctionne essentiellement par apprentissage supervisé : elle trouve des vecteurs supports maximisant la marge sur des petites quantités de données représentatives et assigne toute nouvelle donnée à une partition de l'hyper-espace définie par ces marges.

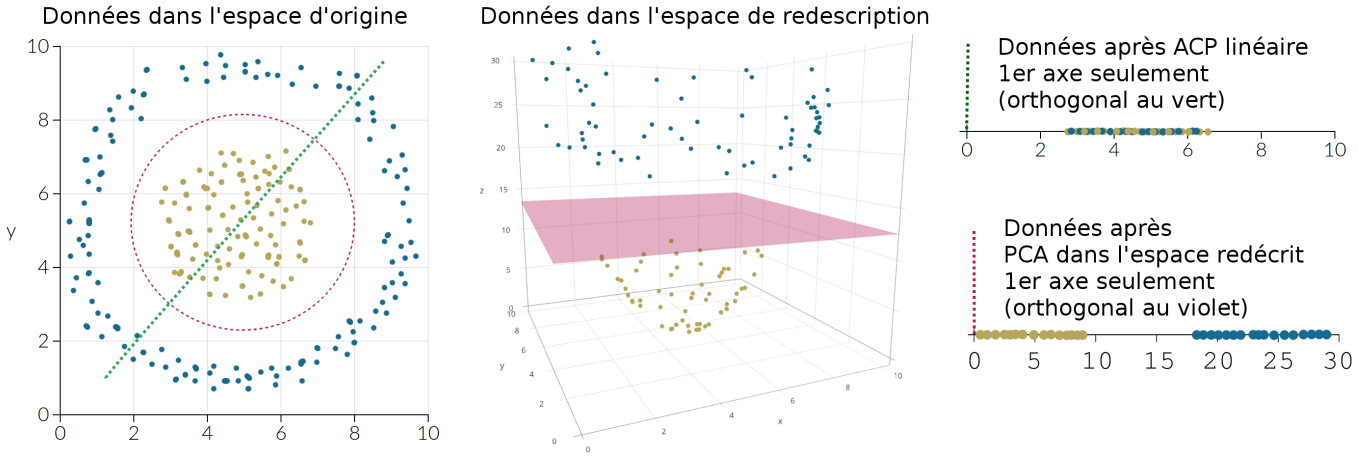


FIGURE 1.25 – Données non séparables linéairement à gauche (coloriées à posteriori). Changement de dimension de l'espace d'entrée par une combinaison non linéaire des axes après centrage des données ($\mathbb{R}^2 \Rightarrow \mathbb{R}^3 : (x, y) \Rightarrow (x, y, z)$ avec $z = x^2 + y^2$). Un kernel remplacerait cette transformation explicite et fournirait le résultat de produits scalaires dans l'espace à grandes dimensions. Résultats de PCA sur le premier axe orthogonal à l'hyperplan (violet ou vert). On observe une amélioration des résultats avec le changement de dimensions (augmentation de la variance sur le premier axe). Note : une analyse en coordonnées polaires des données centrées aurait rendu le problème linéaire.

Word2Vec. Cet algorithme de plongement lexical est très répandu, il fonctionne par apprentissage non supervisé. *Word2Vec* (Mikolov *et al.*, 2013a) fonctionne en prenant des fenêtres de taille donnée autour des mots. La fenêtre est « glissante » comme représenté sur la figure 1.26. Un réseau de neurones utilise alors les collocations pour apprendre une couche intermédiaire d'un nombre prédéfini (hyper-paramètre, par exemple 300) de caractéristiques pour chaque mot. Ces caractéristiques (vecteur contexte) permettent une bonne prédiction du mot. De nombreuses opérations sur ces vecteurs sont possibles. Par exemple une addition vectorielle similaire permet de passer d'un pays à sa capitale pour n'importe quel pays. Les possibilités en termes d'étude de réduction lexicale sont intéressantes : les mots présentant des fortes proximités vectorielles (cf. section 4.1.4) peuvent être groupés.

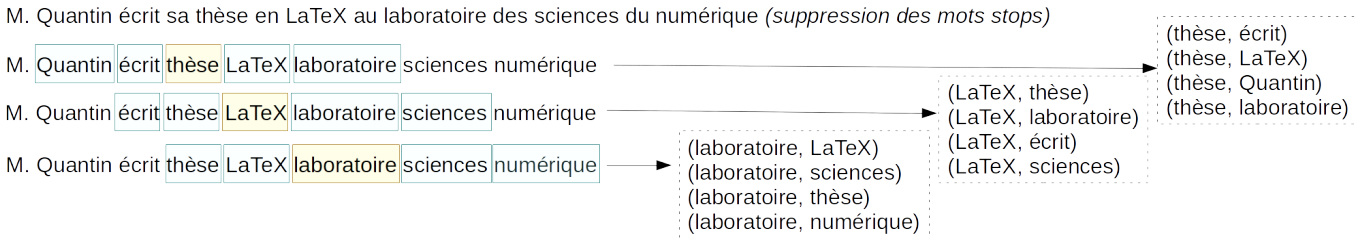


FIGURE 1.26 – Exemple de fenêtre glissante sur une chaîne de mots après suppression des mots stop

Concrètement Word2Vec fonctionne avec un réseau de neurones artificiel (ANN) à 2 couches (une couche intermédiaire). En entrée, il y a tous les mots du corpus, en sortie également. La couche intermédiaire va servir à calculer les modes de collocation des mots, elle contient les vecteurs de mots que nous cherchons à obtenir. Le nombre de neurones intermédiaire correspond à la dimension du vecteur mot que nous obtiendrons (le nombre de mots auquel chaque terme sera lié). Dans notre cas, des corpus de taille réduite et des mots peu fréquents, il est conseillé d'utiliser le modèle de fenêtre skip-gram et la méthode de *negative sampling* pour l'activation de la couche finale. Le modèle skip-gram cherche la probabilité d'apparition des termes de la fenêtre du skip-gram étant donné le terme central. La phase d'apprentissage prend un à un les termes produits par la fenêtre de skip-gram (les ensembles produits sur la figure 1.26) : en entrée le terme central et en sortie le terme associé. Ce sont donc les valeurs de l'apprentissage. Dans la couche intermédiaire du réseau de neurones, les poids initiaux sont aléatoires et ajustés par rétro-propagation à chaque nouvelle combinaison de termes issus des fenêtres. La rétro-propagation s'apparente à la minimisation d'un coût (descente de gradient). Cette légère correction à chaque itération vise à améliorer la prédictibilité du réseau de neurone en ayant connaissance de l'entrée (le mot) et de la sortie (ses collocations). Dans l'idéal cette légère correction touche tous les poids : tous les mots négativement et les mots de la fenêtre positivement. Cependant cette opération est très lourde : potentiellement plusieurs millions de poids à ajuster : 1 poids par mot en entrée et par dimension du vecteur de représentation du mot. Un premier filtre écarte les mots trop génériques du corpus, la probabilité de conserver un lemme l_i est $P(l_i) = (\sqrt{\frac{z(l_i)}{0.001}} + 1) \cdot \frac{0.001}{z(l_i)}$ avec $z(l_i)$ la part des occurrences du lemme l_i dans le corpus et 0.001 le ratio de sampling (valeur par défaut). Pour la partie apprentissage, la solution proposée par le *negative sampling* consiste à choisir quelques mots (entre 10 et 25) comme exemples négatifs (*sample*). Leurs poids sont abaissés (*negative*) tandis que le poids de l'exemple positif est

augmenté. Le choix des exemples négatifs est crucial, puisque n'importe quel mot autre que celui de la collocation est négatif. Les mots les plus courants ont une plus forte probabilité d'être choisis comme exemple négatif : $P(l_i) = \frac{DF(l_i)^{3/4}}{\sum_{j=0}^{n}(DF(l_j)^{3/4})}$ avec TF, le nombre d'occurrences du terme dans le corpus.

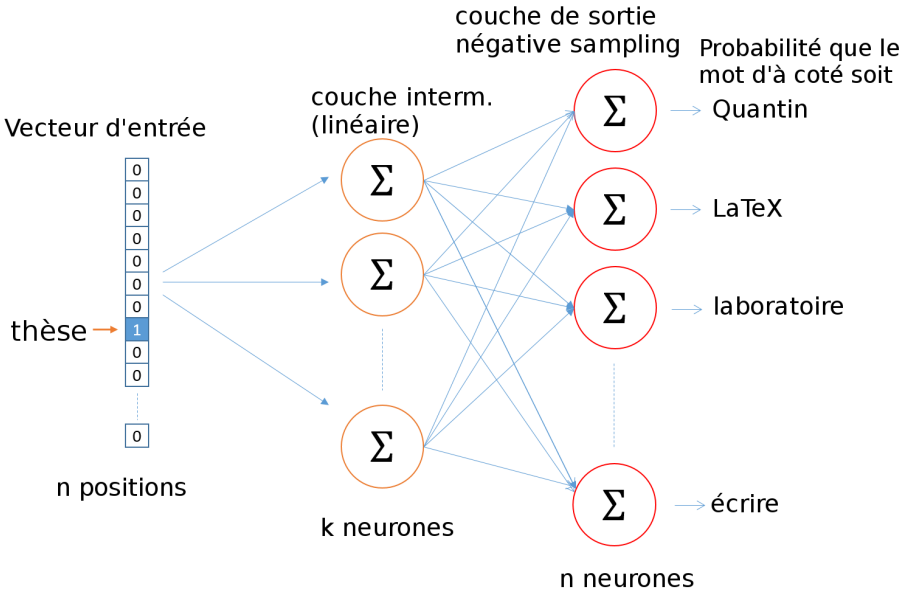


FIGURE 1.27 – Représentation du réseau de neurones par Chris McCormick. En entrée un vecteur de 0 pour chaque lemme du corpus (n) sauf pour une position, en sortie la probabilité d'obtenir un des n lemmes dans la fenêtre du mot d'entrée.

Word2Vec permet de créer des graphes pondérés et orientés de co-occurrence : les termes sont les nœuds et les arcs flous sont tracés avec les pondérations du vecteur du terme (liens que les termes entretiennent entre eux). Puisque tous les mots du corpus sont représentés par un vecteur, il est alors possible de regrouper les mots ayant des vecteurs similaires en utilisant une méthode de clustering ou co-clustering précédent (k-moyennes par exemple) Lorsque ces vecteurs proviennent de différents documents cela peut permettre de détecter et quantifier des usages très similaires de mots : de l'utilisation de notions communes jusqu'au plagiat. On peut parler de regroupements par champs sémantiques, en ce sens Word2vec permet également de désambiguïser des mots ou de trouver des synonymes.

Extensions de word2vec. Une extension de *Word2Vec* vise à construire des vecteurs de représentation de documents (*doc2vec*) qui permettent donc de produire des graphes de proximité/similarité de documents (Le et Mikolov, 2014). À une échelle intermédiaire on peut produire des vecteurs représentatifs de paragraphes ou n'importe quelle unité lexicale composée de mots.

4.2.5 Clustering sur graphes

La détection de sous-graphes directement à partir d'un graphe de documents permet éventuellement de générer des clusters. Ce type de génération peut être obtenu par plusieurs moyens : étude de clique, étude de connectivité (cf. section "Les graphes" (4)). Ce domaine permet la visualisation de cluster sur un graphe en regroupant les nœuds d'un même cluster. Au sens strict, ce domaine ne s'occupe pas de visualisation (mais de la préparation des données). Les cas de clustering de matrices d'adjacence ne sont pas étudiés ici (résolution de clustering de matrices en section 4.2.3).

Une étude de Schaeffer (2007) étudie de nombreux algorithmes de clustering pour des données graphes. Les clusters identifiables dans des graphes ne correspondent pas toujours aux clusters désirables. Par exemple certains clusters de fait sont impossibles à trouver (voir figure 1.28).

Coupe. Un algorithme courant, sous-graphes hautement connectés (HCS) (Hartuv et Shamir, 2000) permet d'opérer directement sur des graphes non-pondérés. Cet algorithme cherche le minimum d'opérations permettant de couper le graphe en 2. Il s'agit du problème de la *coupe minimum*. Une résolution est proposée par l'algorithme de Karger et Stein (1996) qui consiste à itérativement fusionner les nœuds ayant les mêmes voisins dans l'ordre des nœuds ayant le plus de liens et à cumuler les liens à chaque fusion (voir figure 1.29). La technique du *flux maximum* est un autre algorithme utilisé pour cette opération de *coupe minimum*. Il itère sur chaque sous-graphe jusqu'à obtenir des graphes tels que : $\text{nb\_arcs} = \text{nb\_noeuds}/2$ (dit « hautement connectés »).

Des améliorations de cet algorithme pour les graphes flous (pondérés) considère la pondération des liens. Une solution très simple consiste à supprimer les arcs par ordre croissant de pondération. Une autre possibilité consiste à d'abord filtrer le graphe. Cela consiste à supprimer les arcs en deçà d'un seuil, puis à appliquer un algorithme HCS sur les liens restant sans considérer la pondération. D'autres types de filtres peuvent être appliqués au graphe avant application de HSC, afin de le rendre plus creux.

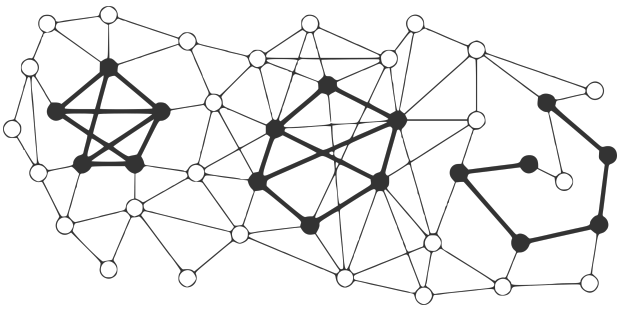


FIGURE 1.28 – Exemple de clusters désirés (en noir) : (1) Identifiable, le premier cluster à gauche est dense : ses arcs internes sont nombreux, ses arcs externes sont peu nombreux ; (2) Difficile à identifier : le second cluster est moins dense, ses arcs externes sont nombreux ; (3) Impossible à identifier : le troisième cluster à droite présente peu d'arcs internes et de nombreuses connexions externes, il présente une inclusion externe.

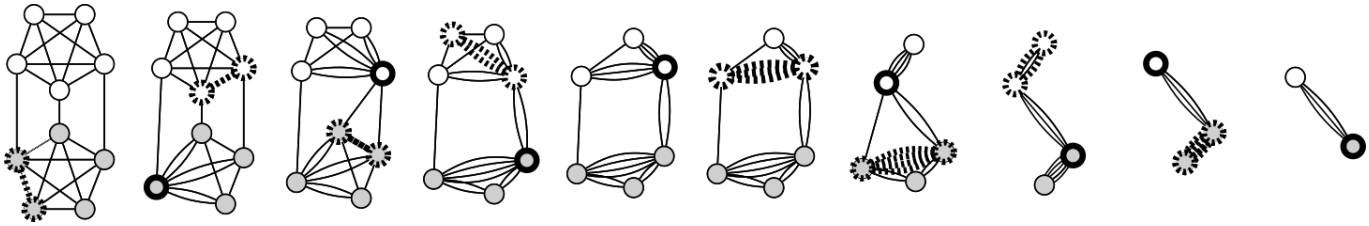


FIGURE 1.29 – Itérations de l'algorithme de Karger sur un petit graphe. Les pointillés indiquent une prochaine fusion, le gras indique le résultat de la fusion précédente.

Densité. La mesure de densité d'un graphe est simple, elle peut permettre de trouver des clusters qui répondent à des sous-groupes maximisant la densité. Sans paramètre supplémentaire ce paramètre revient au problème des cliques (NP-complet). Une clique est un sous-graphe de diamètre 1, qui maximise donc la densité (voir section 4), c'est-à-dire un sous-graphe tel que chacun des nœuds de ce sous-graphe est directement connecté à tous les autres. La recherche de clique peut donc être une condition d'arrêt plus restrictive pour HCS. En fixant le nombre de nœuds minimal au cluster et/ou le nombre minimal de clusters, le problème reste NP-complet, sauf à se satisfaire d'un degré de graphe moyen de 2 (Holzapfel *et al.*, 2006).

Verrou scientifique (6)

Le verrou que nous abordons ici est celui des **échelles**. Les liens qu'entretiennent les documents entre eux nous importe autant que les assignations des documents à des clusters. L'analyse par les clusters doit pouvoir laisser place à une analyse plus détaillée des liens entre quelques documents d'un même cluster, voire une analyse des liens qu'entretiennent 2 documents quel que soit le cluster.

4.3 Analyse spatio-temporelle

Le rapport au temps est un ancrage majeur pour l'historien, en ce sens les analyses diachroniques, éventuellement en lien avec l'espace, semblent intéressantes à étudier. Cette section exploite des données de plus haut niveau de formalisation (donc restriction du domaine d'étude) en utilisant les entités nommées dans le corps du texte, automatiquement ou manuellement extraites, et des métadonnées externes au corps du texte (liées au fichier). Les représentations des documents précédemment utilisées (VSM) doivent être revues pour les représentations spatio-temporelles. En effet, si on peut considérer que la distance entre 2 mots quelconques est constante (ou quantifiable, voir section "Word2Vec" (4.2.4)), la distance entre dates ou lieux implique d'autres opérations. Le problème se résume à un espace à 1 ou 2 dimensions souvent (temps ou espace) parfois 3 dimensions (espace et temps). Alors certaines des méthodes de clustering précédemment évoquées directement appliquées à ces nouvelles représentations des documents produisent des résultats peu intéressants. Par exemple, utiliser les k-moyennes sur des documents représentés par une date et un lieu (vecteur 3) est trivial. Cette section montre donc les défis que représentent les analyses plus fines, de données souvent dynamiques (temporelles) et les visualisations adaptées

4.3.1 Diachronie.

Les séries temporelles sont classiquement analysées dans de nombreuses disciplines (analyses prédictives) principalement en économie / finance où les phénomènes sont stationnaires (persistance de l'objet mesuré). Il s'agit essentiellement de découvrir des règles d'association d'éléments dans le temps. De nombreux modèles sont développés depuis les années 1950 (Whittle, 1951) : modèles autorégressifs (AR) et moyenne mobile (MA), ARMA est une combinaison linéaire des deux. Le modèle Vecteur

Autoregressif (VAR) se focalise sur la prédictibilité d'une variable à partir des antécédents d'autres variables (interdépendances). Sauf dans certains cas rares d'histoire sérielle (Chaunu, 1970) très rigoureux et conscrits, les phénomènes historiques ne sont pas stationnaires.

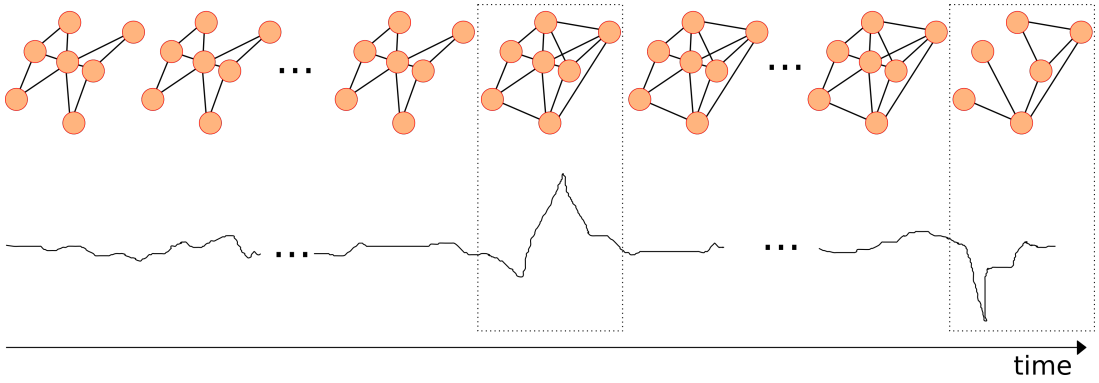


FIGURE 1.30 – Principe de la détection d'événement dans des séries temporelles sur base d'étude de variation de graphes de similarités entre documents

Certaines analyses de motifs temporels s'affranchissent de la non-stationnarité des variables et étudient les co-occurrences, co-évolutions d'événements. Des études de détection d'événements (*event detection*) pourraient intéresser les analyses historiques. Les graphes dynamiques (temporels) (Ren *et al.*, 2017) permettent d'identifier des proximités entre nœuds en étudiant la stabilité temporelle (création et disparition de nœuds ou arêtes). La figure 1.30 illustre le principe. Ces graphes peuvent par exemple être issus de mesures de proximités entre documents. Par une autre approche, l'outil *Diachronic'Explorer* (Lamirel *et al.*, 2016), complémentaire au clustering, permet une analyse de l'évolution des clusters dans le temps (pour les gros corpus textuels), et étudie donc aussi la stabilité d'ensembles textuels, Mei et Zhai (2005) mène une étude similaire en utilisant les modèles de Markhov cachés pour déceler les thèmes transverses aux sous-corpus (obtenus par clustering) malgré des décalages temporels. Enfin de nombreuses études concernent les journaux ou Twitter (Stilo et Velardi, 2016) où les similarités temporelles sont gageures de relations sémantiques (« Time makes sense ») entre termes (cf. figure 1.31). Dans ce cadre ce sont les distributions des occurrences de termes qui sont comparées et qui définissent des événements (clusters de termes situés dans le temps).

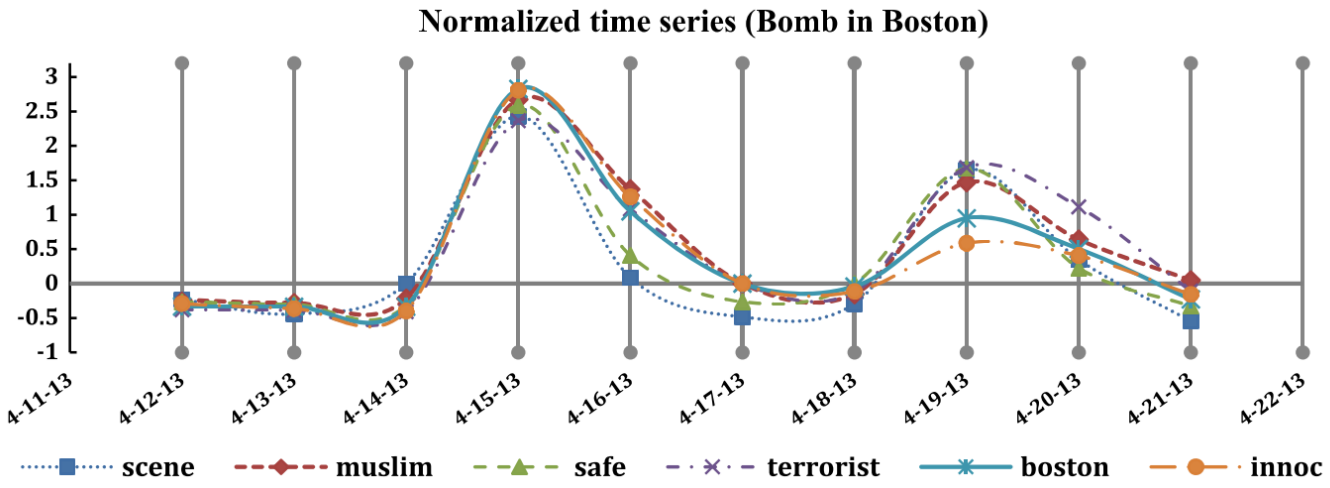


FIGURE 1.31 – Comparaison de la distribution temporelle normalisée des occurrences de certains termes. (Stilo et Velardi, 2016)

4.3.2 Spatio-temporelle

L'analyse des données temporelles en relation avec l'espace produit des motifs dit spatio-temporels. Ces motifs permettent, par exemple, de déterminer au cours du temps les occurrences des événements et leurs localisations. Des analyses peuvent porter sur ces trajectoires. Ces analyses sont principalement menées depuis des données capteurs : GPS des objets connectés, ou depuis des bases de données existantes : étude des espèces invasives, ou depuis les textes (journaux) la diffusion de maladies comme la dengue. La reconnaissance mondiale du livre de Cressie et Wikle (2011) marque l'intérêt pour ce domaine.

La prise en compte de la géographie dans les textes (*textual geography*), davantage utilisée en sociologie qu'en Histoire, fait l'objet d'innombrables études via les réseaux sociaux (twitter principalement). Le standard ISO19108 (ISO, 2015) décrit le

schéma temporel pour les données géographiques, il dépend de ISO8601 (le standard classique pour la description du temps) et s'intéresse plutôt au temps long. Une ontologie (« time » du W3C) complète ces possibilités de description du temps.

Pour les textes bruts (en anglais), de nombreux programmes de recherche développent des outils comme *Textual Geography Analyzer*[6] (Wilkens, 2015) : ils utilisent les NERC (celui de Stanford) et un API de localisation (de Google) pour produire des visualisations des entités géographiques issues de textes.
Le projet international phare dans ce domaine est *mapping the Republic of Letters* (Ceserani et Armond, 2015). Ce projet consiste à créer des représentations spatiales et temporelles des correspondances de grands auteurs européens de la fin du XVII^{e} et début du XVII^{e} siècles. La figure 1.32 montre un exemple de visualisation résultant de ce projet. Notons que ces visualisations constituent la « partie émergée de l'iceberg », issues de nombreuses années de travail en amont sur les corpus d'archive (cf. section "Des données textuelles aux connaissances explicites" (3)) par les groupes *D'Alembert* du CNRS et *Huygens ing* des Pays-Bas.

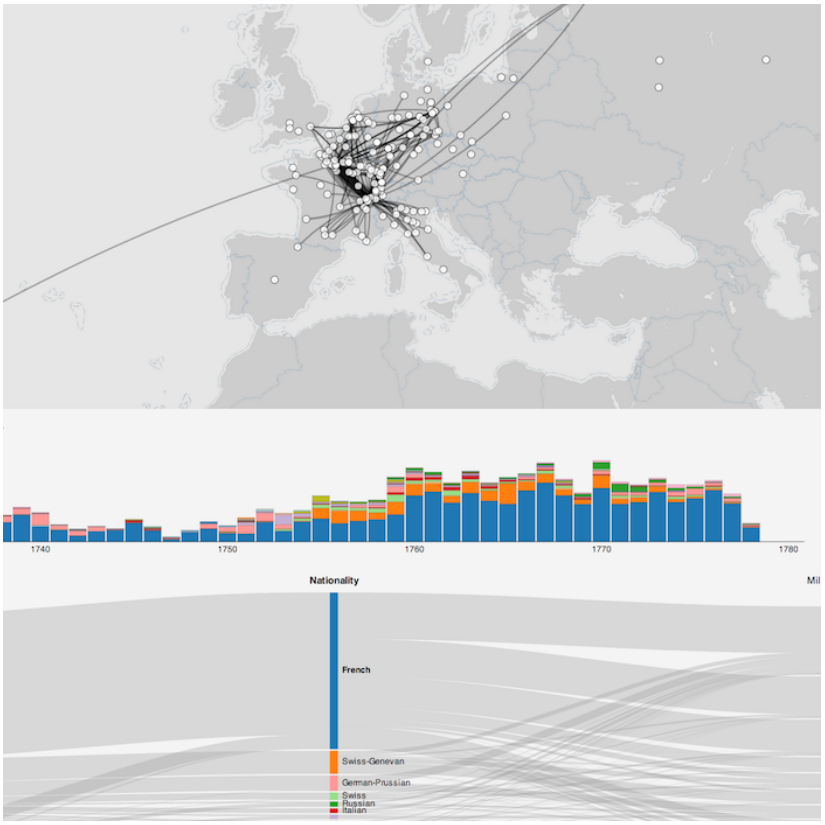


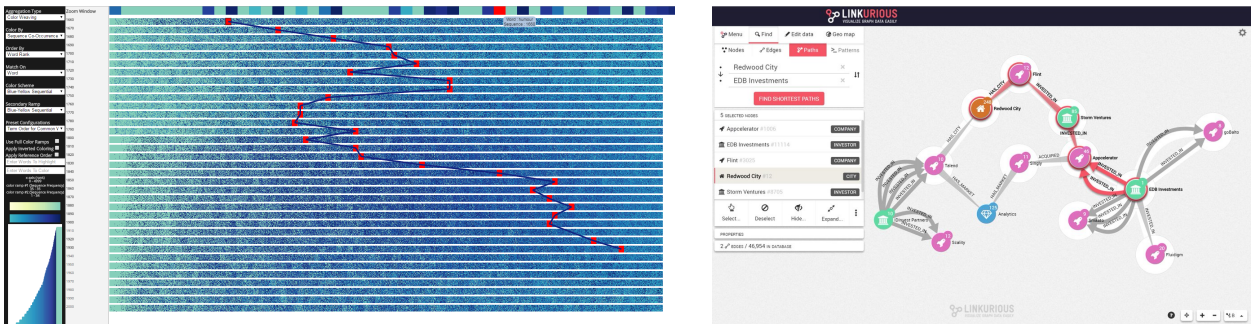
FIGURE 1.32 – Visualisation de la correspondance de D'Alembert (Projet *mapping the Republic of letters*)

Incertitudes. La prise en compte des incertitudes sur les données spatiales et temporelles en contexte archéologique/Historique a été étudié par l'équipe de C. de Runz (Zoghlami *et al.*, 2011). Ces travaux établissent un pont entre la logique spatio-temporelle floue (au sens de la *logique floue*) de l'historien (ex : tournant du XX^{e} siècle) et l'absurde précision du « timestamp » de l'ordinateur (ex : 1905-10-30T10 :45 UTC). Ils complètent la norme ISO8601 (ISO, 1988), qui permet d'encoder des durées, et gère mal l'incertitude temporelle.

Verrou scientifique (7)

Le verrou que nous abordons ici est celui des relations spatiales et temporelles aux données textes. Lorsque ce type de données sont disponibles nous devons être en mesure de les prendre en compte dans leur acceptation incertaine (dans la continuité des graphes flous pour les relations sémantiques), même pour les phénomènes non-stationnaires. Il semble primordial que ces dimensions ne deviennent pas un moyen d'accès exclusif aux données. En effet certaines n'ont pas de dimension spatiale ou/ni temporelles, pourtant elles demeurent centrales pour la compréhension du phénomène historique et patrimonial (ex : les conditions de travail dans les mines de charbon). Il s'agit d'une **approche multi-dimensionnelle** du patrimoine.

6. Projet de l'université Notre Dame (Indiana, USA), récompensé pour son implémentation en 2017



(a) TextDNA (b) Linkurious

FIGURE 1.33 – Exemples d'interfaces de visualisation de deux logiciels

4.4 Outils d'analyses de textes existants

La lexicométrie est une discipline née en France dans les années 1970 et étudie les textes de manière quantitative (*close-reading*). Dans le domaine des SHS, les approches suivantes sont directement opérationnelles à partir des textes, incluant souvent des algorithmes présentés précédemment (étiquetage morpho-syntaxique, PCA, clustering, etc.). Il s'agit d'une revue d'outils. Ces outils sont extrêmement nombreux. En réaliser l'inventaire exhaustif serait une tâche impossible. La revue ci-dessous permet de saisir l'ampleur des travaux déjà menés et le dynamisme de ce domaine. Les classifications sont parfois arbitraires puisque certains outils offrent des fonctionnalités de plusieurs classes. Le tableau récapitulatif 1.8 permet d'établir un classement plus complexe.

Non restreints aux textes. Ces outils proposent des analyses de données diverses : PCA linéaires, ou des clusterings de type k-moyennes (section 4.2.2) sans se restreindre à l'usage de sources textuelles : *ELKI*[a], *ClustVis*[b], etc. Bien que non-spécialisés dans le texte, ces outils clé en main sont des incontournables de l'analyse de données.

Constitution de corpus et formalisation de ressources. *Hyphe*[c] [7] (Jacomy *et al.*, 2016) est un *web-crawler* avec de nombreuses options pour notamment explorer sans tomber dans les grands attracteurs du web (Google, Wikipédia, etc.).
Le Trameur[d] est un outil de ventilation de texte en sous-parties. Il permet de constituer un corpus à partir de sources non structurées. Il constitue la trame et cadre d'un corpus et intègre des fonctionnalités d'analyse statistiques et d'alignement.
NooJ[e] est un logiciel de constitution de ressources linguistiques : filtres grammaticaux et dictionnaires à large spectres (flexions). Il étudie les textes au niveau orthographique, lexical, morphologique et syntaxique.

Visualisation de statistique. Il s'agit d'outils orientés *close-reading* et visualisation de statistiques simples. Ces outils offrent des possibilités d'analyse visuelle. *Voyant Tools*[f] [8] (Sinclair et Rockwell, 2014) propose diverses visualisations d'analyses de document. Il est conçu pour les pages web (via leurs url). De ce fait, son usage combiné avec un outil de constitution de corpus web (*web-crawler*) est intéressante. Il permet par exemple : comparaison de distributions de mots dans un corpus, dans un texte, co-ocurrences de mots dans un corpus, extraction des grands thèmes d'une page web (avec LDA). De nombreux travaux de recherche l'utilisent. Il fonctionne avec de nombreuses extensions.
Google n-gram[g] est un logiciel en ligne qui compte pour chaque année de publication les occurrences de motifs (wilcards et non regex) demandés parmi les textes d'une grande base de textes représentative de la production mondiale (Google Books).
TextDNA[h] est un logiciel destiné à visualiser les segments de textes comme des séquences d'ADN (figure 1.33a) à partir des résultats de Google n-gram par exemple. *KH Coder* étudie la distribution des mots dans un texte. *LATtice* (litterature at the level of the sentence) fait du co-clustering de phrases.
TXM[i] (Heiden, 2010) [9] vise à étudier les caractéristiques du discours via une batterie d'analyses statistiques notamment issues de la linguistique ainsi qu'à gérer le corpus sur la base d'éventuelles métadonnées. On trouve par exemple des statistiques sur la longueur des phrases, sur les pronoms utilisés, sur les conjugaisons, sur les variations lexicales, sur la distribution d'un mot, ou sur les co-occurrences de 2 mots.

Extraction d'expressions complexes. Le projet MONK (Ruecker *et al.*, 2009)[j] propose une interface et une méthode pour sélectionner des skip-n-grams à partir de tris sur n-grams d'ordre (n) variables et manipuler des textes en TEI.
BioTex[k] est un logiciel d'extraction d'expression complexe dans le domaine biomédical dans différentes langues.
FastKwic[l] est un logiciel d'extraction de terminologie pour l'indexation et la création de concordancier en français ou anglais.

7. développé par le Medialab de SciencePo Paris, sous la houlette de Bruno Latour
8. *Voyant Tools* revendique plus de 10^9 utilisations en 1 mois (octobre 2016) est totalement international, porté par Huma-Num et ses équivalents Allemand (DARIAH-DE) et Italien (CNR ILC).
9. TXM, un projet de l'ENS Lyon, a bénéficié de plusieurs ANR depuis 2007 et a su fédérer une communauté d'utilisateurs.

Il est développé par le CNRTL. *AntX*[m] est une suite de logiciels pour la création de corpus disciplinaires, de concordancier, l'extraction de terminologie, pour l'analyse de l'architecture de textes ou de corpus. Fonctionne avec de nombreuses langues.
TAPoRware[n] est un logiciel pour l'extraction de termes et l'analyse des extractions : co-occurrences, acronymes, statistiques, concordances, etc.

Alignements et segments répétés. Ces logiciels permettent d'aligner des textes pour suivre les modifications d'édition et les invariants. Lorsque l'alignement se fait en plusieurs langues, ils peuvent aider à la traduction.
Unitex/GramLab[o] est un logiciel avec interface graphique destiné à extraire la terminologie, à assister la construction d'expressions régulières et à aligner des textes pour la traduction notamment. Certaines de ses fonctionnalités utilisent des sources extérieures : dictionnaires, règles grammaticales, etc.
TermSuite[p] est un logiciel issu d'un projet de recherche, au développement toujours actif, qui permet d'extraire la terminologie de textes en plusieurs langues, de classer les résultats (pondérés) et de procéder à l'alignement des textes.
Sketch Engine[q] est un logiciel propriétaire (et payant) qui permet également d'extraire une terminologie en plusieurs langues, de classer les résultats (pondérés) et d'étudier de nombreuses caractéristiques des textes. Il propose de nombreuses fonctionnalités pour les lexicographes et les linguistes.
Lexico5[r] permet l'alignement de segments, l'analyse en composantes principales de chapitres et d'autres opérations statistiques.
D'une autre manière, le projet *Medite*[s] (Fénoglio et Ganascia, 2008) propose des paramètres pour comparer des textes sur base de segments identiques, il fait partie du projet ObViL, et ne concerne pas la traduction mais les versions de textes littéraires.
ConcQuest[t] est un logiciel expérimental permettant la création d'un concordancier multilingue d'expressions complexes, ce concordancier est facilement requêtable.

Classification descendante. Dans la même veine d'analyse statistique sur des segments : *Tanagra*[u], *IRaMuTeQ*[v] ou Alceste[w] (Reinert, 1999) permettent une « Classification Hiérarchique Descendante » (cf. section 4.2.1) sur un tableau croisant les formes pleines et des segments de textes. Il s'agit de classer des segments de textes de longueur homogène en fonction d'oppositions qu'ils entretiennent. Les oppositions sont calculées sur les mots contenus dans les segments.
À la base créée pour les données moléculaires, *SplitsTree4*[x] construit des similarités de type phylogéniques sur de nombreux types de données.

Graphes. De nombreuses méthodes construisent des graphes de mots-clés (co-occurrence) à partir de textes, comme Textexture[y] (Paranyushkin, 2011) avec un algorithme similaire à celui présenté dans le pseudo-code 1 avec des fenêtres de 3 mots.
WORDij[z] calcule les relations entre les mots d'un texte. Il construit des graphes pondérés de co-occurrence de mots.
Linkurious[aa] est un logiciel d'exploration de graphes multidimensionnels construits à partir de sources textuelles (figure 1.33b). Il permet le travail de groupe et est conçu pour détecter des anomalies.
Netlytic[ab] est tourné vers les données textes issues de réseaux sociaux et produit des graphes de thématiques ou de liens entre individus.
SCITE[ac](Riel *et al.*, 2008b) propose des "topic maps" de textes. Ces sont des graphes de textes partageant des mots clé préalablement extraits. L'implémentation concerne exclusivement des corpus d'articles scientifiques. Nyffenegger *et al.* (2016) produit une expérience similaire de cartographies thématiques (graphes) à partir d'un corpus annoté.

Analyse multi-dimensionnelle. Textobserver[ad] s'intéresse à l'interaction avec l'utilisateur pour la visualisation d'analyses multidimensionnelles. *Prospero*[ae] permet des analyses diachroniques de corpus annotés.
Palladio[af] est un outil en ligne proposé par l'université de Stanford pour créer et visualiser des données historiques à caractère spatial. Il est intégré au projet Mapping the Republic of Letters (figure 1.32).
Le projet Obvil (Observatoire de la vie littéraire) propose de visualiser sous différentes formes des corpus de littérature modélisés en TEI. Plusieurs centaines de documents sont modélisés. Les annotations sont exploitées dans de nombreuses combinaisons : graphes de co-occurrences, frises chronologiques, etc.

Annotation et exploration de textes. *UAM CorpusTool*[ag] est un logiciel d'annotation de texte pour des études directes ou pour produire des corpus d'entraînement destinés à un algorithme avec apprentissage supervisé.
Glozz est un logiciel issu d'une ANR, dont le développement est maintenant arrêté. Il permet d'annoter et d'explorer des corpus de textes.
ANNIS[ah] est un logiciel pour l'exploration de l'architecture de corpus complexe (multi-couche), potentiellement annoté, et multilingue.
CATMA[ai] (Computer Assisted Textual Markup and Analysis) [10] permet principalement de produire (annoter) et gérer des textes

10. CATMA est un projet allemand et international (BMBF, *German Ministry of Science and Research*), ré-implémentation de *TACT* en 2008

en TEI. De nombreuses autres propositions logicielles existent, notamment pour annoter le web (PressForward, Hypothes.is, etc.), mais CATMA a la particularité de se focaliser sur la TEI.

Entités nommées, compréhension de textes. Ce domaine s'éloigne de la textométrie et implique clairement de l'apprentissage supervisé. Ce sont les géants du traitement du langage qui proposent des intelligences artificielles capables de comprendre des questions et d'y répondre. Ils proposent également des analyses de sentiments. Ils fonctionnent tous sur calculateur distant avec une API. Ils sont non-libres mais gratuits jusqu'à un seuil d'utilisation. Les entités nommées extraites sont mises en lien si possible avec des bases de connaissances extérieures (type DBpedia, geonames).
GoogleAPI Natural Language[aj], *Dandelion API*[ak] et *Ambiverse*[al] proposent de classer les entités et d'extraire les sentiments (entités et blocs) dans plusieurs langues.
TextRazor[am], *NetOWL*, et *Rosette*[an] proposent de classer finement les entités et d'analyser les relations qu'elles entretiennent. *Rosette* propose également de la traduction.
Aylien[ao] propose de nombreux outils similaires aux précédents, et intègre des outils de résumés automatiques, de *webcrawler* et de lecture de flux pour analyser les nouvelles sur internet sans développement annexe.
Watson est une solution proposée par IBM, qui a récemment racheté *AlchemyAPI* spécialisé en traitement du langage et propose des services similaires aux précédents. Le traitement du langage est intégré au sein d'un service plus large d'intelligence artificielle.
NERD est un service web qui propose de requêter différents services d'extraction d'entités nommées pour comparer les résultats.

Autres. De nombreux outils originaux existent également, pour des disciplines de niche, par exemple *Poemage*[ap] qui permet la visualisation de topologie sonore pour les poèmes (rimes par exemple).

Frameworks. De nombreux frameworks sont utilisés dans le développement de projet « one-shot » pour l'analyse de corpus suivant des demandes spécifiques : General Architecture for Text Engineering (GATE), Unstructured Information Management Architecture (UIMA), Natural Language Toolkit (NLTK), SpaCy, Apache OpenNLP, Stanford CoreNLP, etc. Ce sont des acteurs incontournables du développement de logiciels et de solutions pour l'analyse de textes. Ils sont également très impliqués dans la recherche (nombreux projets européens et nationaux).

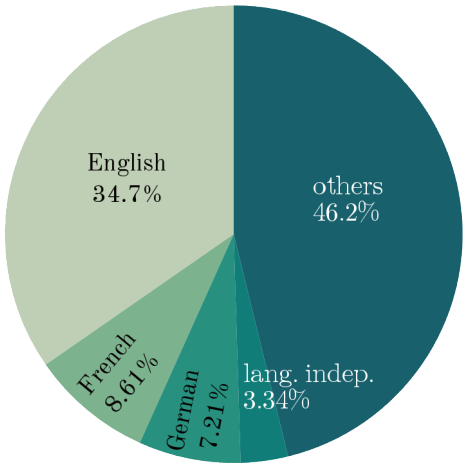


FIGURE 1.34 – Répartition des ressources linguistiques dans les principales langues (LREC map 2010)

Veilles. La profusion des projets de logiciels d'analyse, surtout dans le domaine de la recherche, donne lieu à quelques ambitions de fédération ou recensement. La revue précédente s'inspire de quelques résultats de ces recensements et propose de nombreux nouveaux logiciels à ces listes (qui, à la différence de Wikipedia, ne permettent pas la contribution). Mentionnons ici 3 travaux :

— le recensement du consortium CORLI[aq] de Huma-Num assez lacunaire
— le travail de veille du groupe de recherche canadien TaPoR[ar] très complet et actif
— le récent ouvrage intitulé « Littérométrie » de Bernard et Bohet (2017) qui propose une bonne introduction aux possibilités de traitements et évoque quelques outils.

La veille de référence du domaine du traitement des langues (qui semble ignorée des autres ressources) est la *LREC Map*, qui depuis 2010 cartographie l'ensemble des ressources (outils, librairies, standard, mécanismes) utilisées ou développées par les participants de cette conférence. LREC (International Conference on Language Resources and Evaluation) est une référence internationale. L'initiative est doublée de la création d'un identifiant unique (ISLRN) qui identifie chaque ressource. Cette veille

permet notamment d'estimer les ressources disponibles en fonction des langues. Ainsi l'état des travaux en 2010 annonce que l'anglais, le français et l'allemand sont les langues les plus ciblées (figure 1.34).

On note aussi une rupture des propositions de recherche avec les services en ligne d'analyse détenus par de grandes entreprises. Ces dernières très dynamiques, s'éloignent du *close-reading* et des humanités mais proposent des outils indéniablement issus de recherches poussées en traitement du langage.

5 Conclusion : Les méthodes existantes au défi des humanités

Nous reprenons ici les verrous repérés au fil de cet état de l'art. Pour chaque verrou nous proposons des hypothèses et des contraintes qui guideront la mise en place de la proposition (chapitre 2).

Cloisonnement des pratiques et chaînage des connaissances. Ce cloisonnement entre données sources, expressions de plus haut niveau, raisonnements historiques ou automatiques, etc. (voir figure 1.2) est artificiel. Il occulte les projets qui parcourent plusieurs maillons de la chaîne. La rétro-conception et fabrication de la *salle à manger tournante de Néron* (Quantin *et al.*, 2017) que nous avons réalisée conjointement avec des archéologues, architectes, conservateurs et ingénieurs, illustre cette idée. En effet ce projet débute par l'acquisition de données 2D et 3D, ainsi que de sources historiques, passe par une organisation de ces sources, par des visualisations et par une analyse à la fois archéologique et mécanique, une phase de rétro-conception, une phase de diffusion des données et une phase de valorisation en musée avec un maquette issue de fabrication additive (Bernard et Barlier, 2015). De nombreux allers-retours entre les différentes étapes brouillent les couches *Data Information Knowledge Wisdom* (DIKW). De nouvelles connaissances et hypothèses émergent de cette chaîne, notamment sur le fonctionnement mécanique et la géométrie (voir figure 1.35).

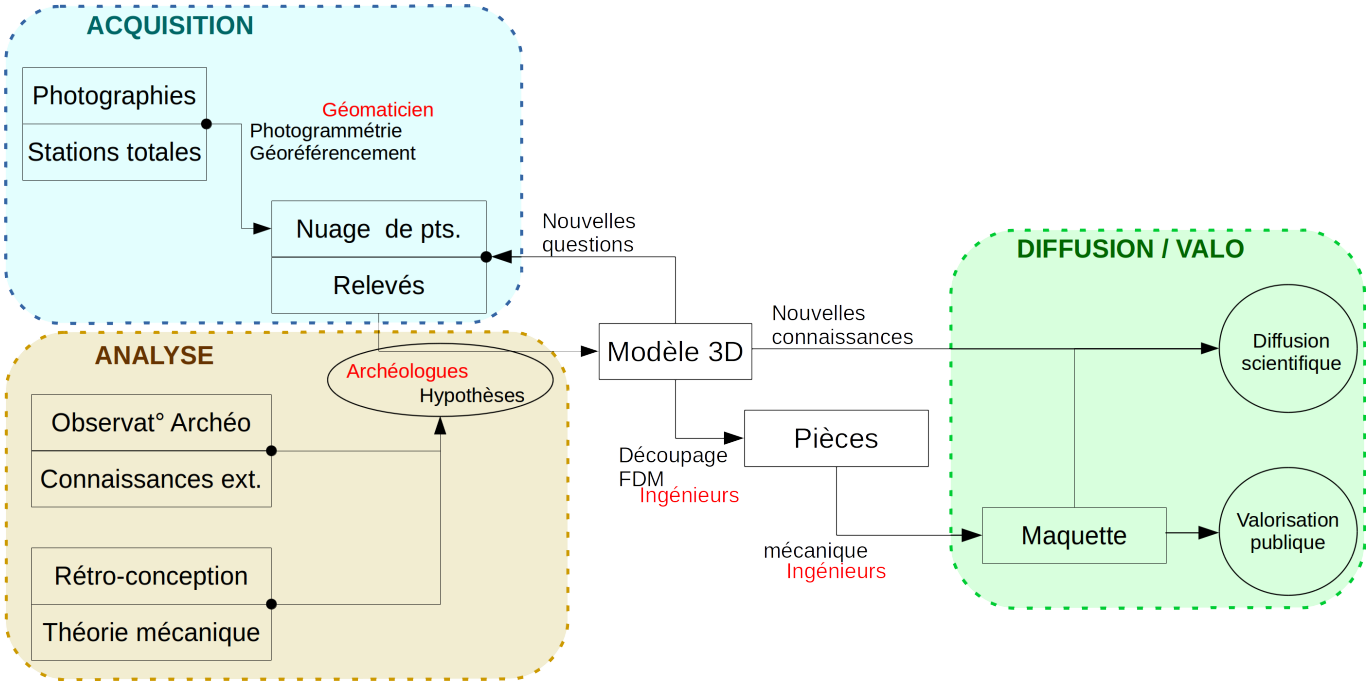


FIGURE 1.35 – Représentation schématique du flux de données dans le projet de la salle à manger tournante de Néron. Ce flux et cette organisation en cascade sont caractéristiques d'un projet en humanités numériques (processus pluridisciplinaire). Il serait adaptable à des projets n'impliquant pas la 3D.

De nombreuses solutions. Nous avons identifié un grand panel de solutions logicielles, de nombreux algorithmes d'extraction de terminologie et de nombreuses initiatives de standardisation et d'uniformisation.

La seconde vague (normalisation, interopérabilité) décrite par Burnard (2012) en section 1.1.1 est bien avancée, la troisième vague (apprentissage et calcul distribué) est opérationnelle du côté des très grosses structures et s'appuie sur les travaux de recherche récents. La première vague (foisonnement de traitement du texte) n'est pas éteinte pour autant. Les 3 vagues se superposent plus qu'elles ne se succèdent.

En effet, de très nombreuses approches numériques s'intéressent aux données textuelles brutes et dépassent l'enjeu soulevé par les humanités. Cette superposition des pratiques (les 3 vagues) sur un très large spectre (3D, ontologies, clustering, etc.) provoque un raz-de marée parmi les humanités qui y succombe totalement depuis quelques années.

Néanmoins l'histoire et le patrimoine ont des finalités propres qui – à leur tour – dépassent l'application en sciences des données. Nous tentons de comprendre les tensions entre ces enjeux, nous identifions également les manques.

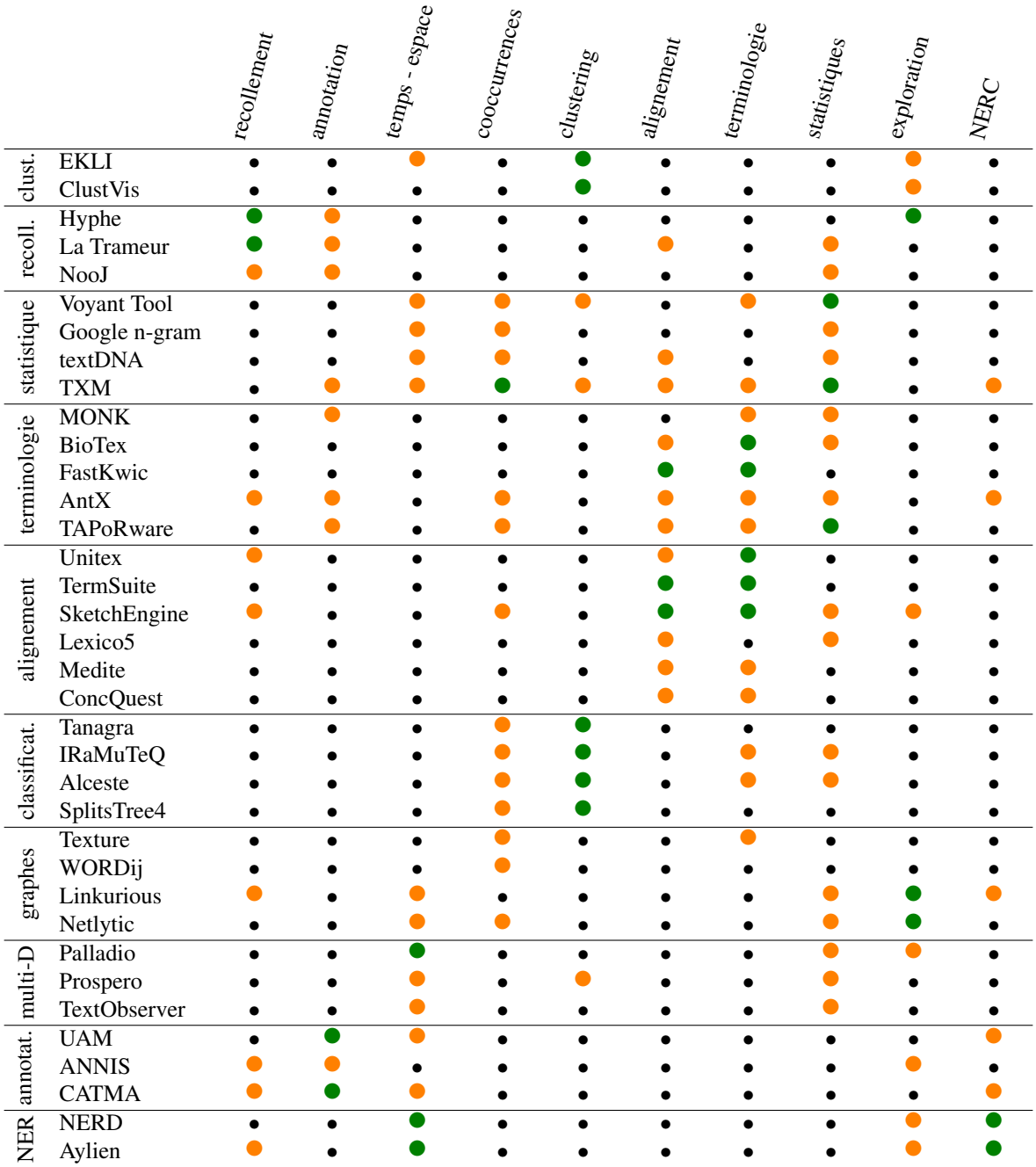


TABLE 1.8 – Comparatif d'une sélection de logiciels d'analyse de textes bruts.
● : cœur du logiciel ● : quelques fonctionnalités • : non proposé

Importance du lien entre 3D patrimoniale et données historiques. Les données historiques (sources) et la documentation (méthodes) ont une importance primordiale dans les projets de 3D patrimoniaux. Ils font la distinction entre la 3D scientifique et la 3D d'animation.

Les méthodes et outils liés à la 3D restent souvent centrés sur les données spatiales et éventuellement temporelles. L'ancrage des données est toujours réalisé via la géométrie ou l'espace. Il s'agit souvent « d'étiqueter le modèle 3D », de l'annoter. Les connaissances non spatialisables (ex : conditions de travail, contexte politique) sont souvent éliminées. Dans une conception externaliste de l'histoire et du patrimoine, cette faille constitue un problème scientifique majeur.

Au cours de la chaîne de traitement pour les projets en 3D, précision et finesse d'analyse des données, propres aux humanités, se perdent. En effet la chaîne de traitement implique de nombreux acteurs, matériels, logiciels tous porteurs de contraintes différentes. En conséquence du point précédent, l'historien n'exploite pas les résultats de projet de patrimoine numérique, même lorsque ceux-ci font une place à la dimension sémantique (ex : Nantes1900).

Tandis que des données scientifiques sont perdues, d'autres sont parfois créées de toutes pièces pour la valorisation. En effet les contraintes matérielles de la valorisation obligent à créer des objets sans aucune information scientifique (ex : supports, texture, complétion de mesh).

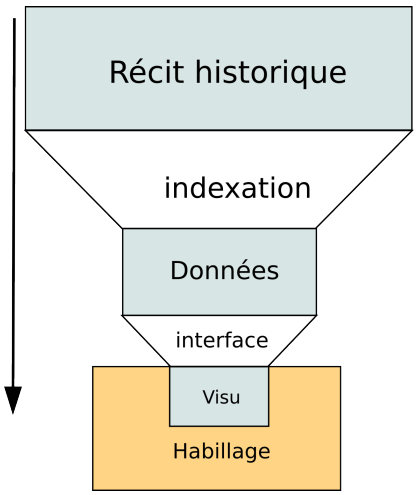


FIGURE 1.36 – Entonnoir des données, de la captation à la valorisation : perte des données à chaque étape, et ajout de données artificielles en bout de chaîne (valorisation, visualisation : textures, extrapolation, etc.)

Nous constatons un cloisonnement entre d'un côté les données et les analyses propres à l'Histoire (voir section "Des données textuelles aux connaissances explicites" (3)) et de l'autre celles propres au patrimoine (voir section "Patrimoine et numérique" (1.2)). Les passerelles existantes ne sont pas numériques. Les effets décrits ci-dessus sont dénommés « entonnoir » des données : de la captation à la valorisation. La figure 1.36 représente cette idée. Face à ce constat, cette thèse fait une proposition scientifique.

5.1 Retour sur les verrous scientifiques

Structure de données. Nous identifions une attente forte en termes de documentation et d'analyse. Dans le cas de la 3D cette attente est émergente. En témoignent les nombreux projets de recherche en « sémantisation » du patrimoine 3D de l'année 2017 seule : Extended matrix, ReSeed, HBIM, Inception, etc. (cf. section 2.3.3). Dans le cas plus général du patrimoine, des outils stables sont déjà largement déployés : les outils officiels pour musées ou les nombreux CMS pour projets hors-institution ; aussi les schémas de structure et de valeurs sont établis. (voir section 2.3.2). Ce besoin de documenter et de connecter diverses sources de connaissances (3D et sémantique par exemple) est renforcé par une perspective externaliste, aujourd'hui dominante en histoire des sciences et patrimoine technique. De ce fait les structures de données classiques (ontologies/thésaurus principalement) sont massivement utilisées comme solutions, elles permettent la documentation et la connexion des ressources (cf. section 2.2) dans une approche qui respecte 2 des défis énoncés mais en soulève d'autres.

La plupart des structure de l'état de l'art répondent à 2 verrous :

— verrou 6 (**approche multi-échelles**)
— verrou 7 (**approche multi-dimensionnelle**).

Mais elles en relèvent pas entièrement les défis de l'analyse pour l'histoire et le patrimoine, en effet :

— verrou 2, **non-restriction du domaine d'étude** : les classes et relations sont pré-établies, ce qui ne correspond pas à une démarche d'historien. Le développement d'une ontologie spécifique à chaque corpus analysé serait chronophage, en rupture avec la notion de standard et peu intéressant pour dégager de nouvelles connaissances historiques ;
— verrou 4, **logique floue** : les standards des ontologies du patrimoine ne permettent que des relations binaires [11], tandis

11. la relation existe ou n'existe pas

que les humanités, nuançant leurs propos, se rapprochent davantage de la logique floue, des relations pondérées [12]. Les ontologies floues du patrimoine restent à établir (voir section 2.2), ce serait le rôle d'un consortium (type CIDOC), hors champs de cette thèse.

Instanciation des structures de données. Une structure de données sans instances et sans utilisation ne serait qu'une démonstration intellectuelle, inutile en HN. Riel *et al.* (2008b) évoque également cette limitation et évite les ontologies. La constitution de ces sources de connaissances est un sujet aux enjeux dépassant les humanités. Des projets gigantesques s'y penchent à travers le monde (DBpedia, Google KnowledgeGraph, NELL, etc.). Peupler les structures de données (souvent des ontologies mais pas nécessairement) est réalisé soit manuellement, soit automatiquement. Cela soulève les problèmes en lien avec les verrous suivantes :

— verrou 3 (**unicité et unité du corpus**) : l'instanciation automatique de structure pré-établies fait nécessairement appel à des algorithmes avec apprentissage supervisé. Or l'utilisation de données extérieures ou d'une partie pour le tout, (phase d'entraînement nécessaire aux algorithmes supervisés) est en contradiction avec le verrou énoncé.
— verrou 1 (**patrimoine dynamique**), l'instanciation « manuelle », précise, mais fastidieuse et avec un fort manque de rappel (faux négatifs) demeure la solution la plus utilisée dans les humanités. La mise à jour permanente des données du patrimoine demanderait un travail fastidieux permanent, en contradiction avec le verrou énoncé.

Nous avons également énoncé une critique de l'utopie de la bibliothèque universelle comme but unique. La bibliothèque universelle motive le développement des modèles de données très structurés. Nous rejetons cette logique. Nous estimons que l'usage doit guider la mise en œuvre, et nous déclarons que l'objet de cette thèse n'est pas la communication mais l'analyse. Alors, pour cette thèse, nous nous autorisons à refuser le support des ontologies et modèles de données trop structurés.

Apprentissage automatique et graphes libres. Nous nous éloignons donc des ontologies et nous nous tournons ver les graphes libres, l'extraction de terminologie et le clustering par apprentissage automatique. Ces méthodes plus « bas-niveau » ne s'intéressent plus directement à la représentation des *connaissances* mais au traitement de l'*information* capable de produire des connaissances après *interprétation*. Il s'agit d'une piste prometteuse pour l'analyse de contenu historique (cf. section 4). En effet ces techniques permettent de conserver la complexité du texte, elles respectent les verrous suivants :

— verrou 3 (**unicité et unité du corpus** : elles n'introduisent pas de biais ;
— verrou 2, **non-restriction du domaine d'étude** : elles n'introduisent pas de classes construites en amont ;
— verrou 4, **logique floue** : elles évitent les arcs binaires en quantifiant les relations entre items.

Nous avons alors identifié une série de précautions qui permettraient d'améliorer les solutions existantes, tout en satisfaisant les verrous pré-cités et en répondant aux critiques des humanités numériques (section 1.1.3). Ces précautions consistent notamment à intégrer les verrous respectés par les modèles structurés dans les graphes libres :

— verrou 6 **approche multi-échelles** : donner accès à plusieurs niveaux de lecture : produire une vision globale d'un corpus et permettre d'investiguer une relation jusqu'à la source. Ceci préserve la qualité des contenus et offre une traçabilité de l'information, la possibilité d'étudier les sources et les transformations.
— Favoriser l'interaction avec l'expert du domaine, seul garant de la qualité des contenus produits, au détriment des inférences automatiques, souvent pauvres au regard des capacités et des connaissances qualitatives de l'historien.
— Permettre l'intervention de l'historien.
— Dépasser la représentation orthogonale des mots (verrou 5 **représentation des mots**) pour prendre en compte la proximité sémantique entre mots en fonction du contexte d'usage local, au sein du corpus (respect des contraintes du verrou 3).

12. La relation existe avec telle intensité ou probabilité

Chapitre 2

Proposition scientifique : Haruspex

> « Ne cherche pas les significations, compte les mentions. »
>
> Karen Spärck Jones

Contents

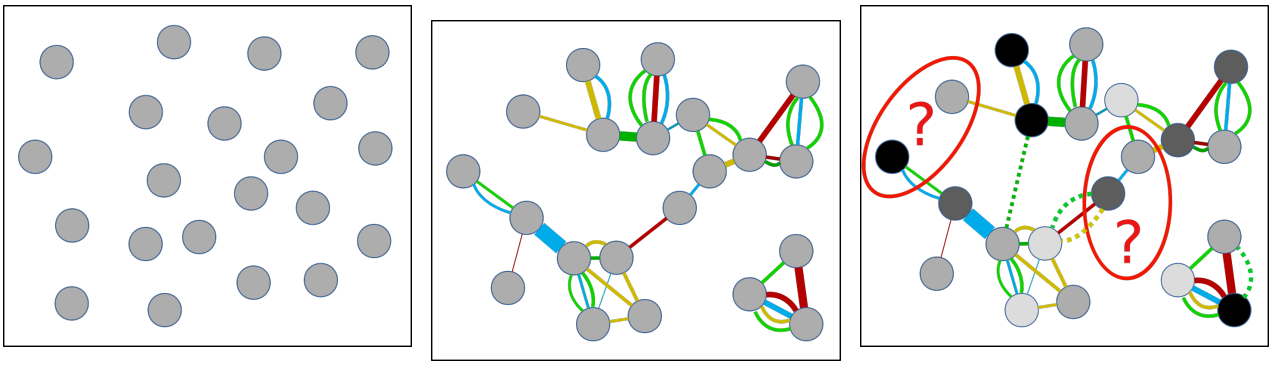
On pourrait plagier la première partie du titre d'un article de Pierre Mounier pour ce chapitre : « Du discours aux données... », à condition de plagier la seconde partie du même titre pour le chapitre suivant (3) « ...et retours » [1] (Mounier, 2011).

Face à la production de textes techniques, riches et non-structurés, nous proposons *Haruspex*, un outil d'analyse et d'intégration de connaissances historiques. Le chapitre précédent (chapitre 1) place les enjeux et les défis à relever. *Haruspex* ne requiert ni classe de vocabulaire cible, ni modélisation de données *a priori*, ni supervision pour entraînement. Indépendant de tout domaine, il opère sur des corpus uniques, c'est-à-dire sur des corpus dont on considère **le contenu seulement mais tout le contenu équitablement**. Il a été conçu pour calculer des proximités entre textes.

En sortie, on obtient une base de données (multi-)graphe (flou) avec les textes en nœuds et les proximités en arêtes (floues non orientées). Des requêtes sur le graphe permettent des résultats quantitatifs et visuels, supports d'interaction avec l'historien, expert du domaine.

1 Introduction

Haruspex est un *pipeline* de type Extraire Transformer Charger (ETL) au sens large, c'est-à-dire un processus de transformation de données et de chargement dans une base de données. Les transformations de données ici sont profondes et impliquent des algorithmes de TAL. En effet : Les données d'entrées sont des textes bruts ou faiblement structurés (figure 2.1a) ; La sortie est un multi-graphe flou de proximités entre les textes d'entrée (figure 2.1b). Ce graphe est alors analysé pour identifier des anomalies ou « chaînons singuliers » qui intéressent le spécialiste du corpus (figure 2.1c).



(a) Corpus de textes homogène, découpés en unités lexicales

(b) Réseau pondéré de textes, liens de thématique(s) commune(s) (proximité)

(c) Analyse du graphe : analyses visuelles et quantitatives : identification de chaînons singuliers

FIGURE 2.1 – *Haruspex* : une méthode de calculs de proximité entre documents et d'analyse de corpus de textes

1.1 Usages, contraintes, hypothèses et objectifs

1.1.1 Contextes d'utilisation

Haruspex est conçu pour assister l'historien dans l'étude d'un ou plusieurs corpus de textes pré-établi(s) en calculant des proximités de contenus entre textes.

Indépendant de tout domaine, en théorie, *Haruspex* peut saisir n'importe quel corpus de textes techniques. Il a néanmoins été conçu pour les corpus en histoire des sciences et des techniques. Les chapitres 3 et 4 recensent quelques usages des résultats de la méthode proposée.

Parmi les problèmes cités par la littérature, *Haruspex* pourrait notamment intervenir pour améliorer :

— les systèmes de recommandation basés sur les contenus (*content based recommendation systems*) pour les textes par exemple, dédiés au *reviewing* d'article scientifiques (Protasiewicz *et al.*, 2016), ou à la constitution de bibliographies (Boonyasopon *et al.*, 2011; Riel *et al.*, 2008b).
— la phase de création de vecteur document pour des applications en patrimoine culturel (Collao *et al.*, 2003).
— Les analyses de cycle de vie produit, notamment l'analyse de contenus non formatés dans la chaine de production comme envisagé par Kassner *et al.* (2014)
— la recherche de cas similaire dans des bases de données de rapport, notamment lié au management des risques (Zou *et al.*, 2017).
— les tâches classiques d'exploration de brevets (*patent-mining*) (Souili *et al.*, 2015)
— la détection fine de plagiat parmi des corpus limités via l'étude de paraphrase (ré-écritures).

1. j'ai retiré le point d'interrogation et accordé retour au pluriel

1.1.2 Contraintes

Le chapitre précédent fixe les contraintes pour un usage en histoire, nous les récapitulons ici :

— Contenus non structurés : les textes en entrée sont bruts.
— Métadonnées : toute métadonnée doit pouvoir être prise en compte, même hors standards de formalisation
— Unité du corpus : aucune partie du corpus ne peut être considérée comme représentative de l'ensemble.
— Unicité du corpus : les biais extérieurs sur les contenus du corpus doivent être minimisés.
— Qualité des données : certains mots peuvent être hors-dictionnaire (néologisme, jargon technique), mal orthographiés, ou manquant (OCR).
— Nuances : Les liens entre les éléments doivent être nuancés (pondérés) et ces pondérations doivent pouvoir être investiguées.

1.1.3 Hypothèses

Le chapitre précédent développe la construction de ces hypothèses à partir de l'état de l'art. Nous les récapitulons ici.

— La notion de récit est fondamentale en Histoire, les faits explicites ne sont pas directement de l'histoire.
— Les capacités de formalisation des connaissances qualitatives sont trop faibles. L'interaction avec l'historien est le seul moyen de produire des connaissances historiques.
— Il existe un intérêt à compléter la lecture qualitative d'un corpus textuel, par une analyse quantitative.
— L'analyse d'un corpus peut se passer de contenus et de schémas de métadonnées en entrée
— La création manuelle de métadonnées est extrêmement fastidieuse et doit être évitée.
— L'historien est principalement confronté à des corpus de moins de 20 000 pages.
— Le corpus est connu qualitativement par l'historien
— L'éthique des humanités évite les processus boite noire.

1.1.4 Objectifs

Objectifs. L'objectif de *Haruspex* est de calculer des distances entre des unités de texte d'un corpus. Ces proximités permettent de créer des graphes multiples pondérés. Les proximités sont multi-échelles, de la vision d'ensemble à l'analyse approfondie de parties identifiées comme anomalies. Certaines anomalies peuvent être identifiées automatiquement. D'autres anomalies ainsi que les régularités sont à détecter via des représentations visuelles.

La liste suivante illustre les objectifs en termes d'analyse historique, et permet de saisir quelques cas d'utilisation. Cette liste est indicative et ne constitue pas l'exhaustivité des défis auquel *Haruspex* pourrait répondre. Elle est plutôt un extrait des questions pragmatiques rencontrées lors de séances interdisciplinaires avec des historiens. Ces questions ont guidé le développement d'*Haruspex* et le différencie d'autres outils de l'état de l'art (Chapitre 1).

— *Proximité multi-échelle* : Quels documents sont les plus proches d'un document donné ? à quel point sont-ils proches ? Que partagent-ils ? Quels sont leurs thématiques communes ?
— *Intra-corpus* : Quels sont les *outsiders* ou les *leaders* d'une thématique, comment sont-il reliés aux autres de la thématique ? du corpus ?
— *Multi-dimensionnel* : Certaines thématiques sont-elles contingentes (dépendante d'un temps ou d'un espace) ? Comment varie le *leadership* d'une vue du corpus à l'autre (par exemple derrière un filtre sémantique) ?
— *Contenus* : Les contenus des textes *outsiders* répondent-ils à une logique particulière ?
— *Co-occurrences* : Existe-t-il des dépendances entre certaines phrases-clés (co-occurrences) ? Est-ce contingent ?
— *Connectivité* : Quelle proximité entretiennent ces phrases-clés ou ces topics ? comment le quantifier ? Si ce sont des ensembles disjoints, observe-t-on une connexion indirecte récurrente ?
— *Anomalies* : Y a-t-il des anomalies (ex : hors-sujets, forte connexion contingente) ? comment les qualifier ?

Ce n'est pas. *Haruspex* n'est pas un outil de formalisation des connaissances. Il n'est pas destiné au partage d'information, il est peu compatible avec les technologies web-sémantique. Il n'est pas non plus un outil permettant de comprendre un corpus sans l'avoir lu.

Il permet néanmoins une abstraction de contenus textuels en représentation de plus haut niveau. Pour cela il produit des (hyper-)graphes (flous) et utilise certaines de leurs propriétés. Dans la mesure du possible les informations issues de bases de connaissances établies sur le web (web-sémantique ou non) sont exploitées.

1.2 Proposition

Étapes. Pour répondre aux objectifs et aux défis, en respectant les contraintes énoncées nous proposons *Haruspex*. Cet ETL diffère du TAL avec apprentissage supervisé. Il est décomposable en la série d'éléments suivants (cf. fig. 2.2) :

A1 La gestion du corpus, incluant une option de *topic-modelling* par NMF. Cette étape est présentée en section "Gestion de corpus" (2).

A2 L'extraction de terminologie non supervisée, grâce à *ANA+* une version améliorée de l'algorithme ANA (Enguehard et Pantera, 1995). Cette étape est présentée en section "Extraction d'expressions-clés" (3).

A3 Le post-traitement des expressions extraites, le calcul d'indicateurs assiste une modération manuelle, présenté en "Post-traitement des expressions" (4).

A4 Le calcul de proximités de documents pair-à-pair. Les mesures proposées incluent et améliorent les mesures classiques (tf-idf, cosinus, etc.). L'amélioration principale consiste à réfuter l'orthogonalité a priori des dimensions de vecteurs termes-documents. Cette partie est présentée en section "Création des liens entre pages" (5).

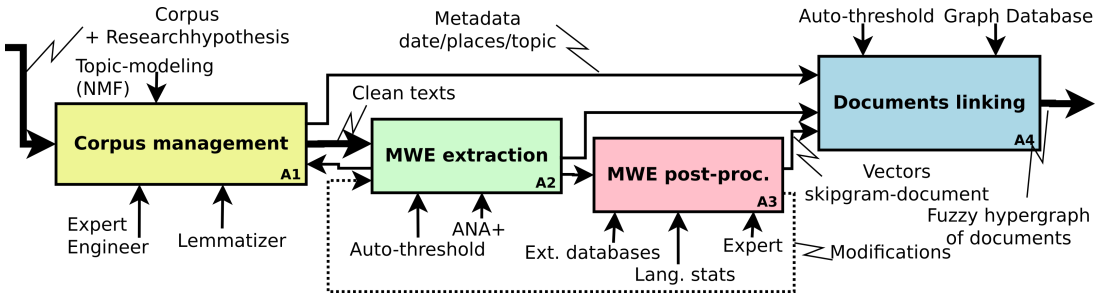


FIGURE 2.2 – SADT décrivant les 4 étapes du processus. Ce SADT vient préciser l'activité 1 du SADT en figure 4

Caractéristiques. D'après la classification établie par Turney et Pantel (2010), *Haruspex* fait partie des VSM (*Vector Space Model*) de type matrice termes-documents. Il est libre de toute pré-modélisation (pas de classes prédéfinies) et de toute dépendance à des ensembles de données étiquetées externes (Enguehard, 2005). L'extraction de *k-skip-n-grams* de longueurs variables (abrégés *skip-grams*) va au-delà des modèles sac de mots (mutliset) de n-gram d'ordres fixes, incapables de maintenir l'ordre séquentiel inhérent à l'information (Wang et Mccallum, 2005). Le graphe multiple flou en sortie correspond au *type V : crisp graph with fuzzy weights* selon la typologie établie par Blue *et al.* (2002).
On note parfois *multiwords expression* (MWE) pour les skip-grams. Les MWE sont définies par Sag *et al.* (2002) comme une acceptation large de composition lexicale : « *idiosyncratic interpretations that cross word boundaries* ».

L'ensemble du processus est modulaire, permettant de mettre à jour les composants avec d'éventuelles avancées de l'état de l'art.

Les résultats sont présentés en 2 temps. Ceux concernant uniquement la partie « performance informatique » et comparaisons d'outils sont en section "Résultats" (6), ils concernent principalement la F-mesure. Les autres résultats (graphes) étant incommensurables. Ces résultats (quantitatifs et visuels) qui concernent l'histoire et le patrimoine sont présentés dans les chapitres 3 et 4. Ces résultats répondent aux défis (section 1.1.4).

2 Gestion de corpus

Dans cette étape décrite par la figure 2.3, les mots du corpus sont simplifiés en lemmes, un étiquetage morpho-syntaxique apporte des informations linguistiques rudimentaires, le corpus est découpé en morceaux et des vérifications sur ces morceaux (*pages*) sont réalisées. En utilisant le *topic-modelling* (NMF), l'homogénéité du corpus est vérifiée, des mesures palliatives sont proposées. Chacune des parties peut alors recevoir des attributs (métadonnées) en fonction des besoins de l'historien. En sortie le corpus est alors explicitement constitué, enrichi et propre.

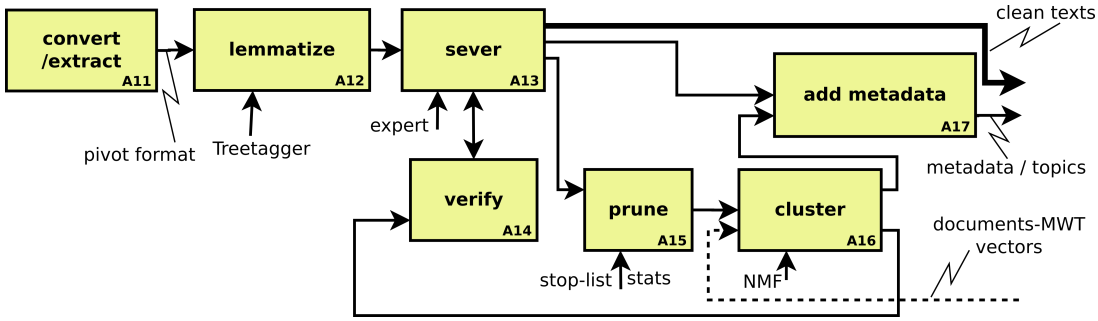


FIGURE 2.3 – Description des étapes de l'activité A1 Gestion de corpus

2.1 Étape préparatoire

Pour ces besoins le texte est converti en un format pivot qui consiste en une syntaxe LaTeX simple, assurant le découpage en unité documentaire et gestion de références.

2.1.1 Les *pages* : unité documentaire du corpus

Il s'agit de définir l'unité documentaire qui sera étudiée. En fonction des cas, cette unité documentaire peut être un fichier (texte) ou un sous partie de fichier (chapitre, section, paragraphe). En fonction de la qualité des données d'entrée, les résultats seront plus ou moins automatiques. Cette étape est primordiale : l'historien construit (ou plus exactement rend explicite) son objet d'étude.

Dans le cas de fichiers LaTeX le découpage est très bon, avec les fichiers TEI, odt, ou doc (MSWord) ce découpage est de qualité variable, souvent mauvais, dans le cas de fichier pdf ou text le découpage est très mauvais. Des options permettent de découper le texte par saut(s) de ligne (paragraphe) et d'agréger les paragraphes trop courts au plus long des paragraphes contigus. Quelques exemples d'objets d'étude :

— relations entre fichiers d'archives textuelles : l'unité documentaire sera le document, plus exactement le fichier texte issu de l'OCR.
— analyse de retranscription d'entretiens : l'unité documentaire pourra être la réponse à chaque question, dans ce cas certaines réponses à des questions pourront se lier avec des réponses à d'autres questions issues d'autres entretiens. Elle pourra aussi être l'entretien dans son ensemble, sans prêter attention aux questions. Des filtres (pattern d'expression régulière) retireront le texte des questions.
— relations qu'entretiennent les sous-parties de plusieurs mémoires : l'unité documentaire sera la sous-partie de mémoire. Un seuil permettra d'éviter les parties introductives non porteuses de contenu.

L'unité documentaire est par la suite appelée *page*.

2.1.2 Contenus extra-textuels

Pour l'analyse de documents hybrides textes/iconographie, cette première étape fait le tri. Seul le texte est conservé et les images sont extraites et mises en lien. Un système de références (pointeurs) similaires est mis en place pour la gestion de citations lors d'analyse de production scientifique (mémoire, thèse, article, etc.). Les références en les notes de bas de page / fin de documents sont supprimées du texte et mises en lien. Les références sont simplifiées et une structuration explicite permet de séparer l'auteur du titre de l'œuvre citée, un champ « autres indications » agrège toute sorte d'informations de la note de bas de page ou de fin de document. Les reprises de citations (*ibidem*, *op.ci.*, etc.) abondamment utilisées en histoire sont remplacés par des liens explicite vers la référence. Des alertes pointent les reprises de citations qui ne semble pointer vers aucune citation pré-éxistante.

Cette gestion très approximative des contenus extra-textuels (références et images) serait un point à améliorer. Concernant les images, le but est de conserver la position dans le texte (*page* associée) pour être en mesure de proposer des illustrations des contenus analysés par la suite. En d'autres mots, on considère que l'image est décrite dans le texte d'une *page* et que l'analyse du texte sera une analyse indirecte de l'image que l'on pourra donc associer. Concernant les références, leur extraction et simplification permet de créer un premier type de réseau entre *pages* : celui de citation. Ce type de réseau est classiquement produit pour analyser à un degré très superficiel la littérature scientifique. Ce réseau est une information supplémentaire pour notre analyse.

Cette étape permet également d'élaguer le corpus de tous les motifs lexicaux récurrents numéro de page ou en-tête par exemple dans le cas de documents PDF.

2.1.3 Lemmatisation et PoS

Cette étape est prise en charge par le lemmatizer *Treetagger* (Schmid, 1994). Les mots initiaux sont conservés, les lemmes et les étiquettes morphosyntaxiques (PoS) y sont associées. *Treetagger* est ici utilisé directement, sans entraînement particulier, ni découpage en mots spécifiques.

À l'issue de la lemmatisation, une matrice de vecteurs (compressée CSR) *pages*-lemmes est produite. Cette matrice filtre les lemmes contenus dans une *stop-list*, les lemmes avec une fréquence d'occurrence supérieure à 0.05 (5%), les lemmes occurrant moins de 5 fois (filtre de *term frequency*), et ceux occurrant dans moins de 3 *pages* (filtre de *document frequency*). Un filtre spécial est conçu pour les mots occurrant dans quasiment toutes les *pages*. Les *pages* ne contenant **pas** les mots (non filtrés jusqu'alors) présent dans plus de 90% des documents sont marquées comme contenant l'absence du lemme. Puis les lemmes occurrant dans plus des 3 quarts des *pages* sont supprimés. Un premier modèle vectoriel du corpus est ainsi produit.

2.1.4 Vérification

Cette étape fonctionne en 2 temps. Dans un premier temps, elle vérifie que les textes ne contiennent pas une proportion trop importante de caractères typographiques hors de l'alphabet français, signe d'une mauvaise OCR. Une rapide vérification concerne

la présence d'anglais dans le texte. Cette vérification est réalisée très simplement : le nombre d'occurrence de quelques termes anglophones de base (« the », « is », « of ») doit être faible. Une vérification de proximité entre les documents est également réalisée pour éviter les documents trop similaires. Cette vérification prend 2 formes : une distance cosinus entre les vecteurs document-lemmes, une vérification par fenêtre glissante de n-gram. Les distances cosinus anormalement faibles par rapport à la moyenne du corpus sont signalées, ce sont souvent des erreurs susceptibles de corrompre les résultats par la suite. Ce sont par exemple des documents contenant les mêmes annexes ou une version brouillon d'un document et sa version finale. La fenêtre glissante de n-gram est une technique classique de détection de paraphrase. Une fenêtre de 100 lemmes est construite et un motif d'expression régulière recherche ces mots dans cet ordre avec de potentielles ellipses (mot manquant). Une alerte signale les recouvrements de plus de 10 mots sur les 100. Cette technique est très rustique, mais elle est suffisante dans la mesure où nous ne cherchons pas à détecter du plagiat à ce stade. Nous cherchons uniquement à éviter les grosses erreurs qui influenceraient l'analyse. Comme indiqué dans la section 1.1.1, *Haruspex* dans son ensemble pourrait être un système de détection de plagiat.

Dans un second temps cette opération vérifie les résultats du clustering (voir section 2.2). Si les clusters sont très disproportionnés ou si la matrice des mélanges document-topic est très creuse, voire séparable alors il faut envisager de séparer le corpus en sous-corpus plus homogènes pour la suite du processus.

2.1.5 Les métadonnées

Si elles existent, les métadonnées du document sont récupérées. Par exemple dans le cas de document TEI ou PDF comportant des balises en en-tête ou associé dans le container sont récupérées : auteur, titre, date, etc. Si aucune métadonnée n'est présente ou détectée, alors nous tentons de récupérer ces données dans le nom du fichier et dans la première ligne du contenu. Les informations ciblées sont les dates et les noms propres. Dans tous les cas ces informations pré-remplissent un tableau que l'historien peut compléter. Le tableau est très libre : chaque ligne est une *page* chaque colonne une métadonnée. Il est alors possible de créer des colonnes selon les dimensions que l'on souhaite investiguer et les informations disponibles.
Une métadonnée additionnelle est produite par l'étape suivante : "Topic-modelling" (2.2). Les topics assignés aux *pages* sont enregistrés comme métadonnées (liste de numéros).

Cette étape est optionnelle, elle permet d'affiner les analyses et de perfectionner les connexions pour la valorisation de contenus.

2.2 Topic-modelling

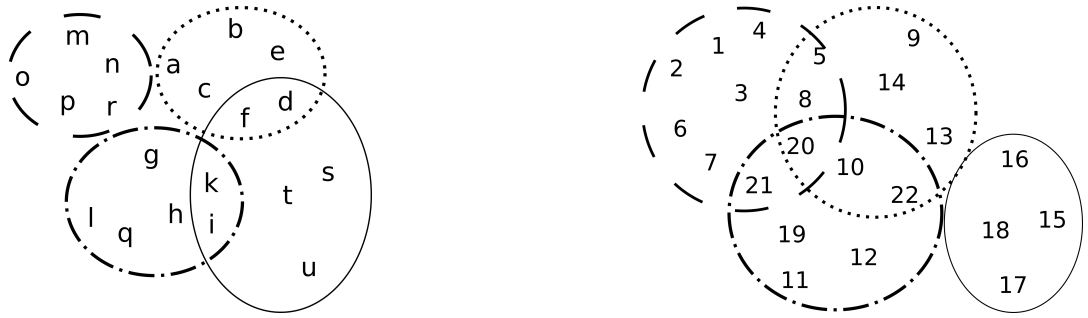
Cette étape du processus est optionnelle mais fortement conseillée. Elle peut être employée avant ou après l'extraction de MWE. Après l'extraction d'expression-clés, on considère que les documents sont représentés par un vecteur d'expressions.

2.2.1 Objectif

Objectifs. Cette étape de préparation du corpus consiste à chercher des catégories latentes dans les *pages* du corpus. L'objectif est triple. (1) D'abord il s'agit de vérifier l'homogénéité du corpus. L'obtention d'un corpus homogène à moindre coût permet de diminuer le bruit des étapes suivantes. En effet à partir de la même *page*, les expressions extraites dépendent du contexte : les autres *pages* du corpus ou sous-corpus (voir section 3. Les *pages* d'un corpus homogène sont peu séparables en topics (figure 2.4b), voire tous les documents comportent une part égale de tous les topics (alors les topics sont peu séparables). Si le corpus n'est pas homogène (il existe des ensembles de documents presque disjoints) alors on sépare le corpus en sous-corpus homogènes avant de passer à l'étape suivante. On travaille alors séparément sur chaque cluster. Concrètement le clustering est conseillé quand le corpus dépasse 3.10^6 mots et 200 *pages* (valeurs empiriques).

(2) Un objectif secondaire et complémentaire est de vérifier la qualité du corpus et éventuellement d'écarter les *pages* qui ne mériteraient pas de former un sous-corpus (hors-sujets), ou de repérer un topic de *pages* comportant une autre langue que le français ou l'anglais. (3) Le troisième objectif est de proposer un outil d'analyse à gros grain pour le corpus. En effet la contribution de chaque document à un topic et la définition de chaque topic en vecteur stochastique de lemmes peuvent produire une vue du corpus pour l'historien, améliorant sa compréhension, ou quantifiant des intuitions.

Choix d'une méthode. Les techniques de co-clustering sont jugées plus adéquates pour notre problème : vérifier la non-séparabilité des documents en topics malgré une séparabilité potentielle de topics (plusieurs topics distincts dans un document). Nous reprenons la notation décrite en état de l'art (section 4.2.3). À partir de notre matrice A, nous visons à obtenir un « topic-models » de 2 matrices de basses dimensionnalités : $A \approx H.W$, avec $W(n \times k$ avec $k \ll h)$ est une matrice *page*-topic et H $(k \times h)$ une matrice topic-lemmes (matrice des coefficients). Suite aux tests réalisés entre les 2 techniques les plus couramment utilisés : LDA et NMF, nous retenons la NMF. En effet cette technique permet un temps de traitement plus court et surtout, une meilleure séparabilité des topics, sans pour autant séparer les documents en topics. Ce choix pourrait être revu en fonction des avancées de l'état de l'art.



(a) Description des topics (ellipses) par des termes (caractères typographiques)

(b) Assignation des documents (numéros) dans des topics (ellipses)

FIGURE 2.4 – Plusieurs topics (ellipses) aux termes (lettres à gauche) distincts peuvent être présent dans une même *page* (chiffres à droite). Par exemple la *page* 8 contient 2 *topics* bien distinct. À l'inverse des *topics* n'ayant aucune *page* commune peuvent partager des termes issus de plusieurs *pages*

2.2.2 Nombre de topics

Le nombre de topics cible est un hyper-paramètre. Demander un nombre trop élevé de topics implique la création de topics étriqués, aux contenus similaires, donc avec de fortes intersections. À l'inverse, demander un nombre insuffisant de topics implique la création de topics au contenu vague et hétérogène, qui sont en fait des regroupements de topics ou de parties de topics. Pour ces raisons, trouver le bon nombre de topics est un enjeu important et non-trivial. Des travaux récents proposent des méthodes pour aider à détecter le « bon » nombre de topics. Nous les présentons ci-après.

Stabilité. Le critère de stabilité proposé par Greene *et al.* (2014) suppose que retirer quelques *pages* du corpus n'affectera pas la définition des topics. D'autres travaux se concentrent sur la connectivité de la matrice *page*-topics (Brunet *et al.*, 2004), ce qui implique l'hypothèse que les topics idéaux sont des topics de *pages* disjoints, c'est-à-dire que la matrice est initialement séparable. À l'inverse, la mesure de stabilité proposée s'intéresse à la définition des topics, la matrice topic-termes, ce qui suppose que les *pages* peuvent appartenir à plusieurs topics mais que ces topics doivent avoir un contour bien défini et stable.
Il propose alors une mesure de concordance entre les topics issus de 2 matrices topic-termes (voir équation 2.1). Cette mesure est l'Index de Jaccard moyen (AJ) sur la permutation optimale (π) maximisant l'Index de Jaccard entre les vecteurs topic-termes. En d'autre mot, il s'agit simplement de retrouver les paires de vecteurs topic-termes issus de 2 NMF et de calculer la variation moyenne.
La permutation optimale (π) (retrouver les paires de vecteurs) peut être trouvée rapidement en résolvant le problème des poids

	A1	A2	A3
B1	0.11	0.8	0.18
B2	0.58	0.15	0.07
B3	0.37	0.06	0.68

(a) Matrice des distances entre topics de 2 échantillons différents

	t11	t12	t13
t21	0	0.74	0.11
t22	0.47	0.09	0
t23	0.26	0	0.61

(b) À chaque ligne puis colonne, soustraire la valeur minimale

	t11	t12	t13
t21	0	0.74	0.11
t22	0.47	0.09	0
t23	0.26	0	0.61

(c) Identification des moindres poids (couleurs)

TABLE 2.1 – Exemple simple de la méthode hongroise. On note t1_ les topics issus du premier échantillonnage et t2_ ceux du second

minimaux (suffisant) pour créer un graphe biparti par la méthode Hongroise (Kuhn, 1955).

$$\text{concordance}(H_x, H_y) = \frac{1}{k}\sum_{i=1}^{k} AJ(R_{xi}, \pi(R_{yi})) \tag{2.1}$$

note : la formule originale de l'article est modifiée, car il semble qu'il y a une erreur : il s'agit de $AJ(R_{xi}, \pi(R_{yi})$ et non pas de $AJ(R_{xi}, \pi(R_{xi})$.
Il propose alors de procéder comme présenté dans le pseudo-code 2.
k_{min}, k_{max} sont les bornes du nombre cible de topics potentiels (définis par l'utilisateur)
Concrètement, les valeurs choisies sont souvent :

— τ le nombre de sub-corpus. $\tau = \sqrt{(n)}$ c'est-à-dire que le nombre de sub-corpus dépend du nombre de *pages* dans le corpus initial.
— β taux d'échantillonnage. $\beta = 0.9$
— k_{min} le nombre minimum de corpus est toujours 2, afin de vérifier qu'on ne peut pas couper le corpus en 2.

Algorithme 2 stabilité

générer τ sous-corpus (A_i ; $i \in [1, \tau]$) composés de $\beta.n$ pages ; $\beta \in [0, 1]$
pour $k = k_{min}$ **to** k_{max} **faire**
  (H_0,W_0) ← NMF(A_0)
  pour $i = 0$ **to** τ **faire**
    (H_i,W_i) ← NMF(A_i)
    S_i ← concordance(H_0, H_i) (voir eq. 2.1)
  fin pour
  S_k ← moyenne$_{i=1}^{\tau}(S_i)$
fin pour
choisir k qui maximise S_k : $k \in [k_{min}, k_{max}]$

— k_{max} est toujours inférieure à $n/4$ et inférieure à 35 ; elle dépend des suspicions. Empiriquement $k_{max} = 35$ (35 subdivision de corpus) est la limite de lisibilité.

Cette mesure peut être combinée avec d'autres mesures indépendantes, dans l'idée que l'intersection de plusieurs prédictions indépendantes est meilleure.

Divergence. La mesure proposée par Arun *et al.* (2010) suppose également que les topics doivent être définis par des ensembles de mots séparés, mais que les documents peuvent appartenir à plusieurs topics.
Il fait remarquer que si la matrice initiale contient uniquement le décompte des occurrences de chaque lemme dans chaque *page* (non-normalisée, donc non-stochastique) ; alors, la norme L_1 de chaque vecteur topic correspond à la proportion de chaque topic dans le corpus. Cette proportion de topic dans le corpus est simplement comptée en quantité, en nombre de mots captés par chaque topic.
Il démontre que la séparation des vecteurs topic-termes est optimale lorsque les valeurs singulières de la matrice topics-termes (H) ont la même distribution que les normes euclidiennes (L_2) des rangs de la matrice W. En effet lorsque les rangs sont bien séparés, les vecteurs sont orthogonaux, alors les valeurs singulières (axes de l'hyper-ellipsoïde dans l'espace de projection) sont ces mêmes vecteurs.
En utilisant la divergence de KL symétrique, l'auteur propose alors de comparer deux distributions :

C_H : la distribution des valeurs singulières de H
C_W : la distribution des valeurs normalisées du décompte de la proportion de chaque topic dans le corpus. Ce vecteur est calculé par $D * W$ avec D un vecteur $1 \times n$ contenant la longueur (en lemmes) de chaque document. Elle est proche de la norme L_1 de chaque vecteur topic.

$$\text{Divergence} = KL(C_H||C_W) + KL(C_W||C_H) \tag{2.2}$$

Lorsque la divergence est la plus faible, c'est-à-dire lorsque la mesure proposée par l'équation 2.2 atteint un minimum alors on obtient la valeur optimale du nombre de topics. En effet, si la matrice initiale n'est pas aléatoire, alors lorsque le nombre optimal de topic est atteint, la divergence augmente. Ceci est dû au fait que la valeur de C_W devient pénalisante, ajoutant un bruit correspondant aux probabilités (faibles mais non-nulle) que des (nombreux) lemmes soient constitutifs de plusieurs topics.

Concrètement cette mesure est moyennée sur plusieurs runs (entre 5 et 10) de la NMF pour chaque valeur de k topics cibles prévue dans l'intervalle (k_{min}, k_{max}).

Exemple d'application. Grâce aux méthodes précédemment énoncées, le nombre de topics cible idéal peut être déterminé. En effet, la complémentarité entre les mesures de stabilité et de divergence est souvent discriminante pour une valeur de k topics Déjà une première analyse *distant reading* peut être développée à partir des résultats de ces mesures. Un arbitrage humain entre la plus forte stabilité et la plus faible divergence est souvent nécessaire, plusieurs solutions sont parfois possibles. La figure 2.5 montre un exemple de mesure de stabilité et de divergence en fonction du nombre de topics. Dans ce cas il s'agit du corpus des expertises de l'ICOMOS relatifs à l'inscription des sites du patrimoine mondial de l'UNESCO. Ce corpus contient 1063 documents après un filtrage des documents anciens dont le contenu texte n'est pas accessible ou de très mauvaise qualité (OCR) ainsi que des documents bilingues. L'interprétation des figures 2.5 montre clairement que le nombre de topics cible doit être 8 ou 9. Pour les départager, on pourra regarder en détails les topics et identifier qualitativement si les lemmes définissant les topics sont plus cohérents avec 8 ou 9 topics.

2.2.3 Sparseness : les matrices creuses

Le corpus est une matrice de *pages*-lemmes, filtrée telle que décrite en section 2.1.3. On calcule que cette matrice est toujours très creuse. En effet, la sparsité est toujours supérieure à 0.4 et souvent à 0.6 d'après la mesure de Hoyer (2004) (voir equation 2.3). Cette mesure indique combien l'information est concentrée dans un nombre réduit de dimensions seulement. La figure 2.6a montre un vecteur pour lequel l'information est équitablement répartie sur ses 13 dimensions, tandis que pour le

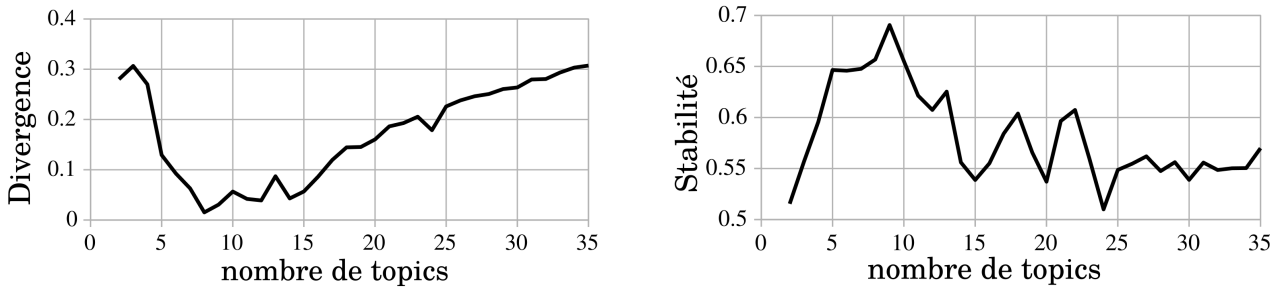


FIGURE 2.5 – Stabilité pour un nombre de topics compris entre 2 et 35 ; corpus des expertises de l'ICOMOS

(a) Sparseness = 0.1 (b) Sparseness = 0.4 (c) Sparseness = 0.7 (d) Sparseness = 0.9

FIGURE 2.6 – Représentation des valeurs (ordonnée) sur les 13 dimensions (abscisse) de 4 vecteurs ayant une sparseness très différente

vecteur représenté en 2.6d, presque toute l'information est contenue dans 2 dimensions seulement. Dans le cas d'une forte sparseness, on parle de matrices creuses (avec beaucoup de zéros). La sparsité d'une matrice correspond à la sparsité de ses vecteurs colonnes. La sparsité d'un vecteur X (un document) comportant n dimensions (lemmes) est donné par :

$$\text{sparsness}(X) = \left(\sqrt{n} - \frac{\sum_{i=0}^{n} |X_i|}{\sqrt{\sum_{i=0}^{n} X_i^2}} \right) \times \frac{1}{\sqrt{n} - 1} \tag{2.3}$$

La sparsité des résultats de la NMF peut également servir d'indicateur, mais souvent elle est difficile à interpréter. On remarque néanmoins sur la figure 2.7 que la valeur de la répartition des *pages* en topics change de variation vers 8 ou 9 topics. Au-delà les topics ne séparent pas mieux les *pages*, mais ce n'est pas notre objectif. Sur cette figure (2.7) on note également que la

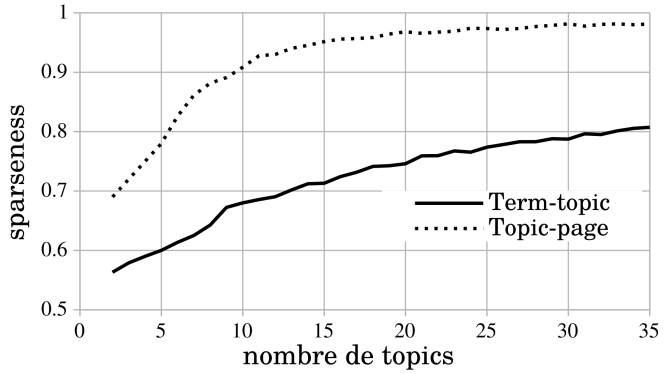


FIGURE 2.7 – Sparseness des matrices produites par la NMF

séparabilité des topics en termes n'est pas excellente. Empiriquement cette valeur est toujours inférieure à celle de la sparseness de la matrice W. Une interprétation serait que la bonne séparabilité des *pages* en *topics* est artificielle puisque les topics restent similaires (partagent de nombreux mots).

2.2.4 La factorisation

La factorisation à proprement parler est réalisée par une implémentation de l'algorithme de Shahnaz et Berry (2006), qui reprend l'algorithme de mise à jour des multiplications (voir section 4.2.3) avec une contrainte de pénalité sur la non-sparsité de la matrice H. Dénommée algorithme du gradient avec une contrainte sur les moindre carrés (GD-CLS), cette méthode prend pour objectif la sparsité de la matrice H, ce qui a pour conséquence d'améliorer (densifier) la localisation des topics dans les *pages* (matrice W).

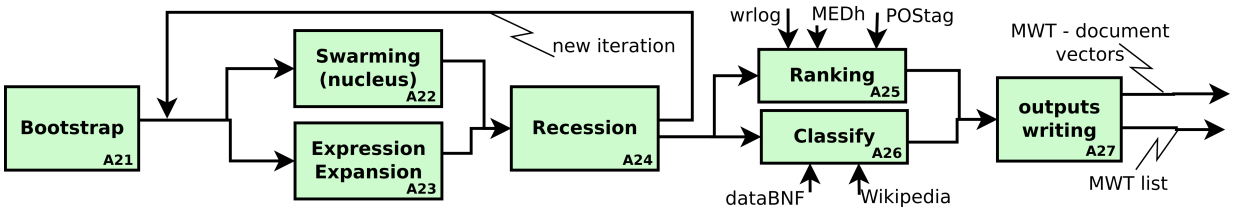


FIGURE 2.8 – Schéma SADT décrivant les activités contenues dans l'activité A2 : Extraction d'expressions-clés

3 Extraction d'expressions-clés

Choix. L'extraction de termes clés ou MWE a été étudiée en section "Extraction de terminologie" (4.1.1). De nombreux algorithmes existent, pour le français peu sont tolérants aux variations de composants : les skip-grams. Ici, nous visons les spécificités intra-corpus via l'extraction de MWE discriminants des sous-parties de corpus. Nous nous intéressons aux composants de liens thématiques latents qui lient les documents. Nous avons besoin d'une certaine flexibilité pour la construction des MWE : les skip-grams au moins. Les méthodes par apprentissage supervisé doivent être évitées. Les déterminants de l'extraction doivent être limités et internes au corpus, par exemple, aucun pré-conçu sur la forme grammaticale des MWE ou les éventuelles imprécisions du marqueur de PoS tagger ne doivent pas déterminer les extractions. Les résultats présentés par l'algorithme *ANA* (Enguehard et Pantera, 1995) était le plus précis sur le corpus de test, l'approche type « apprentissage automatique » semblait intéressante et prometteuse, notamment pour la grande flexibilité de longueur des skip-grams. La terminologie extraite était faiblement ambiguë car très orienté MWE.

3.1 Description de la proposition : ANA+

Caractéristiques. Nous proposons *ANA+* un algorithme probabiliste génératif de terminologie basé sur ANA (apprentissage non-supervisé). Cet algorithme extrait un multiensemble (sac) de skip-grams d'ordres (longueurs) variables et adaptés sans a priori, donc capable de maintenir au mieux l'information contenue dans les séquences de mots. *ANA+* fonctionne en Français et en Anglais. *Haruspex* étant modulaire, les développements proposés ici sont remplaçables par un autre algorithme d'extraction de MWE. Ceci permet d'adapter la proposition à une autre langue par exemple.

Définitions. Dans *ANA+*, nous nommons `candidat` tout skip-gram extrait avant la fin des itérations de l'algorithme (avant l'activité A25 sur la figure 2.8). Nous définissons des `fenêtres` qui sont des aires de recherche de collocations, construites autour des `candidats` (comme sur la figure 1.26). La taille de ces fenêtres se compte en lemmes, les lignes 3 et 4 de l'algorithme 4 décrivent cette construction. Les critères pour évaluer la validité d'une fenêtre dépendent du mécanisme de construction. Les `stopwords` tronquent les fenêtres, les `emptywords` ne sont pas comptés, ils sont transparents, les `linkwords` sont les mots de schéma : {of} en anglais, {de, en, au} et leurs dérivés (du, des, aux) en Français, ils facilitent la validation des fenêtres. Enfin *ANA+* a besoin d'un `bootstrap` de lemmes graines pour démarrer la première itération. Le contenu de ce bootstrap influence sensiblement les résultats. On peut déduire empiriquement que l'algorithme ne converge pas toujours exactement de la même manière, il peut être légèrement orienté par les choix de l'expert.

3.2 Mécanismes de construction

ANA+ itère sur 3 mécanismes de construction et 1 mécanisme de récession. Les 3 premiers mécanismes sont réalisés à partir des fenêtres, ligne 2 et 3 du pseudo-code 3. La phase d'essaimage est réalisée sur une première collecte de fenêtres. Les 2 autres mécanismes (Expansion et Expression) sont réalisés de manière concurrente à partir d'une seconde collecte de fenêtres (voir pseudo-code 4). L'annexe C présente des constructions instanciées. Cette annexe permet de comprendre par l'exemple les

Algorithme 3 ANA+ : mécanisme de construction

1: **pour** $cand$ in $candidates$ **faire**
2: $leftWin \leftarrow (cand_i, lem_{i-1}, lem_{i-2})$
3: $rightWin \leftarrow (cand_i, lem_{i+1}, lem_{i+2})$
4: $win \leftarrow win \cup \{leftWin, rightWin\}$
5: **fin pour**
6: $validWin \leftarrow$ filter(win)
7: $newCands \leftarrow$ count_cases($validWin$)
8: $candidates \leftarrow candidates \cup newCands$

mécanismes à l'œuvre dans *ANA+*.

3.2.1 Essaimage

La phase d'essaimage augmente l'entropie du système. Elle fait l'objet d'une collecte de fenêtres dédiées. Les fenêtres valides répondent aux contraintes suivantes :

— Elles contiennent un `candidat`.
— Elles contiennent éventuellement un `linkword`.
— Elles sont tronquées par un autre `candidat`, un `stopword` ou un `emptyword`.

Les occurrences des lemmes de toutes les fenêtres valides sont comptées quel que soit le `candidat` à l'origine de la fenêtre. Trois paramètres influent la validation du lemme : sa position par rapport au `candidat` (1re ou 2e position) ; l'utilisation systématique d'un même `linkword` (ou son absence) ; l'association avec un même `candidat`. Chaque paramètre est indépendant des autres. Une combinaison linéaire donne un score au lemme, au-delà d'un seuil le lemme devient un `candidat`. Une même fenêtre peut être utilisée plusieurs fois (pour les différents lemmes qu'elle contient).

$$\text{Essaimage}(l_i) = \left(\frac{\alpha}{S}\frac{F_v(l_i)}{C(l_i)} + \frac{\beta}{S}\frac{F_v(l_i)}{L(l_i)}\right) \times \frac{F_v(l_i)}{P(l_i)} \tag{2.4}$$

avec

— $F_v(l_i)$ le nombre de fenêtres valides d'un lemme
— $C(l_i))$ le nombre de `candidats` différents
— $L(l_i)$ le nombre de types de `linkwords` différents, l'absence de `linkwords` est considérée comme un type
— $P(l_i)$ le cumul des positions occupées dans les fenêtres : 1 pour la position la plus proche du `candidat`, 2 pour la seconde position.
— $S = \alpha + \beta$
— α et β méta-paramètres à fixer par l'utilisateur. Des valeurs indicatives sont $\alpha = 10$, $\beta = 3$. β indique la préférence pour utiliser toujours le même type de linkword ; α indique la préférence pour utiliser toujours le même `candidat` initial.

Le résultat du calcul 2.4 est compris entre $1/2$ et $F_v(l_i)$. Le seuil est fixé en fonction de la longueur du corpus. Pour les corpus courts (environ 600.10^3 mots) on valide le lemme l_i lorsque $\text{Essaimage}(l_i) > 3$ environ ; pour les corpus longs (plus de 3.10^6 de mots) on valide le lemme l_i lorsque $\text{Essaimage}(l_i) > 15$ environ. Il serait possible d'automatiser ce seuil (seuil $\approx \frac{mots}{2.10^3}$).

Cette gestion du seuil évite la rigidité des filtre initialement proposés par Enguehard et Pantera (1995) : chaque condition (nombre de `candidats` différents, nombre de *lien-clé* différents) indépendamment. La position dans la fenetre est un nouveau paramètre. Cette version augmente le rappel, sans diminuer la précision.

3.2.2 Développement

Après l'essaimage et après une nouvelle collecte de fenêtres autour des nouveaux `candidats`, la phase de développement augmente la longueur des `candidats` existants en fusionnant des nouveaux composants. De nouveaux `candidats` composés de plusieurs mots (MWE) en résultent.

Les expansions sont composées d'un même `candidat` suivi ou précédé par un même lemme systématiquement. Un lemme peut être ignoré à l'intérieur de la fenêtre (hypothèse skip-gram). Le nombre de motifs [`candidat` (lemme) - lemme] est compté. Un seuil d'occurrences valide les expansions. Ce seuil est d'environ 4 pour les petits corpus.

Le mécanisme d'expression est concurrent avec le mécanisme d'expansion. Il fonctionne de la même manière mais les fenêtres doivent contenir 2 `candidats` pour être valides. C'est à dire que le motif recherché est [`candidat- candidat`]. Le nombre de motifs est compté. Un seuil (environ 3 pour les petits corpus) valide les expressions. Les motifs peuvent contenir un lemme ignoré [`candidat-` lemme - `candidat`] à condition que l'inclusion du lemme ne puisse pas former une expansion.

Gestion des priorités. 2 cas se présentent, illustrés par des exemples en annexe C.

— En cas d'expansion incluse dans une expression ou une autre expansion, la priorité est donnée à l'expansion incluse. Ce comportement suppose qu'il faut construire les expansions élément par élément en évitant les skip-grams si possible.
— Si les expansions sont intriquées : incluses l'une dans l'autre, alors le comportement par défaut donne la priorité à la moins occurrente avec un comportement sans skip-gram (*non-greedy*). Ce dernier comportement suppose que l'expansion la plus occurrente pourra toujours se former, même après que certaines occurrences aient été consommées par l'expansion plus rare.

3.2.3 Récession

Cette étape est réalisée après chaque itération d'expression/expansion. Les candidats nouvellement construits « consomment » leurs composants. Ces derniers ne sont plus comptés individuellement. Les candidats occurrant insuffisamment (sous le seuil fixé) sont désactivés et retournent à leur état antérieur (un candidat plus simple ou un lemme). La désactivation limite les récessions en cascade d'un candidat très composé en lemmes simples. La désactivation permet de reformer les candidats rapidement lorsque le nombre de candidats désactivés dépasse le seuil de récession. Par exemple, 15 des 17 occurrences du candidat « *knowledge database* » sont précédées de « *failure* », alors le candidat « *failure knowledge database* » est construit et « *knowledge database* » est

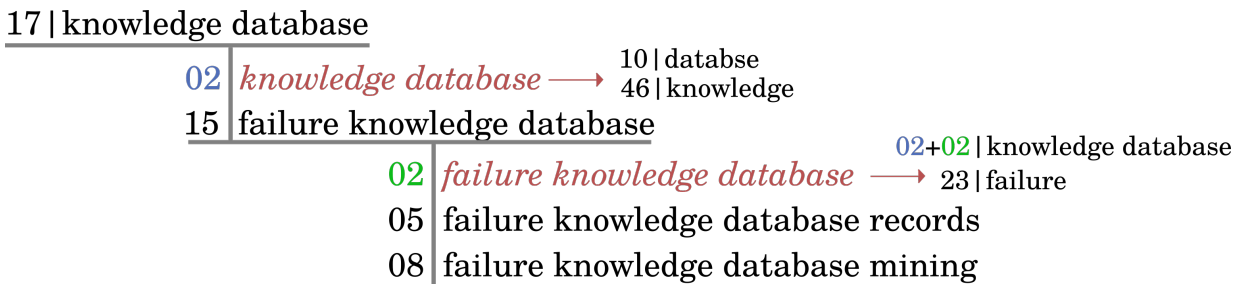
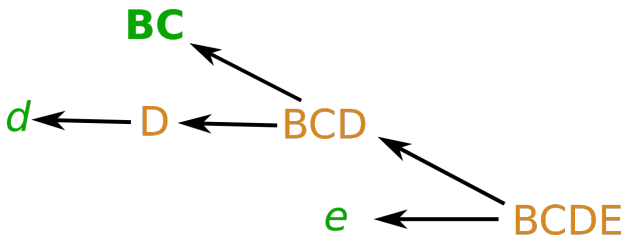


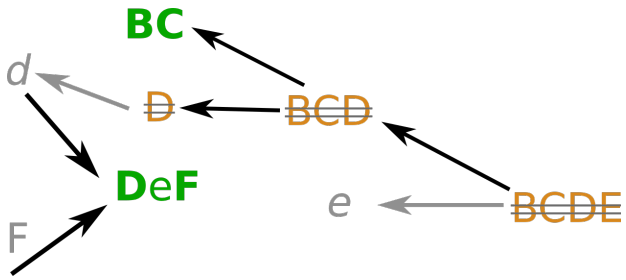
FIGURE 2.9 – Cas classique de récession vers un `candidat` désactivé, puis réactivé, au lieu d'un retour à des lemmes simples.

désactivé, car il n'occurre que 2 fois (en supposant que le seuil est de 3). « *knowledge database* » revient alors son état précédent « *database* » et « *knowledge* ». Mais une prochaine récession vers « *knowledge database* » pourrait réactiver le `candidat`. La figure 2.9 montre ce fonctionnement.

Tout développement d'un `candidat` (activé ou désactivé) détruit l'arbre des occurrences désactivées qui dépendent de lui (voir exemples en annexe C). Les arbres en figure 2.10 illustrent ce propos.



(a) Récession de BCDE jusqu'à des formes stables : le `candidat` BC et les lemmes d et e.



(b) Développement d'une branche (F → DeF, consommant d, et e en skip-gram), destruction de l'arbre en amont

FIGURE 2.10 – Arbre de récession d'une occurrence de `candidat` (BCDE) et destruction : en capitale les `candidats`, en minuscule les formes de simple lemme, en vert les formes stables, en orange les `candidats` désactivés, en gris les occurrences consommées.

3.3 Organisation et produits d'*ANA+*

3.3.1 Amorce, itérations et conditions d'arrêt

Itérations. *ANA+* itère plusieurs fois sur le mécanisme d'essaimage avant la phase d'expansions et expressions. Les phases d'essaimage, de développement (lignes 3 et 5 du pseudo-code 4) sont décrites par le pseudo-code 3).

Algorithme 4 ANA+

```
1: tant que newCands faire
2:     tant que newCands faire
3:         candidats ← essaimage(candidats)
4:     fin tant que
5:     candidats ← développement(candidats)
6:     candidats ← recession(candidats)
7: fin tant que
```

Amorce. On remarque que *ANA+* (pseudo-code 4) construit des `candidats` à partir de `candidats` existants. Il faut donc un `bootstrap` de candidats initiaux pour amorcer la première itération. Ce bootstrap est construit empiriquement à partir des noms propres occurrant davantage que le seuil de récession, des mots définissant les topics de la phase de co-clustering (section 2.2) et des mots ajoutés par l'expert. La construction de ce bootstrap est délicate, car elle peut influer sur les résultats finaux de *ANA+*. En effet l'algorithme ne converge pas toujours exactement sur les mêmes `candidats` lorsque les `candidats` de l'amorce sont modifiés.

Condition d'arrêt. *ANA+* s'arrête lorsque aucun nouveau `candidat` n'est formé, comme décrit par la condition « tant que des candidats nouveaux sont trouvés fait : » (ligne 1 du pseudo-code 4). En réalité on observe régulièrement des oscillations

entre la création de nouveaux candidats et leur destruction par la récession. Ce cycle peut être complexe et se répéter en plusieurs itérations. La condition d'arrêt pratique est donc donnée par un seuil minimal de variation du nombre de candidats sur plusieurs itérations.

3.3.2 Modification par rapport à ANA

6 modifications majeures ont été apportées à l'algorithme initial ANA. Nous les listons ci-dessous dans l'ordre d'importance.

1. La construction des fenêtres et les conditions de validation ont été modifiées (voir travaux de thèse de Enguehard (1993)). La couverture des `candidats` est améliorée, le rappel augmente sans diminuer la précision.
2. La priorité est donnée aux développements ou à l'essaimage de potentiels candidats imbriqué dans un skip-gram
3. Des itérations imbriquées d'essaimage permettent une meilleure couverture du corpus
4. La génération du `bootstrap` est assistée
5. L'algorithme peut passer à l'échelle grâce à l'indexation des lemmes (d'un temps d'exécution quadratique à un temps linéaire) et une phase de calcul parallèle asynchrone.
6. L'utilisation de lemmes au lieu des mots accélère le processus et augmente le rappel. ANA fonctionnait initialement avec l'égalité souple (Enguehard, 2001) basée sur les travaux de Wagner et Fischer (1974) et la distance de Levenshtein (Levenshtein, 1966).

3.3.3 Produits de l'extraction

Chaque *page* est maintenant décrite par un multiensemble de skip-grams. La DF, abscisse de la figure 2.11, est le nombre de

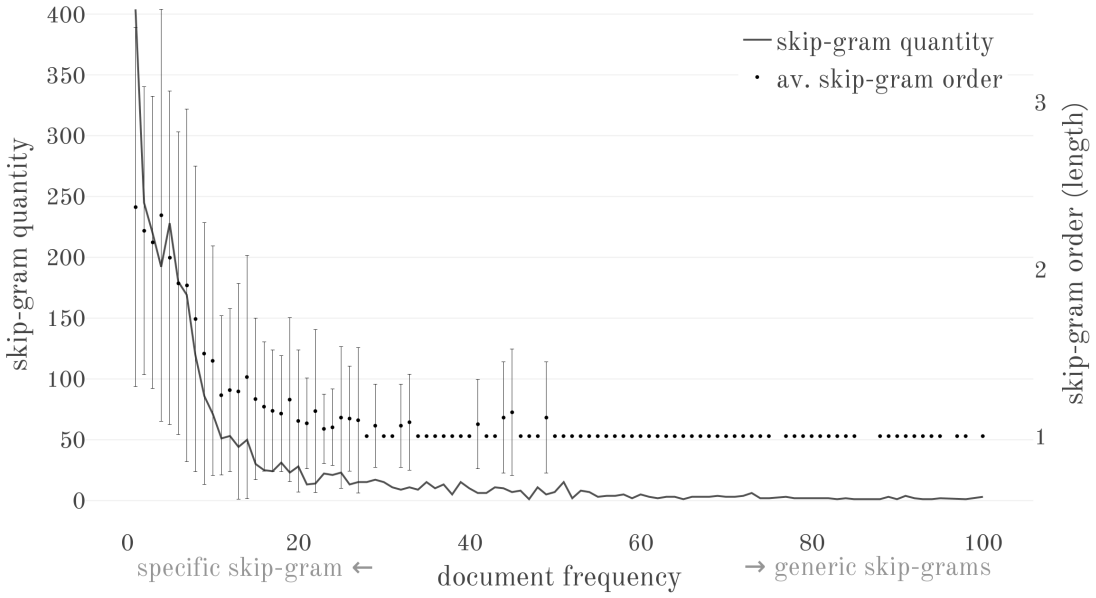


FIGURE 2.11 – Distribution et longueurs (ordre) des skip-grams extraits. Données obtenues sur un corpus de 100 *pages*, résultats similaire avec d'autres corpus.

pages où (au moins une occurrence) du skip-gram t_i est observée. On peut alors définir la quantité Q de skip-grams pour une DF donnée. Le nombre total de skip-grams est noté m (voir équation 2.5).

$$\begin{aligned} DF^{t_i} &= |\{p_j : t_i \in p_j\}| \forall j \in [0, n] \\ Q(DF) &= \sum_{i=1}^{m} DF^{t_i} : DF^{t_i} = DF \end{aligned} \tag{2.5}$$

On remarque que la plupart des skip-grams occurrent dans peu de documents, ils sont spécifiques. La table 2.2 en montre quelques exemples. Ces skip-grams tendent à être plus complexes (contenir plus de mots) que les expressions plus génériques. Davantage de mesures sur les sorties de *ANA+* sont présentés en section 6. Mais nous observons déjà que le type de résultats et la distribution de ceux-ci sont intéressants.

4 Post-traitement des expressions

Les mots clés extraits sont classés dans des catégories si possible et un rang leur est attribué.

Forme la plus occurrente	nb. d'occ.	DF
Chimie de coordination	22	8
Thèse de troisième cycle	16	7
Bronzes de vanadium	26	5
Four solaire d'Odeillo	4	3
Microscopie électronique à transmission à haute résolution	3	2

TABLE 2.2 – Exemple d'expressions extraites par *ANA+*, sur un corpus d'entretiens en chimie du solide.

4.1 Classification des expressions

La classification des skip-grams utilise des bases de connaissances extérieures (Milne et Witten, 2008; Richman et Schone, 2008). S'il existe un titre d'article Wikipédia correspondant au skip-gram, l'idée est que les portails et catégories de cet article puissent servir de classes. Sans restreindre les résultats à des données extérieures existantes, ces classes offrent des informations additionnelles pour les requêtes.

4.1.1 Classes de Wikipédia Français / Anglais

Ce sont les portails et catégories de Wikipédia qui assurent la classification si possible des skip-grams. Or si Wikipédia offre les mêmes moyens de structuration en portail/catégories dans toutes les langues, les pratiques ne sont pas les mêmes en anglais et en français, langues de *Haruspex*. À l'inverse des portails les catégories sont organisées hiérarchiquement.

Wikipédia français est surtout structuré en portails. Les catégories apportent des précisions ou des groupements précis voire surprenants. Les portails français sont directement assimilables à des classes au sens de thème. Wikipédia anglais est davantage structuré en catégories et très peu en portail. L'annexe B montre quelques exemples qui illustrent ces différences et analyse quelques traits de la structuration de Wikipédia.

Les portails de Wikipédia français fonctionnent directement par rapport à nos attentes. Ils permettent un recouvrement correct : un portail concerne plusieurs articles. Nous choisissons donc de les récupérer [2] pour construire des classes de nos instances (expression extraites par *ANA+*).

4.1.2 Indicateur de confiance

Avec les portails Wikipédia. Un indicateur de confiance (*confidence index, CI*) est calculé sur les portails Wikipédia associés au skip-gram. Il mesure l'ambiguïté ou le hors-sujet possible de l'article Wikipédia correspondant au skip-gram extrait. On considère que les skip-gram sont étiquetés (*tagged*) par les portails des articles Wikipédia correspondants.

$$CI(t_i) = \frac{1 + \omega(t_i)/\Omega}{1 + |\overline{G(t_i)}|}$$

avec :

- $G(t_i)$ l'ensemble des portails étiquetant le skip-gram t_i
- $\omega(t_i)$ le nombre cumulé de skip-gram étiquetés avec les portails issus de $G(t_i)$
- Ω la valeur maximum de ω, correspond donc au ω du skip-gram étiqueté avec les portails les plus « populaires » du corpus
- $\overline{G(t_i)} = G(t_i) - G(t_1) \cup G(t_2)... \cup G(t_{i-1}) \cup G(t_{i+1})... \cup G(t_n)$ l'ensemble des portails étiquetant t_i mais avec aucun autre skip-gram.

Cette mesure fonctionne mieux que les indicateurs classiques (index de Jaccard, de Tversky) pour de petits sous-ensembles de multi-sets. En effet, ces derniers sont conçus pour mesurer des écarts entre ensembles similaires et ne sont pas assez sensibles pour discriminer de petites variations parmi des sous-ensembles correspondant sensiblement à l'ensemble mère. Le numérateur croît lorsque les portails de l'article sont fameux (présent dans de nombreux articles) il est compris entre $[1, 2]$. Le dénominateur compte les portails inconnus (isolés), sans doute hors-sujet ou marginaux. Les skip-gram marqués avec un portail isolé ont un indicateur $CI \leq 1$ les autres ont un indicateur $CI \in [1, 2]$. Une variante moins stricte de cet indicateur autorise 1 portail isolé par skip-gram.

En cas de *page d'homonymie* (plusieurs titres d'article Wikipédia correspondent à cette expression), la décision est prise après avoir épuisé tous les skip-gram correspondant à des articles sans ambiguïté. La mesure CI est appliquée pour tous les articles ambigus (qui ont le même titre). L'article avec le score CI le plus élevé est conservé. L'ensemble $\overline{G(t_i)}$ n'est pas mis à jour avec ces nouveaux portails. L'algorithme est fortement conservateur.

Autres classifications. Pour les noms propres en Français, un mécanisme similaire opère des requêtes SPARQL (langage de requête pour serveurs de triplets RDF) vers le serveur de triplet du dépositaire d'autorité *dataBNF*. Les requêtes visent les classifications *Dewey* (voir section 2.2) des instances extraites par *ANA+*.

2. via des requêtes `http` vers l'API de Wikipédia

En anglais, les catégories de Wikipédia sont utilisées. Ce fonctionnement serait à perfectionner. Souvent les catégories des articles sont éclatées et les recouvrements sont rares parmi les données issues de *ANA+*. L'enjeu est alors de remonter l'arbre des catégories jusqu'à trouver des catégories plus génériques, communes à plusieurs articles repérés, en évitant les catégories trop génériques. L'algorithme ici est très naïf pour la complexité de la tâche : il consiste à remonter l'arbre des catégories en agrégeant ces nouvelles catégories jusqu'à ce que le nombre d'article concernés par ces catégories soit supérieur à 1000. Voir exemple et graphique B.1 en annexe.

Enfin en annexe B, un exemple d'analyse des portails sur un corpus permet de comprendre l'intérêt de ce fonctionnement pour regrouper les skip-grams.

4.2 Classement (*ranking*) des expressions

Le classement des expressions (au sens de *ranking*, c'est-à-dire attribuer un score aux expressions) facilite la modération par l'expert et limite ainsi les faux-positifs (améliore la précision). Ce score est calculé par une combinaison linéaire de 5 mesures : MEDh, PoStag, WRlog, TFIDF et CI. Chacune des mesures peut également être utilisée seule pour faciliter la modération.

$$Rank = \frac{PoStag \cdot WRlog \cdot TFIDF \cdot CI}{MEDh}$$

4.2.1 MEDh

MEDh est une proposition pour améliorer la mesure MED (Bu *et al.*, 2010) de cohésion des termes d'une expression. Le fonctionnement de base de MED est expliqué en section 4.1.3. Ici l'objectif est de distinguer les expressions qui cristallisent des associations de mots. Nous mesurons alors la dépendance de mots à l'expression. Lorsqu'un composant d'une MWE (un lemme) n'occurre pas en dehors des *pages* où la MWE occurre, alors on considère qu'il est dépendant du champ sémantique de l'expression. On remarquera qu'il n'est pas nécessaire que ses occurrences forment systématiquement l'expression. L'intérêt de cette mesure est de revaloriser les expressions spécifiques, parfois tronquées (par simplification) ou souffrant de variation. Cet avantage est flagrant avec les noms propres où l'on évoque souvent un « prénom nom » puis uniquement « prénom » ou « nom ». Mais très rarement le nom ou le prénom seuls n'apparaîtront dans une *page* où l'expression n'a pas été formée. L'expression « prénom nom » cristallise cette dépendance. C'est aussi une hypothèse valable avec les noms communs dans les corpus restreints. Par exemple dans un corpus de chimie du solide « valence » est très dépendant de « électron ». La figure 2.12 montre 2 cas, pour

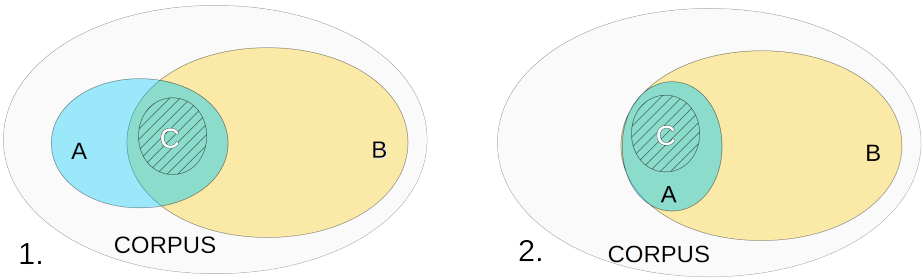


FIGURE 2.12 – A (ex : « information ») et B (« mutuelle ») sont les composants de C (ex : « information mutuelle »). MED trouve la même valeur de cohérence pour ces 2 distributions ; MEDh favorise celle de droite

chacun des cas, nous observons parfois C dans les *pages* où A et B occurrent. A et B sont les lemmes composants C. Dans le cas de gauche les 2 lemmes indépendants A et B sont parfois réunis dans les mêmes *pages* et dans ce cas ils forment parfois l'expression C. Par exemple A = « microscope » et B = « électronique ». Dans le cas de droite, A n'occurre jamais dans les *pages* où B n'occurre pas. Par exemple A= « valence » et B = « électron ». MED attribue une valeur de cohésion égale dans ces 2 cas, MEDh favorise le second cas.

En généralisant aux MWE d'ordres supérieurs, Bu *et al.* (2010) introduit les notions de contexte et de sémantique. En faisant l'hypothèse que la probabilité d'apparition est proportionnelle à la cardinalité, la sémantique notée $\mu(t_i)$ correspond à la distribution des occurrences de t_i, c'est-à-dire aux *pages* où le skip-gram t_i occurre (soit C en figure 2.12). Le contexte noté $\phi(t_i)$ correspond à la distribution jointe des occurrences des composantes de t_i (soit $P(A \cap B)$ en figure 2.12). On note $(dl_1, dl_2, ..., dl_k)$ les distributions (indépendantes) des lemmes $(l_1, l_2, ..., l_k)$ composant le skip-gram t_i (d'ordre k).

$$MEDh(t_i) = MED(t_i) \times (1 + \min_{p=1..k}(KL(\mu(t_i)||dl_p)) \tag{2.6}$$

En approximant la complexité de Kolmogorov par le codage de shanon-fano. on écrit MED :
$MED(t_i) = (\phi(t_i)|\mu(t_i)) = \log|\mu(t_i)| - \log|\phi(t_i)| = \log(DF(l_1, l_2, ..., l_k)) - \log(DF(t_i))$

Nous avons donc $MEDh(t_i) \geq MED(t_i)$. MEDh varie comme MED, la valeur idéale étant 1, cette valeur augmente lorsque la cohésion des composants de l'expression diminue. La divergence de KL [3] est choisie, car elle est fortement discriminante pour les valeurs nulles d'une distribution et faiblement pour les variations de valeurs autres. Ce n'est pas le cas d'autres mesures de déviations (entropie de Shannon par exemple). Ainsi la représentation d'une dimension influe davantage que la norme du vecteur. Par exemple le vecteur [1, 2, 0, 1] sera proche de [2, 4, 0, 4] mais distant de [1, 2, 1, 1].

4.2.2 Marques linguistiques

Les marqueurs linguistiques ou PoS tagger consistent à repérer les valeurs grammaticales des composants de l'expression. De nombreux algorithmes d'extraction de terminologie utilisent ces valeurs comme filtres en amont. Nous les utilisons comme critère de notation : les expressions hors filtre sont moins bien notées mais ne sont pas exclues. Les valeurs sont arbitraires. Trois cas sont distingués :

— L'expression contient un verbe ou un mot suspect ({et, pour, avec, ... }) : $PoS = 0.5$
— L'expression contient au moins 1 nom et exclusivement des adjectifs ou noms. La MWE correspond au motif (Nom|Adjectif)*Nom(No
$PoS = 2$
— Dans tous les autres cas : $PoS = 1$

4.2.3 Ratio d'étrangeté

L'étrangeté est présentée en section 4.1.3 sous le nom deWRlog (Cram et Daille, 2016). Elle mesure la surreprésentation d'un mot dans le corpus par rapport à la moyenne dans la langue. Cette mesure est simplement adaptée au MWE (t_i) en conservant le lemme (l_i) le plus étrange de l'expression. Par exemple dans le skip-gram « supraconducteur à haute température », le WR de « supraconducteur » prévaudra. En effet les autres composants (même communs) ne dévalorisent pas la rareté de l'expression. Une mesure plus fine où chaque composant augmente la rareté pourrait être envisagée.

$$WR(t_i) = \max WRlog(l_i) : l_i \in t_i$$

4.2.4 Autres indicateurs déjà présentés

TF-IDF Cette mesure est également présentée en section 4.1.3. Par la suite l'importance des expressions est proche du calcul de TF-IDF, c'est pourquoi cet indicateur est utilisé ici. En réalité, il sert à détecter les MWE faux-positifs avec un gros potentiel de nuisance.

Indicateur de confiance du groupe. Le calcul de cet indicateur est précédemment expliqué lors de la capture des portails Wikipédia. La valeur de cet indicateur (CI) varie entre 0 et 2. Si aucune page Wikipédia ne correspond au skip-gram, alors $CI = 1$, qui est une valeur neutre.

4.3 Fusion

Les expressions peuvent fusionner avec les parents. Une suggestion automatique des fusions est proposée. les fusions doivent être validées par l'expert. Ces suggestions facilitent la modération : augmentent la précision en réduisant les faux-négatifs sans diminuer le rappel. Ces fusions concernent différents types d'expressions, les critères sont les suivants :

— les expressions ayant un score (voir section 4.2) faible. Par exemple : « hydrogène pour véhicule électrique » fusionne avec « hydrogène » et « véhicule électrique ».
— les expressions dont la racine (*stem*) du lemme est identique. Par exemple « supraconductivité » et « supraconducteur ».
— les expressions ayant une mauvaise cohérence ($MEDh$). Par exemple « première année de thèse » fusionne avec « thèse ».

4.4 Modération

Les résultats de *ANA+* (section 3) sont modérés par l'expert. Cette modération est assistée par les indicateurs produits en section 4.2. Les éléments présentés ne sont pas censurés. Il y a donc du bruit. Les indicateurs permettent une décision rapide et les propositions de fusions améliorent les résultats. À cette étape il est possible de :

— Refuser une fusion (acceptées par défaut)
— Supprimer une expression
— Modifier les classes de l'expression (portails Wikipédia).

Les modifications de l'utilisateur sont enregistrées. En prévision d'éventuelles modifications du corpus (qui impliquerait de relancer l'extraction), les skip-grams validés sont inclus dans le `bootstrap` (section 3) et les éléments supprimés seront automatiquement censurés des résultats. La valeur de l'indicateur de confiance CI est recalculée après modération (modification des groupes).

3. nulle pour des distributions identiques, vaut 1 pour des distributions fortement divergentes

À partir de cette étape, on considère la notation suivante. Pour chaque *page* p_j, on définit le multiensemble $\{\{T_{p_j}\}\} = \{\{t_i : t_i \in p_j\}\}, \forall i \in [0, m]$ de skip-grams t_i situés dans p_j parmi les m skip-grams modérés.
On définit $O^{t_i}_{p_j} = |\{\{t_i : t_i \in p_j\}\}|$ le nombre d'occurrence du skip-gram t_i dans la *page* p_j.

5 Création des liens entre pages

Cette partie vise à calculer des proximités entre les documents sur la base des MWE précédemment extraites.

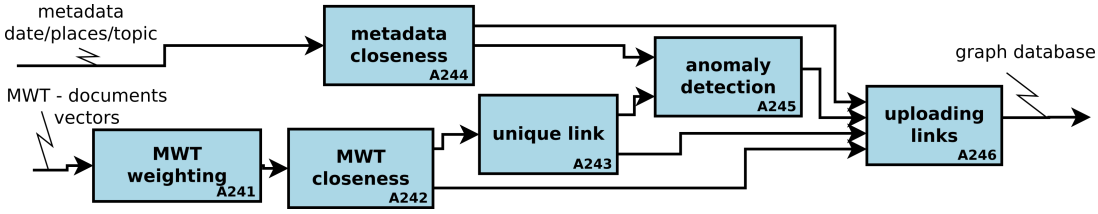


FIGURE 2.13 – Description SADT des sous-activités de l'activité 4 : lier les *pages*

5.1 Approche classiques et problèmes

Le calcul de proximités entre documents permet de former des graphes. La section 4.1.4 traite de certaines mesures classiques.

Les calculs de proximités classiques proposent une mesure globale entre 2 documents. Chaque *page* est représentée par un vecteur des poids de termes occurrant dans la *page*. Les mesures sont donc des opérations entre les vecteurs : cosinus, index de jaccard, divergence de KL, etc. Ce type de distances présente 2 inconvénients majeurs dans notre cas, nous les présentons ci-après (section 5.1.1).

5.1.1 Problèmes

Problème des échelles. Les mesures proposées ne donnent qu'une seule distance entre 2 vecteurs (*pages* dans notre cas), sans plus de précision. Une approche moins globale serait adaptée à la faible taille des corpus traités (jamais plus de 1000 documents, souvent autour de 100). Pour résoudre le verrou 6, et résoudre ce problème des mesures classiques, nous proposons de décomposer cette mesure en sous-ensembles plus précis : distances entre documents sur une combinaison de thématiques ou sur une notion précise.

Problème de distance entre mots. Selon les techniques classiques précédemment évoquées, la distance entre les *pages* p_1 et p_2, respectivement représentées par les ensembles {engrenage, modélisation, monument} et {pignon, ordinateur, patrimoine} n'est pas celle que l'on perçoit qualitativement, intuitivement. En effet, même s'ils sont syntaxiquement différents, les mots « fronton » et « pignon » sont proches sémantiquement (idem pour les autres paires, à vous de jouer), mais cela n'est pas toujours vrai (« pignon de pin »).

Conformément au verrou 5, nous établissons que :

— L'hypothèse que tous les mots sont fondamentalement différents n'est jamais vérifiée : un calcul de distance purement lexical semble non-représentatif.
— Les notions de synonymie (mots ayant le même sens) et de champs lexical (mots appartenant au même domaine de sens, liés par un hyperonyme) dépendent des contextes d'usage : il n'est pas possible d'établir une distance entre les mots *a priori*.

5.1.2 Résolution des problèmes pré-cités

La section 4.2.4 propose une approche **quantitative** du problème. Des techniques de *word-embedding* construisent des vecteurs de lemmes : chaque lemme du corpus est représenté par un vecteur d'autres lemmes. Ces vecteurs sont associés à une probabilité de collocation avec le mot. Ainsi, avec certaines limites, cette technique quantifie la distance entre les mots d'un corpus. Pour obtenir des proximités entre mots, 2 options ont été envisagées :

A Les mots similaires apparaissent dans des contextes similaires. Par exemple : avec les phrases « Le fruit du poirier est comestible » et « le fruit du pommier est comestible » nous obtenons des contextes identiques pour « poirier » et « pommier », leur distance est nulle.
B Les mots similaires apparaissent régulièrement ensemble. Avec l'exemple précédent, les mots « fruit » et « comestible » apparaissent ensemble.

Nous retiendrons l'hypothèse A. De plus, les données dont nous disposons ne sont pas des mots simples mais des skip-grams déjà constitués. La constitution de ces skip-grams correspond justement à une fréquence élevée de collocation.

Pour éviter de construire des résultats que nous connaissons déjà (les composants du skip-gram co-occurrent), nous considérons le skip-gram extrait par *ANA+* comme un terme unique. Par ailleurs, la création de vecteurs représentatifs pour les termes que *ANA+* n'a pas extrait ne nous intéresse pas.

Nous utilisons la méthode *Word2vec*. Les données pour l'apprentissage dans notre cas seront les couples {skip-gram, contexte}. Le contexte est défini par des fenêtres standards de 5 (lemmes) maximum, tronquées par les `stopwords` (voir section 3.1). Les données d'apprentissage étant en volume très limité, et ayant déjà filtré les `emptywords` et `stopwords`, nous ne retiendrons pas la technique d'échantillonnage proposée par Mikolov *et al.* (2013a). Nous obtenons un vecteur représentant chaque skip-gram.

5.1.3 Valeurs d'occurrences d'un skip-gram dans une *page*

Le principe de valeur d'occurrence. Les occurrences des skip-gram dans les *pages* sont redéfinis en *valeurs d'occurrence* (scalaires). On note $O_{p_j}^{t_i}$ la *valeur d'occurrence* d'un skip-gram t_i dans une *page* p_j. Cette valeur compte le nombre de fois que le t_i occurre et rajoute la projection P des autres termes sur cette composante.

$$O_{p_j}^{t_i} = |\{\{t_i : t_i \in p_j\}\}| + P \tag{2.7}$$

P est la composante du terme t_i dans les vecteurs représentant les autres skip-grams. Ainsi par exemple, dans un corpus de 6 mots (pour l'exemple), la table 2.3a montre que chaque occurrence du mot « pignon » comptera aussi pour les valeurs d'occurrences d'autres termes (en moindre proportion). Réciproquement, d'autres termes compteront aussi comme valeur d'occurrence de « pignon ». Toujours dans notre exemple de corpus à 6 lemmes différents, une *page* qui comporterait 5 occurrences de « engrenage »

	pignon	*ordinateur*	*patrimoine*	*engrenage*	*modélisation*	*monument*
pignon	1	0.02	0.12	0.84	0.31	0.11
ordinateur	0.02	1	0.02	0.13	0.61	0.01
patrimoine	0.12	0.02	1	0.09	0.23	0.89
engrenage	0.84	0.13	0.09	1	0.42	0.09
modélisation	0.31	0.61	0.23	0.42	1	0.24
monument	0.11	0.01	0.89	0.09	0.24	1

(a) Exemples de distances cosinus entre les vecteurs des différents mots

pignon brut	*pignon et similitude*	*page exemple*
1	1	4.58
0	0.02	1.71
0	0.12	3.47
0	0.84	5.4
0	0.31	3.4
0	0.11	3.14

(b) Représentations brute et avec similitudes d'une occurrence de « pignon », puis d'une *page*

TABLE 2.3 – Exemples de représentation de mots et *pages* dans un corpus de patrimoine industriel.
Note : ce corpus ne traite pas de pignon de bâtiment, les vecteurs traduisent cette désambiguïsation

et 3 occurrences de « patrimoine » et 1 occurrence de « ordinateur », sera représentée par le vecteur à droite du tableau 2.3b.

Algorithme. Pour éviter de manipuler des matrices très denses (faiblement compressibles), l'algorithme projette uniquement les dimension des mots suffisamment proches (proximité supérieure à un seuil). Le seuil est par exemple fixé à 0.5 Dans le cas des mots composés (MWE) : la composition ne compte pas comme occurrence de chacun des composants. On distingue alors les expressions dites « figées » de leurs composants (qui ont des contextes d'occurrences très différents). Par exemple dans un corpus de chimie du solide : « chimie de coordination » ne sera pas proche de « organisation ».

5.2 Proposition de création de liens

Dans notre cas de VSM, les nœuds sont des *pages* représentés par des vecteurs de skip-grams. Nous visons à construire un graphe de proximités (voir section 4.1.4) entre *pages*, la pondération des arcs est calculée à partir des skip-grams. Il s'agit de l'inverse du graphe de cooccurrence dans les documents (*term maps*, voir section 4.1.4).

Notre but ici n'est pas de créer des clusters de document mais d'établir des liens plus précis entre documents éventuellement issus de clusters différents. La création de cluster est établie par NMF (section 2.2).

Sept types de liens sont créés à cette étape, ils sont décris ci-après :

— : lien établi entre 2 documents sur la base d'un skip-gram commun
— *lien unique* : lien entre 2 documents, basé sur l'ensemble des skip-grams en commun.
— *lien négatif* : lien entre 2 documents où un skip-gram générique est absent.

Algorithme 5 Vecteurs des valeurs d'occurrences

Entrée: Seuil ; Vecteurs de collocations pour chaque MWE : $collocMWE_x$
Sortie: Vecteurs des valeurs d'occurrence pour chaque MWE : $vecMWE_x$

pour chaque $collocMWE_h$ **faire**
  pour chaque $collocMWE_i$ **faire**
    si $vecMWE_h[i] \neq 0$ **alors**
      $dist_{hi} \leftarrow \cos(collocMWE_h, collocMWE_i)$
      si $dist_{hi} \geq seuil$ **alors**
        $vecMWE_h[i] \leftarrow dist_{hi}$
        $vecMWE_i[h] \leftarrow dist_{hi}$
      fin si
    fin si
  fin pour
fin pour

— *lien-boucle* : boucle (*loop*) sur un document généré par un skip-gram occurrant seulement dans ce document.
— trois liens de proximité sur les métadonnées : espace, temps et topic.

5.2.1 Création des *liens-clés*

Un *lien-clé* par skip-gram occurrant dans une paire de *pages* est construit : $\forall t_i \in p_j, p_k \exists ! P^{t_i}_{p_j p_k}$

Nous étendons la notion d'occurrence d'un skip-gram à celle de « valeur d'occurrence » (section 5.1.3). Ainsi une expression similaire de l'expression cible peut compter occurrence. Un seuil γ de similarité est défini. On prend souvent $\gamma = 0.8$. Une valeur de γ plus faible amalgame un trop large champs lexical autour de chaque mot.

Nous supposons qu'un petit nombre de *pages* contenant de nombreuses occurrences d'un skip-gram partagent une caractéristique saillante et sont liées. Le calcul de la pondération du *lien-clé* est fonction de la distribution de l'expression au sein du corpus (noté W) et au sein de chaque paire de *pages* concernées (noté F) :

W : Au sein du corpus, une distribution entropique (Shannon, 1948) des skip-grams parmi les *pages* est favorisée.
F : Au sein de la paire de *pages* concernés, la quantité et l'équipartition des occurrences du skip-gram sont favorisées.

En W, une fonction à seuil inspirée de IDF(Salton *et al.*, 1973) permet de séparer les expressions génériques des expressions spécifiques à certaines parties du corpus (discriminantes). En F, nous utilisons une fonction logarithmique impliquant le nombre minimal d'occurrence de l'expression dans la paire. Cette fonction est inspirée de la TF. La mesure que nous proposons est donc

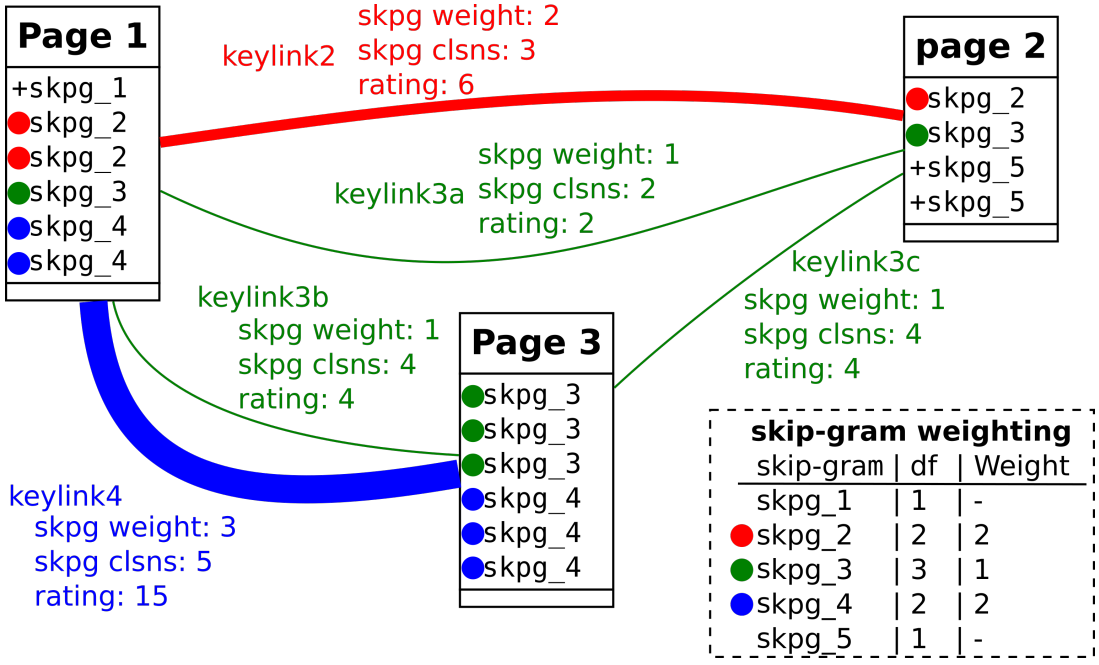


FIGURE 2.14 – Schéma simplifié de la création de *lien-clé*. *Weight* indique le poids d'un skip-gram. *Clsnss* indique le volume commun à 2 *pages* ; *rating* la pondération du lien, proportionnelle à l'épaisseur du trait.

inspirée de TF-IDF. Pour chaque t_i, occurrant dans une paire de *pages* p_j et p_k, le *lien-clé* est pondéré par :

$$P^{t_i}_{p_j p_k} = W_{t_i} \times F^{t_i}_{p_j ; p_k} \forall t_i \in p_j et p_k \tag{2.8}$$

Le poids du skip-gram : W_{t_i}. Il s'agit de la partie IDF de la pondération. Cette mesure vise l'entropie dans la distribution des skip-grams. Les skip-grams présents dans de nombreuses *pages* (génériques) sont peu nombreux (à droite du graphique en figure 2.11. Pourtant la plupart des *liens-clés* tels que définis par l'équation 2.8 sont issus de ces rares skip-grams génériques. Les barres sur le graphique de la figure 2.15 correspondent à la quantité de liens en fonction de la DF et de la quantité de skip-

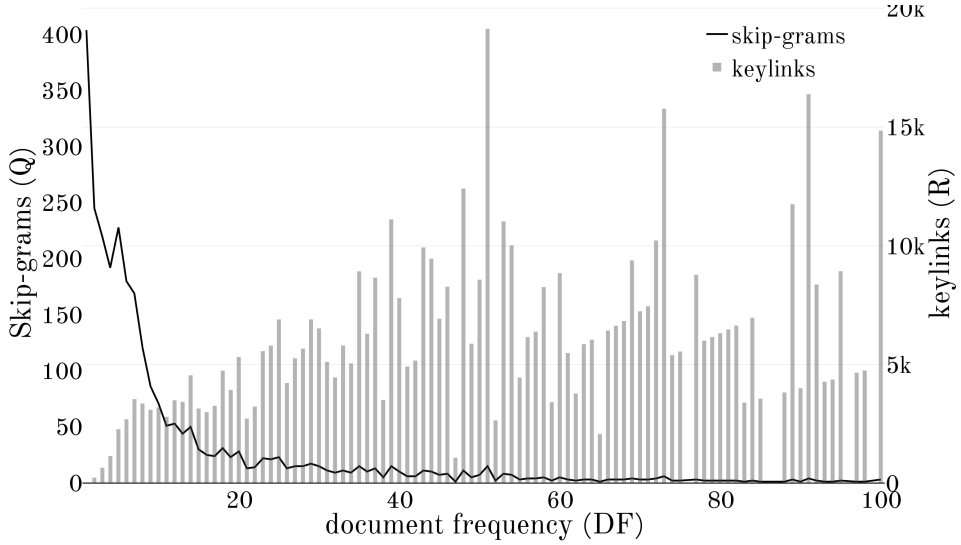
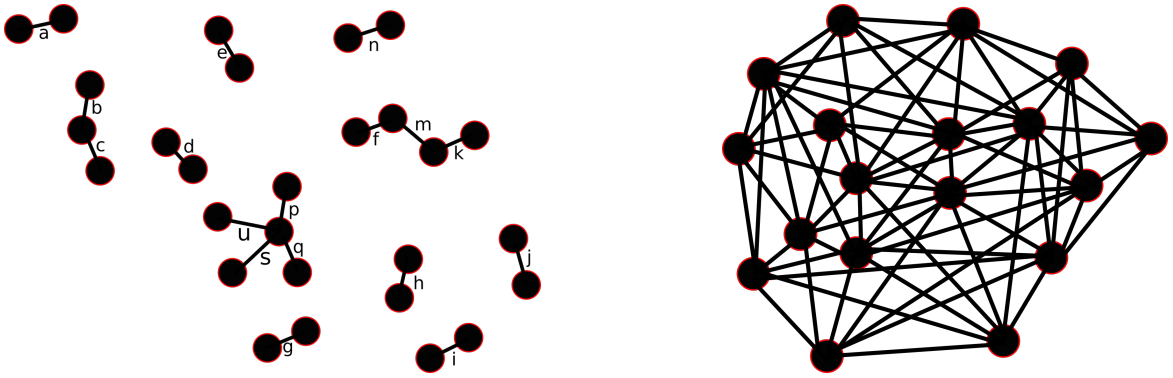


FIGURE 2.15 – Le nombre de *liens-clés* dépend du nombre de skip-grams et de DF. Pas de keylink pour $DF < 2$. Exemple issu d'un corpus de 100 *pages*. Résultats similaires pour tous les corpus.

grams. Cette quantité est une combinaison (écrite avec C) de 2 éléments parmi les *pages* concernées. $R(DF) = Q(DF) * C_2^{DF}$ avec $R(DF)$ la quantité de liens, $Q(DF)$ la quantité de skip-grams en fonction de la DF. Ce phénomène est une explosion combinatoire. Par exemple, 21 skip-grams occurrant chacun dans 2 documents produisent 21 liens (figure 2.16a), tandis que 1 skip-gram occurrant dans 21 documents produit 253 liens (figure 2.16b). Pour éviter les liens issus de skip-grams génériques,



(a) 18 skip-grams ∈ 2 *pages* chacun : 18 liens

(b) 1 skip-gram ∈ 18 *pages* : 153 liens

FIGURE 2.16 – Explosion combinatoire des liens issus de skip-grams génériques. En figure b, tous les liens ne sont pas représentés.

donc peu discriminants, nous proposons un sigmoïde pour la pondération des skip-grams en fonction de leur DF. Cette fonction à seuil est paramétrable (α et β) peut s'exprimer indépendamment du nombre total de document dans le corpus.

$$W_{t_i} = \frac{1}{1 + e^{\alpha(DF_{t_i} - \beta)}}$$

Les paramètres sont calculés en fonction de caractéristiques du corpus ou définis manuellement. α paramètre la pente de la transition, β paramètre le centre de la transition (en DF). L'idée est de saisir la zone de transition et d'éviter les grandes quantités d'arcs non discriminants (qui lient chaque *page* à toutes les *pages*). La solution consiste à trouver une valeur de DF limite. Les valeurs manuelles du cas précédent seraient $\alpha = 1$ et $\beta = 15$. Pour cela, les valeurs de la variation absolue du nombre de skip-grams sont lissées par un filtre linéaire (convolution gaussienne).

Lorsque le nombre de skip-gram varie faiblement alors le nombre de *liens-clés* est fortement influé par la combinaison. Ce seuil de variation minimal est arbitrairement fixé à 1% du nombre de skip-grams extraits.

Comme le montre les graphiques de la figure 2.17, les sigmoïdes proposées évitent le « bruit » (zone rouge) de la pondération IDF. en effet, IDF est peu discriminante pour les skip-grams ayant une DF intermédiaire. Par ailleurs, les sigmoïdes proposées

permettent une stabilisation des valeurs (plateau initial à gauche) pour les skip-grams discriminants (apparaissant dans peu de documents).

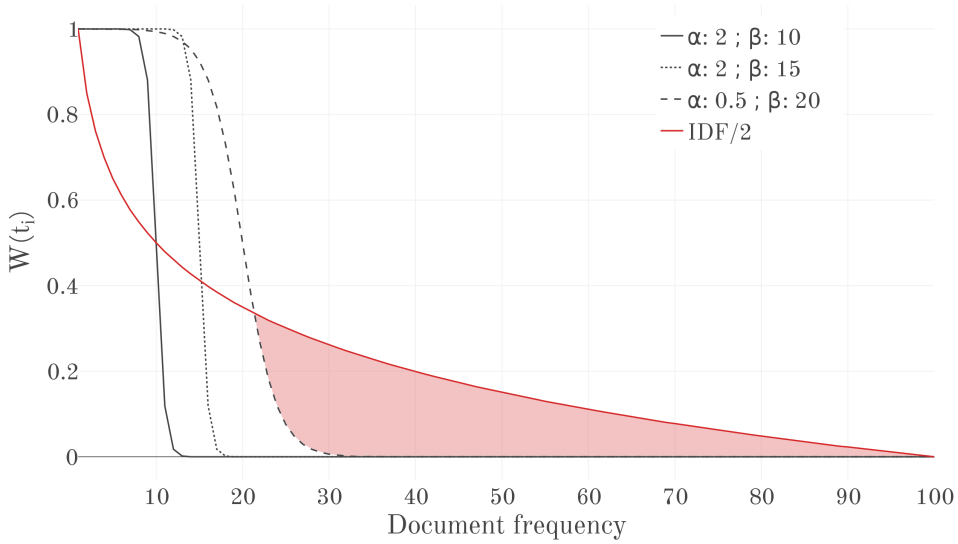


FIGURE 2.17 – Pondération ($W(t_i)$) des skip-grams en fonction de la DF. La zone rouge montre le bruit évité en choisissant la dernière sigmoïde plutôt que IDF

Les valeurs de α et β sont déterminantes sur les résultats, l'application sur *Nantes1900* en section 1 montre comment il est parfois nécessaire de modifier ces valeurs en fonction des résultats désirés. β détermine la taille maximale d'un cluster sur un sujet donné (une expression). Si $\beta >> \alpha$, ces valeurs déterminent faiblement (très indirectement) la taille des clusters résultants des liens uniques car les groupes présentent systématiquement un fort recouvrement de différents sujets (union de plusieurs expressions). Ce recouvrement, propre au contenus du corpus, est prépondérant dans la forme du graphe des liens uniques.

L'importance de la relation en *pages* : $F^{t_i}_{p_j;p_k}$. Cette autre composante de la pondération du *lien-clé* est en lien avec la TF. Il s'agit donc des occurrences du skip-gram t_i dans les *pages* p_j et p_k, ainsi que des occurrences pondérés des termes équivalents (voir section 5.1.3). Nous notons ces quantités respectivement $O^{t_i}_{p_j}$ et $O^{t_i}_{p_k}$.

La fonction TF somme les quantités : $TF = O^{t_i}_{p_j} + O^{t_i}_{p_k}$. Elle présente 2 inconvénients :

— ne pas différencier la distribution des occurrences dans la paire. Cela ne prend pas en compte $|O^{t_i}_{p_j} - O^{t_i}_{p_k}|$. Une *page* où un skip-gram occurre beaucoup attire à elle toutes les *pages* où ce skip-gram occurre.
— croître de manière régulière, quelle que soit la valeur. Ainsi la variation de pondération est la même entre 2 et 3 occurrences du terme qu'entre 25 et 26 occurrences.

Pour pallier à ces inconvénients, nous proposons de favoriser une équi-répartition des occurrences d'un skip-gram dans une paire de *pages* et de limiter les variations de pondération pour les occurrences très nombreuses. Pour cela nous utilisons la fonction log et le minimum d'occurrences partagées.

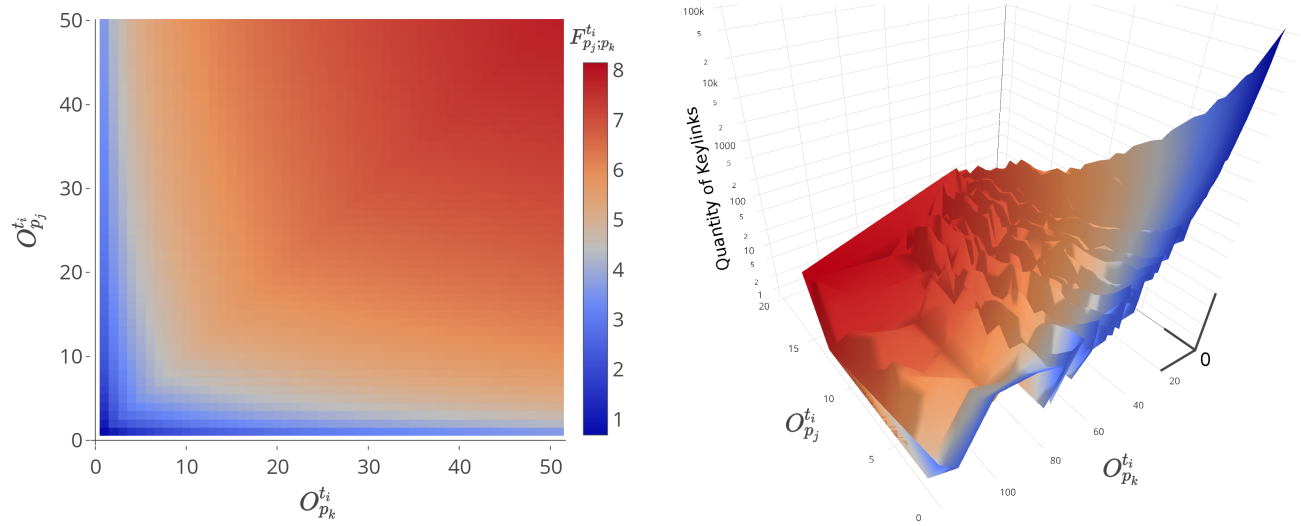
$$F^{t_i}_{p_j;p_k} = \log((O^{t_i}_{p_j} + O^{t_i}_{p_k}) \times S^{t_i}_{p_j;p_k})$$

avec $S^{t_i}_{p_j;p_k} = \min(O^{t_i}_{p_j}; O^{t_i}_{p_k})$ La figure 2.18a montre la pondération $F^{t_i}_{p_j;p_k}$, la figure figure 2.18b montre le drapage du gradient de pondération sur les quantités de liens en fonction de leur répartition. Pour la facilité de requête, les valeurs sont normalisées entre [0,1]. Différentes propriétés sont remarquables (figure 2.18b) :

— Le pic bleu : la plupart des liens sont issus de 2 *pages* partageant peu d'occurrences du même skip-gram ($O^{t_i}_{p_j}$ and $O^{t_i}_{p_k}$ proches de 1). Dans ce cas la pondération est faible (bleue).
— La bordure bleue coté lecteur : peu importe qu'une des pages contienne de nombreuses occurrences d'un skip-gram (ici la *page* p_k contient jusqu'à 100 occurrences), la pondération restera faible (bleue) si l'autre page (p_j) contient peu d'occurrences du même skip-gram.
— La zone rouge : au-delà d'un seuil d'occurrences communes (environ $\min(O^{t_i}_{p_j}; O^{t_i}_{p_k}) > 5$), une pondération stable et élevée concerne moins de 10 *liens-clés* pour chaque couple (p_j, p_k) (petits pics rouges).

5.2.2 Création des liens uniques

Comme le nom l'indique, le lien unique est un unique arc pondéré calculé pour chaque paire de *pages*. Ce lien permet une vue plus macroscopique corpus. En effet, chaque *page* peut entretenir un grand nombre d'arcs avec une autre page, rendant la lecture du graphe difficile. Un lien unique entre 2 *pages* correspond à la somme des *liens-clés* entre ces mêmes *pages*. La valeur



(a) Gradient de pondération F en fonction de la répartition des occurrences dans la paire de *pages* $O^{t_i}_{p_j}$ et $O^{t_i}_{p_k}$

(b) Gradient de pondération corrélé à la quantité de liens dans un corpus de 100 documents, résultats similaires sur les autres corpus.

FIGURE 2.18 – F est calculé en fonction de la distribution du skip-gram dans la paire de pages.

	page A (raw TF)	page B (raw TF)	page A (occ. val.)	page B (occ. val.)
pignon	0	3	4.58	4.66
ordinateur	1	0	1.71	3.12
patrimoine	3	0	3.47	2.4
engrenage	5	0	5.4	4.71
modélisation	0	5	3.4	6.17
monument	0	1	3.13	2.53
cos (A, B)	0		0.94	

TABLE 2.4 – Distance cosinus : vecteurs d'occurrences VS. vecteurs de valeurs. À gauche les documents sont orthogonaux ($\cos(A,B)=0$) ; à droite les documents sont proches ($\cos(A,B)\approx 0.94$). Pourtant ce sont les mêmes documents, seul le calcul des représentations change.

du lien est rapportée à la longueur (en nombre de mots) de la paire de page. En effet, les pages comportant davantage de mots, sont potentiellement plus liées aux autres :

$$\forall (p_j,p_k), \exists!\ L_{jk} = \sum_{i=1}^{n} \frac{P^{t_i}_{p_j;p_k}}{m_j + m_k}$$

avec m_j le nombre de lemmes (y compris différents) contenus dans le document j.

Les mesures citées en section 4.1.4 combinées avec les valeurs d'occurrences définies en section 5.1.3 permettraient d'obtenir ce type de résultats. Par exemple des distances cosinus entre documents représentés par des vecteurs de valeurs d'occurrences des termes du corpus. Le tableau 2.4 reprend l'exemple des sections 5.1.3 et 5.1.1. Ce tableau montre les différences entre les distances cosinus (voir section 4.1.4) de 2 documents en fonction des représentations. Pour la cohérence entre les niveaux d'échelles, c'est-à-dire entre le niveau *lien-clé* et le niveau lien unique, nous utilisons la somme des *liens-clés*. Par ailleurs, cette technique évite un calcul supplémentaire. L'ensemble des classes identifiées sur les *liens-clés* est reporté sur le lien unique, si bien qu'il est possible de requêter ces liens par classe.

5.2.3 Création des liens négatifs

Nous proposons les liens négatifs comme une mesure complémentaire de la rareté d'un terme (*lien-clé*, section 5.2.1) : la rareté de l'absence d'un terme est l'absence d'un skip-gram générique et de tout skip-gram similaire (suivant les résultats des valeurs d'occurrence, section 5.1.3). Lorsque
1. presque toutes les *pages* mentionnent un skip-gram
2. quelques *pages* ne le mentionnent pas ni aucune variantes (formes composées), ni terme similaire (synonyme)
alors on considère que les quelques *pages* qui ne contienne pas l'expression forment une caractéristique saillante du corpus. Par exemple, dans un corpus de 200 articles en gestion des connaissances, seuls 2 articles ne mentionnent pas le lemme « connais-

sance », alors ils sont liés par un lien négatif.

$$LienNeg^{t_i}_{p_j,p_k} = 1 \ \ if \ \ \overline{DF^{t_i}} \leq a$$

avec :

— $a \in [0, 3]$ la zone stable. Plutôt entre 1 et 2 pour éviter le bruit. 0 désactive ce type de liens.
— γ la valeur d'occurrence seuil de t_i pour considérer que t_i et tout terme similaire (synonyme) est absent. On fixe $\gamma = 0.8$ généralement.
— $\overline{DF^{t_i}_{\gamma}} = |\{p_j : t_i \notin p_j\}|; j \in [0, m]$ les *pages* où la valeur d'occurrence de t_i est inférieure à γ.

Puisqu'il est impossible de compter le nombre d'occurrences d'un skip-gram absent, et que leur valeur d'IDF est nulle (car générique), le lien est binaire.

5.2.4 Liens en boucle

Les skip-grams n'occurrant que dans une seule *page* ne forment aucun lien (besoin d'une paire de *pages*) en propre. Pourtant ils sont porteurs de sens. Des liens boucles (*loop*) sont alors créés. Les termes similaires (valeur d'occurrence) sont également pris en compte, en effet un skip-gram peut former un lien avec un skip-gram différent si sa valeur d'occurrence est proche (synonyme), on fixe généralement ce seuil (γ) à 0.8. Seuls le skip-grams suffisamment occurrents dans la *page* sont pris en compte.

$$Loop^{t_i}_{p_j} = O^{t_i}_{p_j} \ \ if \ \ O^{t_i}_{p_j} \geq a$$

avec :

— $p_j : \forall j \in [0, m], \exists! \ j : t_i(\gamma) \in p_j$
— γ la valeur d'occurrence (section 5.1.3) seuil de t_i pour considérer qu'un autre terme est équivalent à t_i (synonyme). On fixe $\gamma = 0.8$ généralement.
— $a \in [2, 10]$: en fonction de la longueur du corpus, environ $n/100$ avec un minimum de 2.

5.2.5 Proximité sur les métadonnées

Trois autres types de liens sont produits. Ils sont optionnels, car basés sur les métadonnées des *pages* spécifiées précédemment (section 2.1.5). Si les métadonnées ne sont pas mentionnées, les liens ne sont pas créés. Ces mesures sont très simples.

Proximité temporelle. Dans un espace unidimensionnel, 3 cas existent : intervalle-intervalle, intervalle-point, ou point-point. Lorsqu'il y a inclusion d'un élément (date ou intervalle) dans un intervalle, la proximité est maximale ($= 1$). Il ne s'agit pas d'une distance, car cette mesure ne respecte pas la condition de séparation (section 4.1).

Proximité spatiale. Dans un espace plan, une *page* peut être localisée à plusieurs endroits si nécessaire (spécifier plusieurs localisations en métadonnées). La proximité entre les points les plus proches est calculée. Il n'existe pas de zones, ni d'incertitudes. Il ne s'agit pas d'une distance, car cette mesure ne respecte pas la condition d'inégalité triangulaire (section 4.1), en effet les proximités des couples de *pages* (A, B) et (B,C) peuvent être élevées, et celle (A, C) faible, si B a plusieurs localisations.

Proximité de *topics*. La proximité de *topics* est la somme des coefficients topic-*page* de la matrice W (section 2.2) pour chaque *page* de la paire. Un ratio pénalisant est appliqué lorsque les *topics* ne sont pas identiques. Trois cas existent pour la proximité de *topics* (k) entre 2 pages p_1 et p_2 :

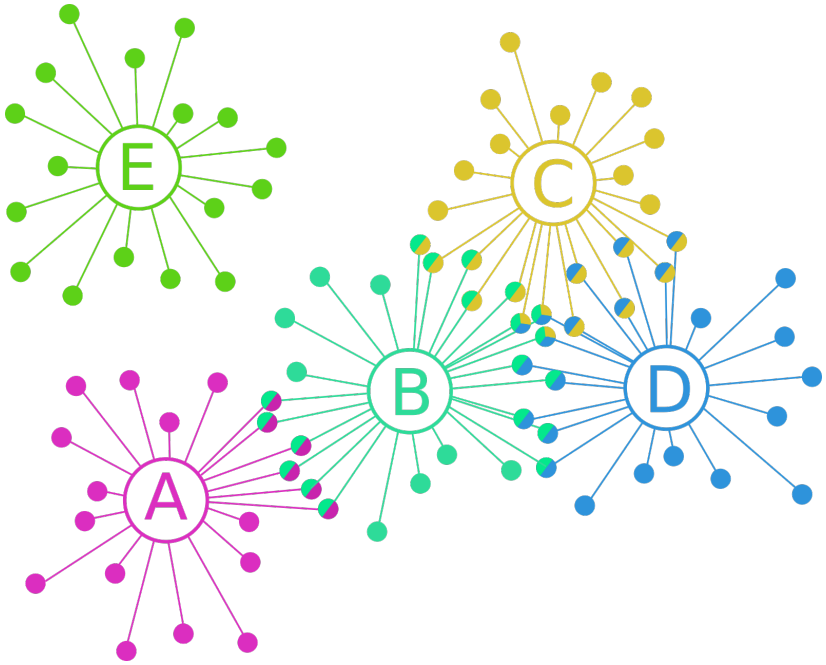
— Pour les *topics* (k) communs à la paire de *pages* : la proximité est calculée en sommant les pondérations des coefficients topic-*page* de chacune des 2 pages dans tous les *topic* communs. Par exemple, sur la figure 2.19a, entre une page mixte représentant les topics A, B et une page représentant A et/ou B. C'est un cas particulier du troisième cas. $P(p_1, p_2) = \sum^{k} W[p_1, k] + W[p_2, k]$ avec k les topics commun à p_1 et p_2.
— Pour les *topics* disjoints de la paire de *pages*, la proximité est alors minimale ($= 0$). Par exemple, sur la figure 2.19a, entre une *page* de E et n'importe quelle *page* d'un autre *topic* (non-E).
— Pour les autres topics (connectés mais non-identiques), chaque *page* de la paire se lie à ses *topics* respectifs, puis une pénalité (le carré de la distance entre les *topics*) est appliquée.

$$P(p_1, p_2) = \sum^{i,j} \frac{W[p_1, k_i] + W[p_2, k_j]}{2(1 + D(k_i, k_j))^2} \tag{2.9}$$

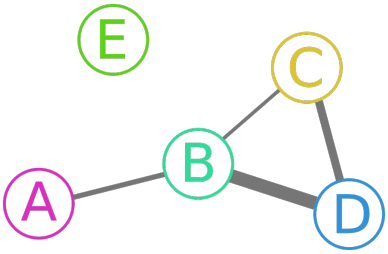
avec

— $W[p, k]$ le coefficient de la matrice W entre la page p et le *topic* k
— $D(k_i, k_j) \in \mathbb{Z}$ la longueur du chemin le plus court entre les *topics* k_i et k_j.

Le ratio de 2 compense les paires symétriques (ex : $i = 1, j = 2$ et $i = 2, j = 1$). Le carré du chemin le plus court pénalise fortement les *topics* distants. Pour calculer $D(k_1, k_2)$ on utilise l'algorithme de Dijkstra (1959) sur le graphe des proximités entre *topics* (figure 2.19b). Un arc entre 2 *topics* (graphe de proximité, figure 2.19b) est obtenu en sommant les pondérations des arcs des *pages* communes à ces 2 *topics*. Ce calcul suppose que 2 pages moyennement proches de plusieurs topics communs sont proches si ces topics communs sont proches. Un calcul plus fin de la proximité entre pages prendrait en compte les mots reliant les *pages* aux topics. Ainsi 2 pages distantes de plusieurs topics distants entre eux, pourraient néanmoins être proches entre elles (d'un topic virtuel non détecté), par exemple des pages de A et C (mais non-B) sur la figure 2.19a. À l'inverse, des pages distantes de plusieurs topics communs pourraient s'avérer éloignées par exemple celles aux confins des *topics* B, C et D sur la figure 2.19a. Ce type de détection correspond aux liens uniques, affranchis de l'existence de topics pré-établis, calculant la proximité entre *pages* sur la base de mots discriminants communs.



(a) Représentation en graphe des *pages* (points de couleurs) et topics (nommés par un lettre capitale) à partir des pondérations de la matrice de proximité W.



(b) Les distances entre *topics* : la somme des pondérations des arcs des *pages* communes.

FIGURE 2.19 – Calcul de proximité de *topics* entre *pages*

5.2.6 Détection d'anomalies

Les anomalies sont des points d'intérêt pour l'expert du corpus. Leur détection est basée sur l'entropie des arcs d'une paire de *pages*. Les poids des différents types d'arcs (calculés précédemment) sont considérés. On parle d'entropie de connectivité multidimensionnelle. 3 cas sont vérifiés :

— chaînon manquant : tous les arcs sauf 1 sont fortement pondérés.
— extrême : généralise le cas précédent, chaque arc est soit fortement, soit faiblement pondéré.
— jumeaux : un arc de lien unique (section 5.2.2) présente une pondération anormalement élevée.

Concrètement ces anomalies concernent une paire de *pages*, elles sont renseignées par un arc supplémentaire de type *anomalie* entre les pages concernées. Ce type de détection d'anomalies permet une première approche automatique. L'expertise humaine est nécessaire pour déceler d'autres anomalies, souvent plus intéressantes.

6 Résultats

6.1 Performance de l'extraction de terminologie

6.1.1 Temps d'exécution

Les résultats de l'extraction de terminologie peuvent être mesurés. *ANA+* réalise une extraction rapide avec un temps d'exécution linéaire (voir figure 2.20). L'implémentation actuelle dépend de la mémoire (environ 0.5Go par million de mots). Cette dépendance n'est pas problématique dans le cadre de la destination d'usage. Le nombre de skip-grams extraits dépend des seuils. Les seuils automatiques visent à en extraire environ 3k et jusqu'à 10k. La précision est influencée par ces seuils, une application requérant un rappel élevé et un autorisant une basse précision, peut abaisser ces seuils.

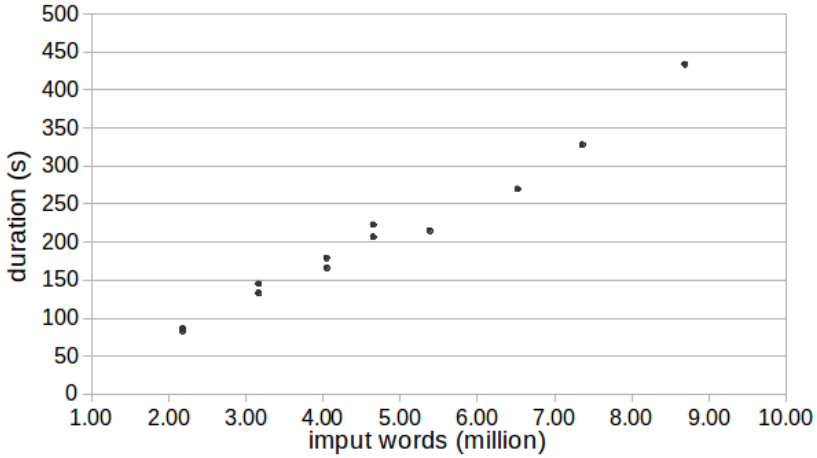


FIGURE 2.20 – Le temps d'extraction de la terminologie par *ANA+* est linéaire, proche de 1 seconde pour 20k mots.

corpus	top 100	top 25%	top 50%	all	fusion.
SSC	0.99	0.89	0.78	0.59	0.81
UN	0.96	0.87	0.76	0.54	0.78
CIRP	0.91	0.84	0.71	0.48	0.70

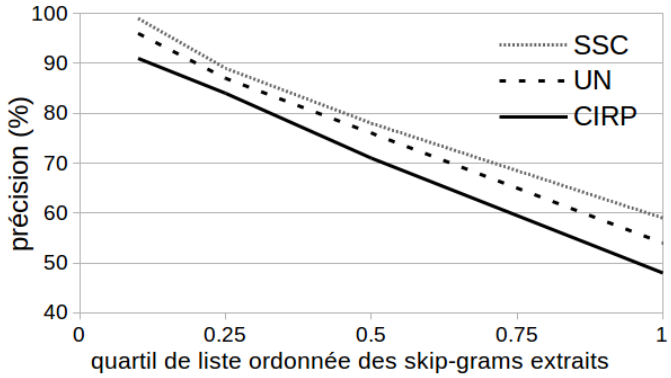


TABLE 2.5 – Précision de *ANA+* sur différents corpus

6.1.2 Précision de *ANA+*

La précision présentée dans le tableau 2.5 est organisée par tranches de skip-grams classés (hiérarchisés).

Calcul de précision. La précision est le ratio entre la liste de skip-grams censurés (après avoir retiré manuellement toutes les extractions incorrectes) et les listes de skip-grams initiales. Les corpus sur lesquels la précision a été mesurée sont : Chimie du solide (CdS) en Français ; expertise de l'ICOMOS pour le patrimoine mondial de l'UNESCO (UN) en Français, et *CIRP annals manufacturing technology* (CIRP) en anglais. Toujours dans le tableau 2.5, la colonne « fusion. » (pour fusionnés) considère les résultats sur l'ensemble de l'extraction après avoir accepté toutes les propositions de fusions (voir section 4.3). La précision semble meilleure (peu de faux positifs) en français, et chute de manière constante (voir le tableau et la figure 2.5). Cela signifie que le classement (*ranking*, section 4.2) est efficace. En corollaire, la dernière colonne montre que les fusions proposées améliorent beaucoup la précision. En effet, la plupart des faux positifs sont fusionnés avec leurs parents et disparaissent donc de la liste.

Précision et fusions. Une mesure de la précision des fusions donne un score approximatif de 0.85, c'est à dire que les propositions de fusion qui concernent des skip-grams qui auraient du être conservés sont rares. Les fusions erronées (c'est-à-dire les 0.15 restantes) impliquent une baisse de rappel. En effet, ces fusions erronées font disparaître de la liste des skip-grams qui auraient du être trouvés (faux négatif). En production, nous voulons éviter cet effet à tout prix. Pour cette raison, les fusions ne sont jamais automatiquement appliquées, l'expert valide toujours les propositions. Pour cette raison, la mesure de précision des fusions ne nous intéresse pas davantage. Néanmoins, pour la mesure des performances d'*ANA+*, nous retiendrons ces fusions automatiques qui améliorent le F-score.

6.1.3 Rappel et F-mesure

Comme proposé par Frantzi *et al.* (2000), pour faire face à la difficulté de compter les faux–positifs, nous mesurons une F-mesure relative. La procédure est la suivante :

1. Nous réalisons l'extraction de terminologie sur le même corpus avec *ANA+* et avec *TTC Termsuite* (Cram et Daille, 2016). Les résultats sont 2 listes de skip-grams ordonnées.
2. Les fusions de la liste issue de *ANA+* sont toutes acceptées à l'aveugle.

	rappel		F-mesure	
corpus	TS	*ANA+*	TS	*ANA+*
exact	0.60	0.61	0.58	0.69
racines	0.60	0.63	0.58	0.71
imbriqués	0.64	0.64	0.60	0.72

TABLE 2.6 – Résultats comparés de rappel et F-mesure pour *Termsuite* (TS) et *ANA+*.

3. Seuls les skip-grams les mieux notés de la liste la plus longue sont conservés de sorte à ce que les 2 listes soient de même longueur.
4. Les 2 listes sont modérées par l'expert. La précision de ces listes est donc maximale (précision = 1).
5. Une liste de référence est construite de l'union (fusion) de ces 2 listes modérées.
6. Selon plusieurs filtres (lignes du tableau 2.6) les mots manquants dans une liste par rapport à la liste de référence sont considérés comme des faux négatifs. Le rappel et la F-mesure sont calculés sur cette base.

Le tableau 2.6 présente les résultats de rappel relatif pour le corpus de chimie du solide. Sur ce corpus la précision de *ANA+* est 0.81 (après fusions, voir tableau 2.5) ; la précision de *TTC termsuite* est 0.57 par le même procédé (modération manuelle des résultats).

La ligne « exact » mesure l'écart brut entre les listes et la référence. La ligne « racines » considère les mots qui ont la même racine comme équivalent (ex : « supraconductivité » et « supraconducteur »), cela a pour effet de réduire légèrement la contribution de *Termsuite* à la liste de référence, et d'augmenter donc le rappel d'*ANA+*. La ligne « imbriqués » accepte les références incluses dans un skip-gram de l'une des 2 listes. L'avantage de *Termsuite* sur ce point signifie que *Termsuite* compose des MWE plus complexes que *ANA+*, par exemple *Termsuite* ne trouve pas « propriété mécanique » mais « propriété mécanique des fibres ». Une étude plus approfondie montre que « propriété mécanique » fait partie des éléments tronqués de la liste car mal notés.

6.2 Résultats de *Haruspex*

6.2.1 Interactions, interface, visualisation

Un des objectifs de *Haruspex* est d'offrir à l'historien de nouvelles vues de son corpus, de soulever de nouvelles questions de recherche et de proposer un moyen d'enquête quantitative. L'enjeu de cette étape, production de résultats de pipeline (ETL) est de proposer des **vues** adéquates. Cette notion est abordée plus précisément dans les chapitres 3 et 4. Dans un premier temps, la notion d'interface intervient fortement. En effet l'objectif n'est pas de produire des résultats de qualité historique « en autonomie » mais d'interagir au mieux avec les connaissances et les modes de travail de l'historien.

Jusqu'à maintenant nous entendions le terme « interface » dans une acceptation plus large que la forme graphique. Il s'agissait principalement d'expérience et d'interactions. Il faut maintenant intégrer la composante graphique, sans délaisser les interactions. La forme du résultat de *Haruspex* est un graphe multiple pondéré contenant tous les liens créés (section 5) entre les *pages*. Ce graphe n'est pas visualisable dans son ensemble malgré le faible nombre de nœuds (quelques centaines, rarement milliers, de *pages* maximum). En effet, il contient plusieurs dizaines de milliers de liens (essentiellement des *liens-clés*). Ce type de visualisation est compliqué à produire et offre de très faibles capacités d'interaction avec l'expert (illisible, difficile à manipuler).

L'enjeu se situe donc entre le graphe impossible à visualiser dans l'ensemble et les connaissances qualitatives de l'historien. Cet enjeu prend la forme d'interactions médiatisées par une interface de visualisation de données. Les notions de multi-échelles et de multi-dimensions interviennent. L'interface est un nouveau support pour les interactions : elle donne les mots pour poser les questions.

6.2.2 Formalisation du processus

Le processus de production de graphes intéressant pour l'historien est donc réalisé avec l'historien. Ce processus itère sur la suite d'interactions suivantes :

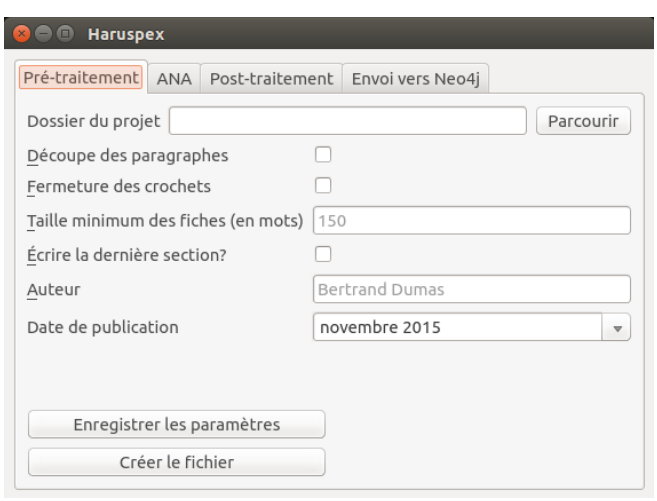
1. Poser une question de recherche à partir des analyses existantes
2. Reformuler la question en une requête explicite vers la base de données
3. Visualiser les résultats de la requête (graphe par exemple)
4. Analyse quantitative et visuelle des résultats

Une représentation plus détaillée du processus est établie à partir d'un cas concret d'utilisation en section 1.2.

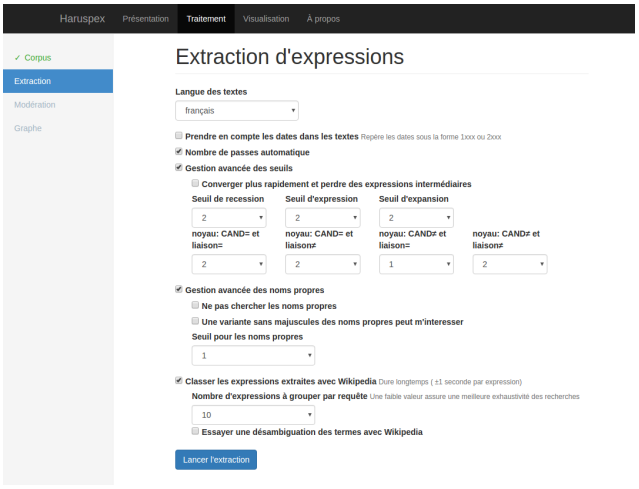
6.2.3 Mise en forme des résultats

Le résultat d'*Haruspex* est un graphe multiple pondéré contenant tous les liens créés (section 5) entre les *pages*. Il est stocké dans une base de données orientée graphe (*neo4J*). Des requêtes vers ce graphe permettent d'extraire certains aspects qui intéressent l'historien.

Pour visualiser les graphes on doit quitter l'interface d'*Haruspex* (voir figure 2.21) et utiliser un outil extérieur. De nombreux



(a) Première interface d'*Haruspex*, réalisée avec Qt5 en 2015



(b) Seconde interface d'*Haruspex*, utilisant les technologies web en 2017

FIGURE 2.21 – Interfaces d'*Haruspex*

logiciels et librairies permettent de visualiser des graphes. La contrainte de visualiser de multi-graphes pondérés, c'est-à-dire plusieurs arcs de taille variable entre les nœuds réduit ce choix. *Cytoscape* (Shannon *et al.*, 2003), initialement prévu pour la visualisation de molécule en bio-informatique, répond aux besoins. La figure 2.22 montre un exemple de graphe produit sur un corpus de 44 documents. Une analyse plus approfondie de ce corpus est proposée en chapitre 3. Sur cette figure, le corpus étudié est celui de chimie du solide, qui contient 44 *pages* (nœuds). Les arcs bleus appartiennent à la thématique de « physique », en rose à la « chimie ». La couleur des nœuds est liée à la discipline officielle du texte (avec le même code couleur). La bordure du nœud indique l'ancienneté (les foncés sont des témoignages plus anciens).

Conclusion

Le processus implémenté permet de traiter des corpus de texte techniques, de volume limité, spécifiques à un domaine, et sans interférence avec des données extérieures et sans préparation des données en amont. *Haruspex* propose une adaptation ou une amélioration de résultats précédents (ANA, TF-IDF, MED) combinés dans une perspective innovante : la création de graphes multiples flous. Sa modularité permet une mise à jour des composants (*topic-modelling*, extraction de terminologie). Dans le cas de nos besoins, les résultats de l'extraction de terminologie améliorent les récentes propositions de l'état de l'art.

Haruspex ne propose pas de raisonnement en soi. Le chapitre 3 montre un exemple d'interfaçage avec les connaissances de l'historien dans un raisonnement aboutissant à de nouvelles connaissances historiques.

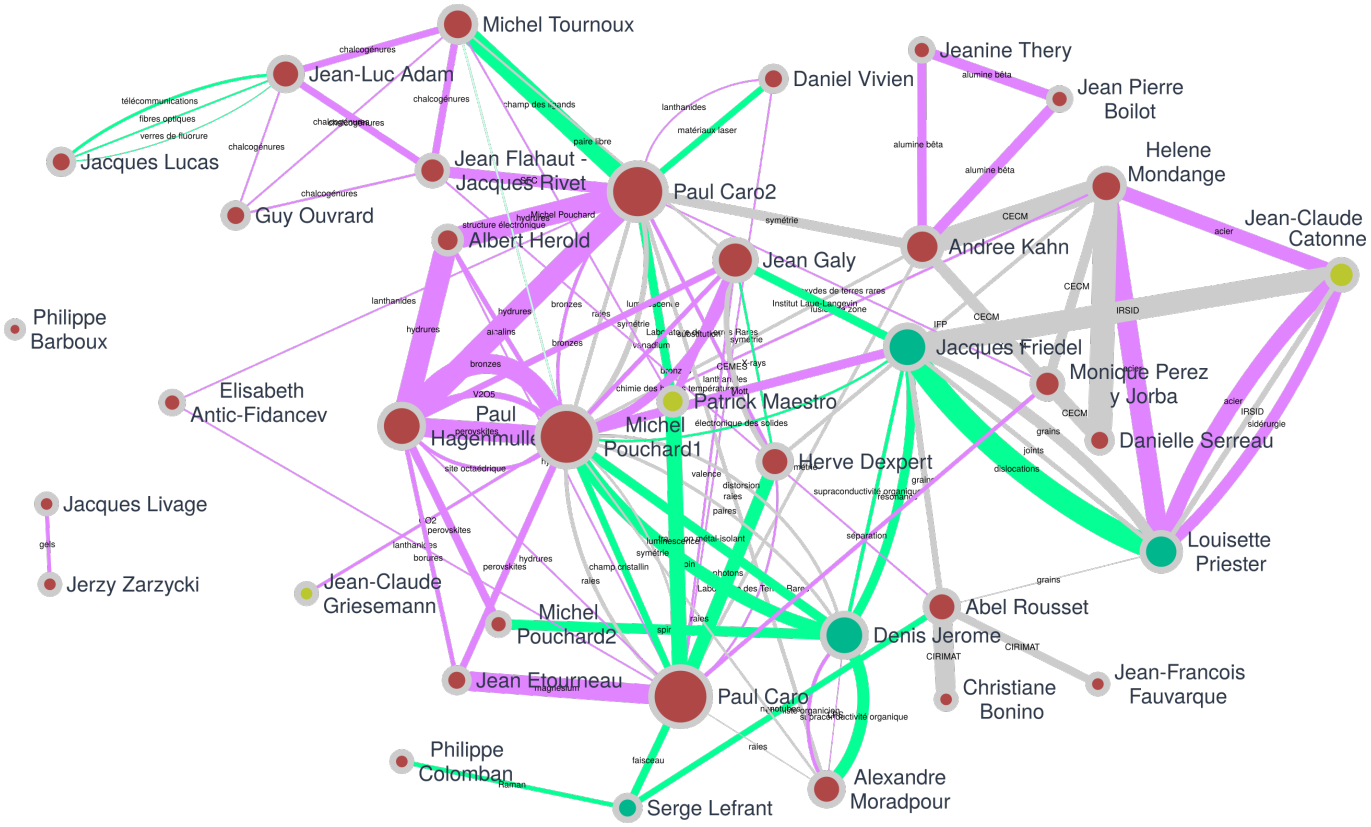


FIGURE 2.22 – Exemple de graphe multiple flou labellisé produit par une mise en forme de la réponse à une requête sur les résultats de *Haruspex*

Chapitre 3

Application : Cartographie d'un corpus en Histoire de la Chimie du Solide

> Computers act like « a telescope for the mind [...] enlarge the whole range of what its possessors could see and do [and so change] their whole picture of the world »
>
> Masterman (1962)

Contents

Cette étude est organisée en trois parties. Nous commençons par présenter l'objet d'étude — les archives orales —. Nous lui appliquons ensuite *Haruspex*. Ceci permet de fabriquer trois types d'image numérique : une cartographie générale du corpus ; des zooms sur des territoires plus restreints du corpus ; des visualisations spécifiques du corpus par la sélection de critères particuliers. Ces artefacts numériques sont confrontés aux connaissances historiennes soit pour confirmer des acquis antérieurs, l'artefact numérique jouant alors le rôle d'indice supplémentaire pour le raisonnement historien, soit pour donner des représentations surprenantes, l'artefact numérique jouant alors un rôle heuristique. Cette étude de cas se prolonge par une réflexion épistémologique et réflexive sur les relations entre les sphères numériques et historiques (Chapitre 5, section 5). Nous établissons ainsi une typologie des inférences permises par l'interaction hommes/machines et nous formalisons un schéma de fonctionnement de la méthode numérique dépassant le cadre de cette expérience.

Notes :

— Ce chapitre est une ré-écriture partielle d'un article publié conjointement avec Benjamin Hervy et Pierre Teissier.
— Le fonctionnement d'*Haruspex* n'était pas abouti lors de l'écriture. Parmi les plus gros décalage, on note l'absence de calculs de distance entre mots ("Valeurs d'occurrences d'un skip-gram dans une *page* " (5.1.3)).

1 Présentation du corpus d'archives orales

1.1 Histoire et mémoire de la recherche sur les matériaux au XXe siècle

Notre corpus d'archives orales s'inscrit dans un programme de recherche plus large consacré à l'histoire et à la mémoire de la « recherche sur les matériaux » ou *Materials Research* en anglais Bensaude-Vincent et Teissier (2015). Ce programme collectif est incarné depuis 2011 par le site internet Sciences : Histoire Orale [1], simultanément « lieu de mémoire » et espace de réflexion épistémologique. Dans cet ensemble étendu et bigarré, nous avons délimité un corpus plus homogène selon deux critères. Premièrement, nous avons choisi les seuls entretiens que l'un d'entre nous connaissait suffisamment pour les avoir menés ou les avoir utilisés de manière soutenue. Ceci est nécessaire pour pouvoir discuter les résultats numériques à partir d'une connaissance qualitative préalable. Deuxièmement, nous n'avons retenu que les textes en français car l'analyse sémantique par mots-clés exige une unicité de langue.

1.1.1 Composition

Formellement. Nous obtenons ainsi un corpus de 41 entretiens retranscrits, identifiés par le nom des personnes interviewées. Seules dans la plupart des interviews, on compte deux exceptions pour lesquelles deux personnes ont été interrogées ensemble. Les entretiens étaient semi-directifs, c'est-à-dire que l'interviewer(e) interroge l'interviewé(e) à partir d'une grille de questions préétablies. Les personnes interviewées ont ainsi été amenées à raconter leur carrière et leurs environnements professionnels, pour les plus anciens depuis les années 1940. Ces personnes sont à la fois témoins et actrices de la recherche sur les matériaux de la seconde moitié du XX^{e} siècle, que ce soit dans la sphère académique (chercheurs, enseignants-chercheurs, administratifs) ou industrielle (administrateurs, ingénieurs).

Qualitativement. Le corpus d'entretiens n'est pas d'un seul tenant. Il est composé de trois sous-corpus présentés dans la table 3.1. Ces trois sous-ensembles se distinguent par la date de constitution, l'interviewer principal, la grille de questions et le thème de recherche. Détaillons-les pour expliciter leur teneur. Le sous-corpus 3.2a est le plus ancien. Il est composé de 7 entretiens, réalisés entre 2000 et 2003, pour la plupart par B. Bensaude Vincent. Il provient d'un programme d'histoire de la science et ingénierie des matériaux, parrainé par la *Sloan Fondation* et le *Dibner Fund* du MIT (Cambridge, MA) [2]. Le sous-corpus 3.2b est le plus étendu et le plus homogène. Il est composé de 26 entretiens, réalisés entre 2004 et 2007, pour la plupart par P. Teissier dans le cadre d'une thèse sur l'histoire de la chimie du solide en France Teissier (2007). Le sous-corpus 3.2c est le plus hétérogène. Il est composé de 8 entretiens, réalisés entre 2009 et 2016, par P. Teissier, en lien avec plusieurs recherches historiques : chimie, matériaux, batteries et piles à combustible, voitures électriques et hydrogène.

1.1.2 Un corpus d'historien

L'étude du corpus se justifie, au niveau des sciences humaines, par une cohérence temporelle, géographique et thématique. En effet, les témoignages des acteurs concernent la même période, du milieu du XX^{e} siècle au début du XXI^{e} siècle, un même espace, le territoire français pour l'essentiel, et un thème général unique, la « recherche sur les matériaux » ou « recherche en matériaux » ou encore *Materials Research* en anglais Bensaude-Vincent et Teissier (2015). Cette appellation générique a l'avantage de couvrir le spectre des disciplines scientifiques liées aux matériaux : mécanique, physique, chimie, biologie, sciences pour l'ingénieur, électronique, etc., sans favoriser des spécificités nationales. Ainsi, utiliser *Materials Science and Engineering* aurait induit un biais historiographique en prenant les États-Unis comme modèle planétaire par rapport auquel les autres nations mesureraient un

1. https://www.sho.espci.fr/?lang=fr
2. Les principaux chercheurs du programme, H. Arribart, B. Bensaude Vincent et A. Hessenbruch, ont ainsi collecté une trentaine d'entretiens de chercheurs en matériaux ayant fait leur carrière en Europe, aux États-Unis ou au Japon. Voir http://authors.library.caltech.edu/5456/1/hrst.mit.edu/hrs/materials/public/general.htm

(a) Corpus 2000-2003 (B.B.V.)

Interviewé-e(s)	Date interview
Barboux, Philippe	2000
Boilot, Jean-Pierre	2000
Colomban, Philippe	2003
Friedel, Jacques	2001
Griesemann, Jean-Claude	2001
Livage, Jacques	2001
Zarzycki, Jerzy	2001

(b) Corpus 2004-2007 (P.T.)

Interviewé-e(s)	Date interview
Beuzit, Pierre	2010
Catonné, Jean-Claude	2012
Fauvarque, Jean-François	2013
Lisse, Jean-Pierre	2009
Lucchese, Paul	2016
Poulain, Marcel et Michel	2016
Priester, Louisette	2011
Vitet Sylvain	2009

(c) Corpus 2009-2016 (P.T.)

Interviewé-e(s)	Date interview	Interviewé-e(s)	Date interview
Adam, Jean-Luc	2006	Lefrant, Serge	2006
Antic-Fidancev, Élisabeth	2006	Vitorge, Marie-Claude	2007
Bonino, Christiane	2006	Vivien, Daniel	2004
Caro, Paul	2005	Tournoux, Michel	2006
Dexpert, Hervé	2006	Théry, Jeanine	2004
Étourneau, Jean	2005	Serreau, Dannielle	2004
Flahaut, Jean et Rivet, Jacques	2005	Rousset, Abel	2006
Hérold, Albert	2006	Pouchard, Michel	2004
Galy, Jean	2006	Perez y Jorba, Monique	2004
Hagenmuller, Paul	2004	Ouvrard, Guy	2006
Jérôme, Denis	2006	Moradpour, Alexandre	2006
Kahn-Harari, Andrée	2004	Mondange, Hélène	2004
Lucas, Jacques	2005	Maestro, Patrick	2007

TABLE 3.1 – Constitution du corpus d'entretiens

« retard ». C'est contre cette téléologie spontanée des acteurs des matériaux que se situe le programme de recherche collectif, incarné depuis 2011 par le site internet [3].

Études préalables. Hébergé par l'ESPCI ParisTech, ce site ouvre « un lieu de mémoire » et un espace de réflexion épistémologique. C'est un lieu de mémoire parce qu'il a pour ambition de mettre à disposition, de manière libre, un ensemble de corpus constitué sous la coordination de B. Bensaude-Vincent, depuis le début des années 2000. Par ordre de constitution, on trouve les domaines suivants : science et ingénierie des matériaux (B. Bensaude-Vincent, puis E. Bertrand), chimie du solide (P. Teissier), catalyse (B. Voillequin), nanomachines moléculaires (X. Guchet et S. Loeve), spectroscopie théorique (S. Loeve), toxicologie des nanoparticules (E. Malcotte), biopuces d'ADN et protéines (X. Guchet, R. Leroux et S. Pellé), cellules souches cancéreuses (L. Laplane), etc. Le site incorpore aussi des notes et réflexions épistémologiques sur le statut et la fonction des archives orales dans le travail historique. Il a donné lieu à plusieurs communications orales articulant histoire, mémoire et humanités numériques Teissier et Loeve (2013); Bensaude-Vincent et Teissier (2015).

Remarque sur les archives orales. Les archives orales constituent une mise en mémoire de l'histoire par la collecte de témoignages d'acteurs du processus historique. Elles sont devenues courantes pour étudier les sciences contemporaines même si elles posent des problèmes spécifiques par rapport aux archives classiques : témoignage fabriqué suite à la demande de l'historien, ce qui inverse la temporalité de production de l'archive, position irréductiblement située du témoignage, dimension interpersonnelle entre l'interviewé(e) et l'interviewer(e), etc. de Chadarevian (1997). Mais ce sont ces mêmes ambiguïtés qui en font aussi tout l'intérêt pour la recherche historique parce qu'elles permettent de rentrer dans l'époque contemporaine par la « micro-histoire » en mettant en valeur les contradictions de mémoire, les subjectivités individuelles mais aussi les récurrences de langage, d'explication et de mythes propres à une communauté donnée.

1.2 Découpage du corpus

Ajustements. Une première version du corpus contient 44 documents (interviews). Un premier test permet d'identifier que les interviews de certains chercheurs constituent des anomalies : il s'agit de paires d'interviews d'un même chercheur (voir section 3.1). Les deux paires de nœuds les plus importantes étaient alors formées par ces entretiens réalisés à deux moments différents : M. Pouchard (septembre et décembre 2004) et P. Caro (2002 et 2005). La diversité des interviewers et des questions

3. Sciences : Histoire Orale `http://www.sho.espci.fr/spip.php?article3`

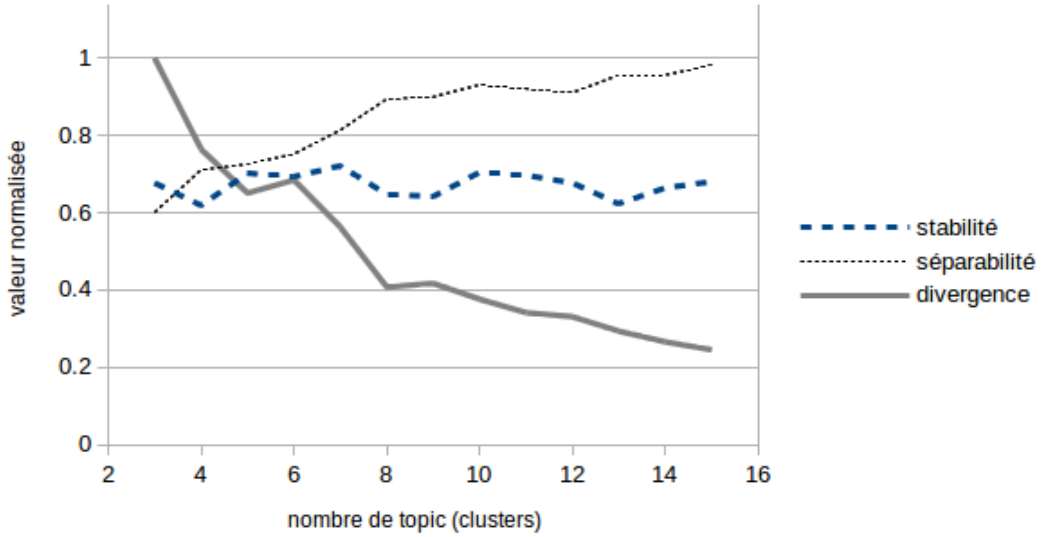


FIGURE 3.1 – Résultats de test pour determiner le nombre de *topics*

et l'éloignement temporel des deux entretiens de P. Caro ne modifiaient pas sensiblement leur proximité, ce qui signifie que le témoignage de chacun porte une forte dimension idiosyncrasique. Il n'en est pas pour autant isolé du corpus puisque chacun des entretiens par paire établissait des liens différents avec des entretiens tiers. Ainsi, l'interview Caro 1 était lié à l'interview P. Maestro sans que Caro 2 le soit. C'était l'inverse entre Caro 2 et Pouchard 2. Ces paires de témoignages conduisaient à la structuration de mini-cluster trompeurs, structurés autour d'une pensée qui aurait artificiellement fait école. Nous décidons de réduire ces paires à une seule interview en supprimant celle se référant à une grille de questions différente du reste du corpus.

Homogénéité. Une analyse rapide du corpus par *topic modelling* permet de s'assurer de l'homogénéité du contenu des documents constituant le corpus : le corpus semble indivisible sur la base des mots pris un à un. Aucun paramétrage de l'algorithme ne parvient à discerner de catégories latentes. En d'autres mots, il n'est pas possible d'identifier de "thématiques" discriminantes au sein du corpus sur la seule base des mots employés. Cette caractéristique est nécessaire d'un point de vue méthodologique pour l'analyse des résultats produits par *Haruspex*.

Pour ce corpus, aucun découpage en sous-corpus (*topic modelling*) n'est effectué. Les test effectués tendent à montrer qu'un tel découpage serait compliqué à réaliser automatiquement. Les résultats de ces tests (présentés en figure 3.1) montrent que les résultats sont continuellement meilleurs en augmentant le nombre de *topics* jusqu'à 15. Découper un corpus de 41 documents en plus de 15 *topics* semble inutile. Une première analyse concerne l'hyper-spécialisation des textes du corpus, possédant des affinités par petits groupes, voire parfois isolés. Une analyse complémentaire concerne l'homogénéité du corpus, dont le partitionnement n'est pas trivial. Ces tests de stabilité et de divergence sont présentés en section 2.2.2.

1.3 Enjeux de l'analyse numérique d'archives orales

Le corpus est intéressant pour l'analyse numérique dans la mesure où il offre un niveau intermédiaire de granulométrie : suffisamment cohérent autour de la recherche sur les matériaux et suffisamment hétérogène par la présence de trois sous-corpus. Les caractéristiques quantitatives de base du corpus sont présentées dans la table 3.3.

format de fichier	Open Document
structure interne (titre, sous-titre, etc.)	non
nombre de documents	41
nombre de mots	339k
nombre de mots après filtre	87 316
nombre de lemmes différents	9 884
moyenne du nb de mots par document	8268
écart-type du nb de mots par document	4837

TABLE 3.3 – Caractéristiques du corpus étudié

Pluridisciplinarité. Jusqu'à présent, ce collectif de recherche n'a pas analysé de manière quantitative les différents corpus constitués par chacun de ses membres. C'est l'une des motivations de ce chapitre de fournir des outils numériques pour une telle analyse. Plutôt que de traiter l'ensemble des corpus disponibles qui représentent plus de deux cents entretiens, nous avons préféré nous focaliser sur le seul corpus que l'un d'entre nous a constitué depuis une douzaine d'années. Il semblait en effet indispensable de pouvoir tester la robustesse de la méthode numérique par une connaissance qualitative préalable. Nous avons aussi choisi d'agréger trois sous-ensembles, relativement différents, pour amplifier l'hétérogénéité du corpus (voir tableau 3.1). Ceci permet de disposer de textes proches au niveau sémantique (au sein du sous-ensemble 2), de textes éloignés (de 1 à 2) et très éloignés (de 1 à 3). Nous pourrons ainsi vérifier que la méthode numérique est capable de mettre en évidence la distance sémantique plus ou moins élevée entre des documents appartenant à des sous-ensembles. Néanmoins, nous devrons aussi garder à l'esprit que le nombre de variables extérieures est augmentée d'autant. Ainsi, la variabilité des grilles de questions d'un sous-ensemble à l'autre, voire d'une interview à l'autre, a une influence certaine sur les relations entre les entretiens, mais difficile à quantifier. C'est un risque à prendre si nous voulons développer des outils numériques pertinents, qui pourront servir à l'analyse étendue de la mémoire en sciences des matériaux.

L'entretien forme le nœud de base de notre analyse quantitative. Il est repéré par un certain nombre de méta-données sur l'âge du témoin, son genre, sa nationalité, son affiliation professionnelle, etc.

Enjeux épistémiques. Un lien entre deux nœuds est fortement déterminé par le contenu sémantique des nœuds. Or, un entretien est un texte extrêmement complexe. Il contient des informations : itinéraire professionnel de l'interviewé(e), ses institutions de rattachement, des collègues plus ou moins proches, des thèmes de recherche, des laboratoires, des communautés, des théories et instrument, des politiques scientifiques, des stratégies commerciales et industrielles. Il contient en outre une vision spécifique — celle de l'individu — du champ de recherche à travers l'espace social (institutionnel), géographique (villes et pays) et mémoriel (durée et précision des souvenirs variables en fonction des individus et des générations). Il entrecroise ainsi des questions techniques (recherche), des énoncés relationnels et affectifs (interpersonnels), des positionnements identitaires (discipline, génération, genre, etc.) et des jugements culturels. En résumé, un entretien tisse une multitude de niveaux de discours, de références plus ou moins explicites et d'interprétations des mondes humains et naturels. Il contient un nombre indéfini de significations. Au contraire, l'analyse numérique construit ses données comme des nombres et les relations entre les entretiens comme des calculs de nombres.

Il y a donc une irréductibilité des données entre sciences informatiques et sciences humaines. Ces données sont numériques et calculatoires pour les premières, sémantiques et herméneutiques pour les secondes. Cette rupture soulève quatre enjeux épistémiques de difficulté croissante :

- Identifier différents niveaux de discours : scientifique, technique, institutionnel et politique certes, mais aussi épistémologique, parfois moral ou affectif.
- Saisir, à un niveau de discours donné, des structures : par exemple, au niveau institutionnel, repérer une école de recherche ou une communauté scientifique.
- Faire apparaître la dimension temporelle, qui focalise, sans doute, la principale interrogation des historiens : la structuration des discours pourrait par exemple faire apparaître des éléments propres à chaque génération de témoins.
- Appréhender un méta-discours comme celui des épistémologues, qui produisent des discours savants sur des discours savants : par exemple, l'analyse du travail du physicien qui décrit le comportement d'un nuage d'électrons.

2 Application d'*Haruspex* à la cartographie de la mémoire sur les matériaux

2.1 Extraction de terminologie par *Haruspex*

L'originalité principale d'*Haruspex* tient à la deuxième étape de son fonctionnement : l'extraction automatique d'expressions spécifiques d'un corpus, c'est-à-dire sans spécification *a priori* de ce qui constitue une expression à retenir. La table 3.4 donne quelques exemples d'expressions extraites de notre corpus auxquelles elle associe deux caractéristiques quantitatives : leur nombre d'occurrences dans le corpus et le nombre de documents concernés. Elle intègre aussi la troisième étape d'enrichissement des expressions par la qualification thématique (manuelle) des expressions au moyen de requêtes vers *Wikipedia*.

L'extraction automatique de données du corpus donne une liste brute de 2327 expressions. L'un d'entre nous a traité cette liste afin d'éliminer les expressions trop générales (articles, pronoms, etc.) et de valider les proposition de fusion de l'algorithme (ex : l'expression « supraconductivité » est fusionnée avec « supraconducteur »). Une modération de trois heures environ a ainsi permis de réduire la liste à 1372 expressions pertinentes.

2.1.1 Chaînage des connaissances pour l'exploration de corpus

Les liens entre document proviennent de cette terminologie (section 2.1). La présence d'une expression spécifique dans deux documents forme un lien entre les nœuds correspondant. L'intensité de ce lien spécifique est pondérée en fonction de la répartition statistique de l'expression au sein de l'ensemble du corpus. Par exemple, une expression comme « laboratoire » est trop répandue dans notre corpus pour révéler des relations signifiantes. Elle est donc rejetée. Ce sont les barres bleues à droites de la figure 3.2. À l'inverse, une expression comme « bronzes de vanadium », qui n'apparaît que dans cinq documents (cf. table

Forme extraite	Nb d'occ.	Nb. de doc. concernés	Thématique
Bronzes de vanadium	26	5	Sciences
Chimie de coordination	22	8	Chimie
Thèse de troisième cycle	16	7	France, Éducation
Microscopie électronique à transmission à haute résolution	3	2	Physique
Four solaire d'Odeillo	4	3	Industrie, Énergie, Sciences

TABLE 3.4 – Exemples d'expressions extraites, informations numériques associées et thématiques relatives

3.4), fournit des informations relationnelles. Le lien est donc validé. Son intensité est pondérée de manière d'autant plus forte que le nombre d'occurrences de l'expression est élevé et proche de l'équipartition entre documents concernés. Ainsi, pour « bronzes de vanadium », Michel Tournoux mentionne l'expression une seule fois alors que les quatre autres interviewés le mentionnent cinq fois ou plus : ce lien spécifique entre le nœud Tournoux et les quatre autres nœuds sera donc beaucoup moins fort que celui qui unit les quatre autres entre eux. Ceci évite de valoriser la simple évocation d'un sujet.

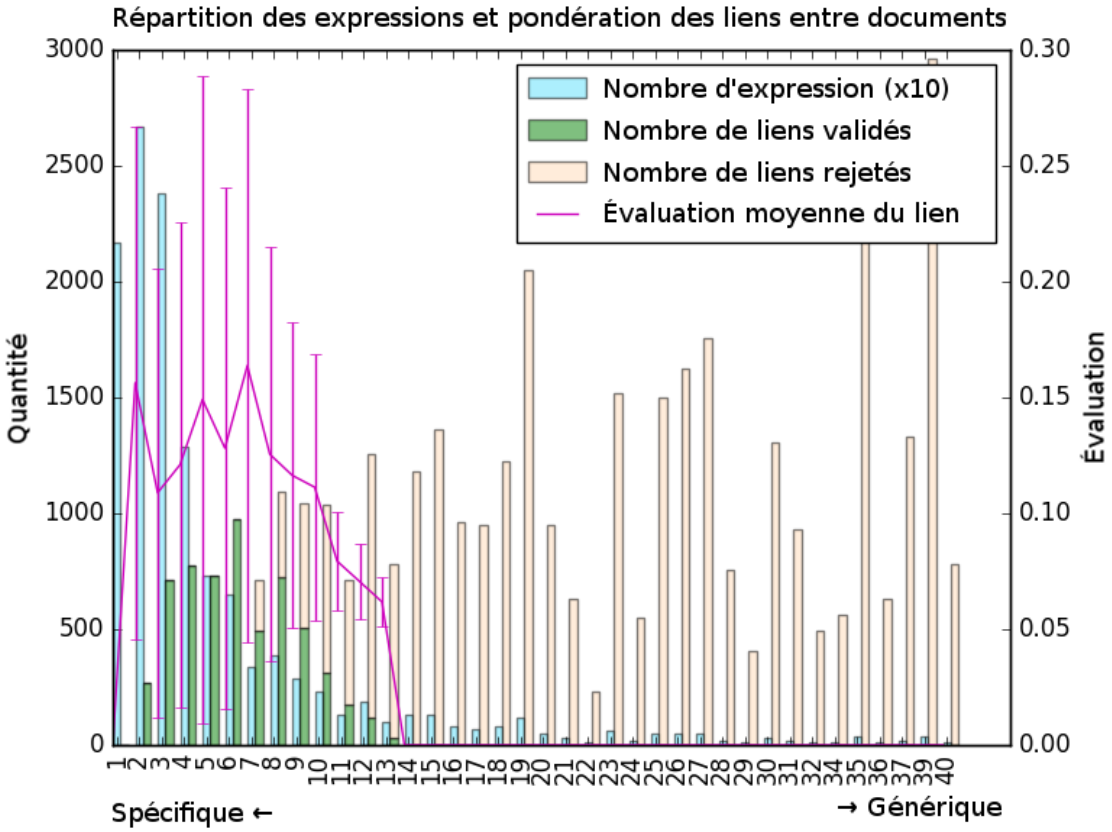


FIGURE 3.2 – En bleu le nombre d'expressions-clés (×10) en fonction du nombre de documents où l'expression apparaît. En vert le nombre de liens (valides) créés par ces expressions, en jaune le nombre de liens rejetés. En rose (ordonnée secondaire) la notation moyenne de ces liens.

Un pic de score moyen (en rose) est observé pour les expressions apparaissant dans 6 entretiens, cela signifie que ces expressions sont plutôt occurrentes, avec une bonne répartition des occurrences parmi les 6 textes concernés. On note aussi l'importante proportion d'expressions présentes dans 2 et surtout 3 documents (barres bleues), associé à une chute des scores des liens (rose). Cela indique un corpus constitué de trios et de duos avec une concentration des occurrences sur un seul des 2 ou 3 entretiens. Le corpus fonctionne donc à deux vitesses : d'une part des sujets spécifiques à un duo ou trio porté par une seule personne, et d'autre part des sujets plus généraux, spécifiques à un groupe de 6 chercheurs environ sans prépondérance d'aucun d'eux.

La pondération des liens est appelée « chaînage » car elle consiste à enchaîner les nœuds du corpus les uns aux autres en supervisant la façon dont les liens sont fabriqués et leur intensité calculée (voir section 5). Dans le cas de notre corpus, le « chaînage » conduirait à créer 324 876 liens spécifiques sans filtrage, le calcul de pondération permet d'écarter les liens les plus faibles. Il reste alors 15 242 liens dans la version épurée. Cet ensemble astronomique de relations ne peut être saisi par

l'entendement humain. Il doit être simplifié par des procédures de filtrage et de visualisation des données graphes. Ceci conduit, par le réglage de différents paramètres de filtrage et de visualisation, à produire des représentations numériques (graphes) qui servent de base à la confrontation avec l'analyse qualitative de la section 3.1.

3 Confrontation des analyses quantitatives et qualitatives

3.1 Cartographie générale du corpus : chimie du solide, électrochimie et automobile

L'utilisation de filtres et de métadonnées permet de visualiser un réseau global de relations entre les entretiens du corpus. L'intensité des liens spécifiques a été sommée pour calculer un lien résultant entre chaque nœud et un seuil de visualisation a été imposé (valeur numérique de 0,3) pour ne faire apparaître que les liens de plus forte intensité. Nous supposons que la figure résultante (figure 3.3) constitue une cartographie mémorielle de la recherche sur les matériaux. Cette hypothèse sera validée si l'interprétation qui en découle est possible et convaincante par rapport à la connaissance historienne du corpus d'archives orales.

La figure 3.3 montre des inhomogénéités dans la position, la taille et les relations réciproques entre nœuds, que nous appelons *clusters*. Un *cluster* désigne un regroupement particulier de nœuds en raison de leurs proximités sémantiques par rapport au reste du corpus. Ce sont des clusters binaires (voir définition en section 4.2) identifiés visuellement sur les représentations graphiques des données quantitatives. La figure peut être interprétée de manière qualitative grâce à la connaissance historienne préalable que nous avons du corpus. L'analyse qualitative suivante fait apparaître deux types de commentaires.

3.1.1 L'analyse sémantique comme discriminant thématique

Le premier commentaire concerne le rapport entre l'histoire de la constitution du corpus et sa représentation numérique. Le corpus a été formé par le regroupement de trois sous-corpus d'archives orales (cf. table 3.1) différenciés par la période de collecte (2000-2003, 2004-2007, 2009-2016), le principal interviewer (B. Bensaude-Vincent pour le premier puis P. Teissier pour les deux suivants) et le thème de l'entretien. De ces trois différences *a priori*, la représentation numérique ne retient, au niveau de filtrage qui est le nôtre, que la dimension thématique. En effet, elle mêle les entretiens du sous-corpus 1 (science et ingénierie des matériaux) et du sous-corpus 2 (chimie du solide) sans qu'il soit possible de distinguer leur origine sans connaissance du corpus [4]. Les témoins correspondant appartiennent à la communauté de « recherche sur les matériaux ». Par rapport à cette partie centrale du corpus, cinq entretiens forment un *cluster* périphérique, isolé en bas à droite de la figure. Tous ont été interviewés sur le même thème spécifique : l'industrie des voitures électriques et à hydrogène en lien avec les réseaux électriques. Quatre d'entre eux (P. Beuzit, J.-P. Lisse, P. Lucchese, S. Vitet) appartiennent au sous-corpus 3, le cinquième (J.-C. Griesemann) au sous-corpus 1. En outre, les trois entretiens qui leurs sont le plus liés concernent la recherche et le développement des batteries et piles à combustible, deux d'entre eux issus du sous-corpus 3 (J.-C. Catonné, J.-F. Fauvarque), le troisième du sous-corpus 1 (P. Barboux).

Cette première confrontation entre analyses quantitative et qualitative montre deux atouts d'*Haruspex*. D'une part, l'outil est particulièrement adapté à l'analyse sémantique de textes puisqu'il parvient à isoler des *clusters* centrés sur des thèmes périphériques par rapport au reste du corpus : industries automobile et électrique pour le *cluster* de cinq nœuds, isolé en bas à droite ; recherche et développement en électrochimie si on rajoute les trois nœuds les plus proches du *cluster* isolé. D'autre part, l'outil semble relativement indépendant des conditions extérieures de l'entretien (date, projet) et de la subjectivité de l'interviewer (formulation des questions, mode d'expression) comme le montre l'intégration des deux premiers sous-corpus réalisés par deux historiens différents ayant des projets différents à des dates différentes.

3.1.2 Analyse de la communauté française de chimie du solide

Laissant de côté le *cluster* périphérique, notre deuxième ensemble de commentaires analyse les trois quarts restant du corpus. Ces 33 entretiens appartiennent à la communauté internationale de « recherche sur les matériaux », mais avec une forte focalisation sur la chimie du solide française : 27 chimistes, 4 physiciens et 2 techniciens, travaillant tous en France.

L'école de recherche de R. Collongues. La partie principale du corpus forme cinq *clusters*. Une première partition distingue les parties droite et gauche de la figure 3.3 par rapport à l'axe vertical reliant D. Vivien et J. Livage. A droite, se trouve les représentants de l'école de recherche initiée par Robert Collongues regroupés en deux *clusters*. Il est important de rappeler ici que R. Collongues, acteur majeur de la formation de la discipline, décédé en 1998, n'a pas été interviewé et n'est donc pas représenté dans le corpus. Son influence peut néanmoins se mesurer au « trou » qu'il laisse en haut à droite de la figure 3.3. Le premier *cluster*, composé de cinq nœuds (A. Kahn, H. Mondange, M. Perez, D. Serreau, D. Vivien), correspond au laboratoire historique de Vitry-sur-Seine puis de l'École de chimie de Paris. Le second *cluster*, composé de trois nœuds (J.-P. Boilot, P. Colomban, J. Théry), quatre si l'on ajoute P. Barboux, correspond à peu près à un groupe de recherche héritier de Collongues établi à l'École polytechnique. Cette école de recherche, cluster diffus autours de l'absence de Collongues, est exploré en figure 3.4. Ce graphe fait ressortir les liens fragiles entre les chercheurs, liens écrasés par la magnitude des clusters voisins. On découvre que les liens

4. Dans une perspective réflexive, on peut noter ici que le second interviewer a réalisé une thèse sous la direction de la première, ce qui, à n'en pas douter, a généré des similitudes de méthodes et de questionnements. Ils ont, en outre, réalisé certains entretiens ensemble.

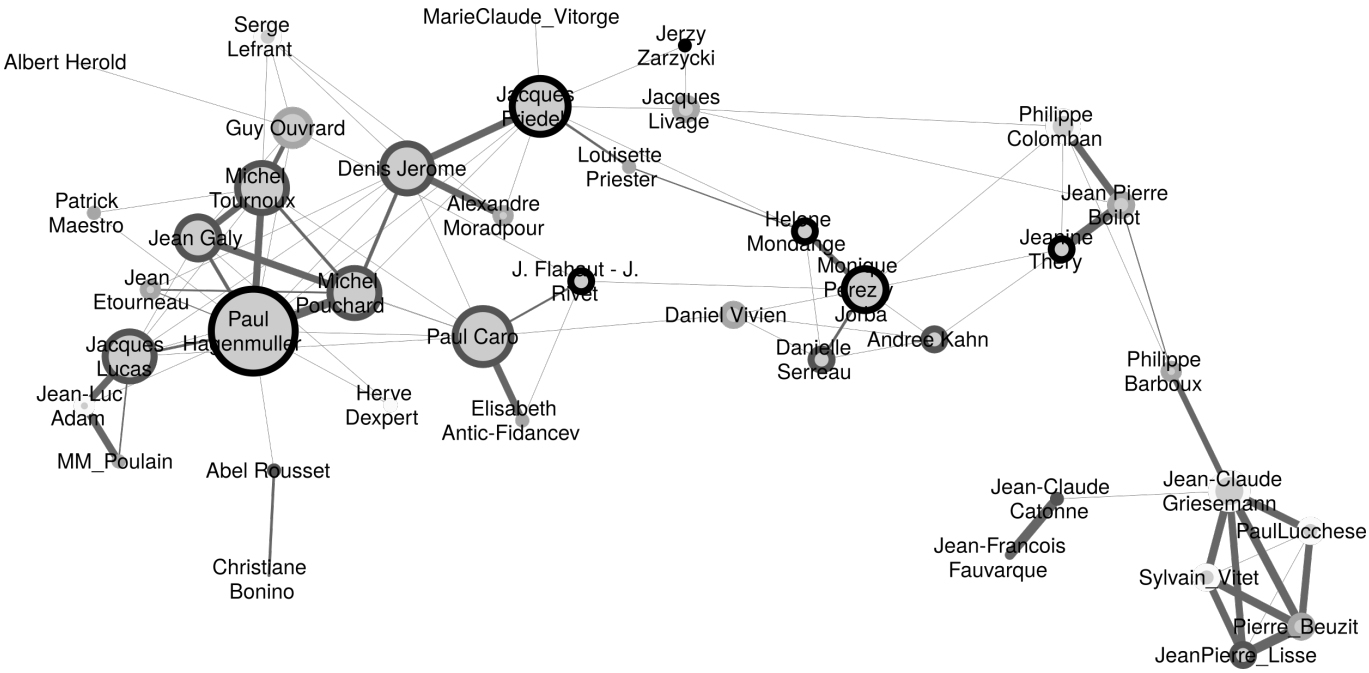


FIGURE 3.3 – Chaque nœud représente un entretien, identifié par le nom du témoin interviewé. Les liens sont la superposition (somme) de tous les liens créés précédemment (1 lien par expression partagée). En dessous d'un seuil de pondération (0.3), les liens ne sont pas affichés, mais continuent d'influencer la forme du graphe. La taille du nœud indique son degré de connectivité : plus le témoignage a de connexions, plus il est volumineux. La couleur des nœuds renseigne sur la date de thèse de l'interviewé : plus les nœuds sont foncés, plus la décennie de soutenance est ancienne (noir dans les années 1950, blanc après 1980).

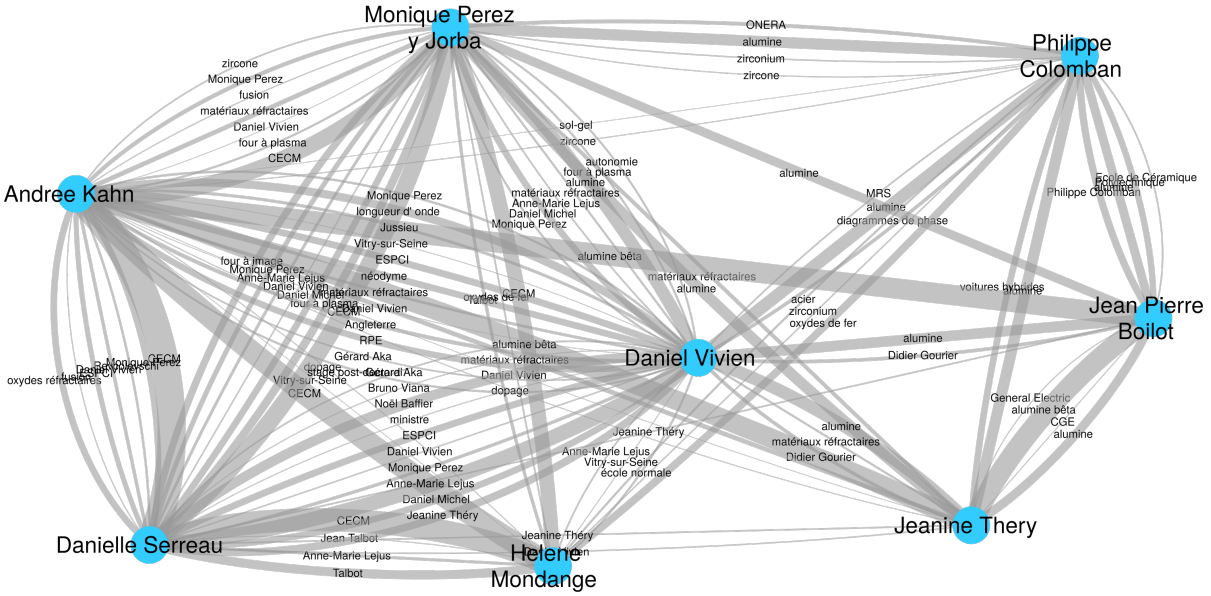


FIGURE 3.4 – Les membres identifiés de l'école de Collongues. Ces 8 membres se divisent en sous-groupes : à gauche 5 nœuds fortement liés (liens sociaux et scientifiques) et à droite, 3 chercheurs périphérique (principalement liés aux autres chercheurs par l'alumine).

de type scientifiques sont plutôt répartis, que l'évocation de l'alumine, voire de l'alumine bêta constitue le principal lien entre des chercheurs périphériques avec le noyau et que ce sont des proximités sociales (nom de personnes et d'institutions) qui marquent le rapprochement différencié des membres du noyau (à gauche). Les 5 nœuds du noyau appartiennent au même laboratoire, ce qui explique cette proximité.

La gauche de la figure 3.3 est plus dense, même si trois *clusters* peuvent être repérés. Le premier d'entre eux est central dans le corpus. Il est composé, en son cœur, de cinq nœuds : P. Hagenmuller, qui constitue le nœud le plus connecté du corpus, et quatre de ses disciples directs (J. Etourneau, J. Galy, M. Pouchard, M. Tournoux). Des lignées peuvent ainsi être tracées d'une génération à la suivante : P. Hagenmuller - M. Tournoux - G. Ouvrard (S. Lefrant étant un collègue physicien d'Ouvrard à Nantes). On retrouve ici, comme dans le cas Collongues, une structuration en école de recherche, qui a essaimé depuis Bordeaux

vers d'autres villes universitaires comme Nantes (M. Tournoux) ou Toulouse (J. Galy). Certains « anciens de Bordeaux » sont faiblement liés au *cluster* central, l'un (H. Dexpert) parce que ses choix thématiques l'ont éloigné de l'école, l'autre (P. Maestro) parce qu'il a fait carrière dans l'industrie des matériaux. Un « héritage indirect » est aussi repérable entre P. Hagenmuller et J. Lucas à Rennes [5], celui-ci étant lui-même lié à son disciple (J.-L. Adam) et plus faiblement à deux chercheurs de son laboratoire (Michel et Marcel Poulain).

Les deux derniers *clusters* se trouvent au milieu de la figure. L'un, en bas, s'organise autour de P. Caro, lui aussi, figure centrale du réseau. Ses relations avec les autres sont peu marquées par des relations de filiation d'école : il est certes lié à son ancienne doctorante (E. Antic-Fidancev), qui reste dans son laboratoire, mais pas du tout à un autre chercheur de son laboratoire (H. Dexpert). Il est davantage lié à des pairs de la même génération que lui (D. Jérôme, J. Flahaut et J. Rivet) avec qui il partage des liens de type épistémique (ex : « four solaire », « isolant »).

Des personnalités influentes n'ayant pas fait école. Contrairement à la structuration par écoles de recherche, déjà mentionnées, l'entourage de Caro pourrait correspondre au *cluster* d'un chercheur influent (académicien et original) qui n'a pas fait école au sens d'une filiation (de thèse) mais qui a structuré le champ au niveau plus épistémique. Le dernier *cluster* identifié rassemble les physiciens du corpus (D. Jérôme, S. Lefrant, L. Priester) autour du physicien J. Friedel, qui jouerait le même rôle structurant que P. Hagenmuller si le corpus contenait plus d'entretiens de physiciens. Il agrège aussi des chimistes qui n'étudiaient pas des cristaux minéraux, ce qui était rare à l'époque concernée : solides amorphes (J. Livage, J. Zarzycki) ou organiques (A. Moradpour, M.-C. Vitorge). D'autres hétérodoxes par leurs positionnements thématiques ou sociologiques (C. Bonino, A. Rousset et A. Hérold), cependant, sont plus liés au *cluster* Hagenmuller qu'au *cluster* des physiciens.

En comparaison avec l'analyse qualitative. Ainsi, la partie centrale de la figure 3.3 confirme deux résultats principaux de la recherche historique préalable Teissier (2014) :

1. une autonomie des communautés de physique et de chimie du solide en France
2. une structuration de la communauté de chimistes du solide polarisée par la forte opposition entre deux mandarins tout puissants (Hagenmuller et Collongues) et clairsemée d'une multitude d'écoles de recherche moins puissantes (Caro, Flahaut, Hérold, Lucas, Rousset, etc.)

Il est intéressant de remarquer que les liens entre les deux écoles de recherche les plus puissantes se fait via les écoles de puissance intermédiaire (Caro, Flahaut) et par les physiciens, qui collaborent avec les chimistes depuis l'extérieur. L'analyse numérique révèle ainsi la mise en connexion des centres d'une communauté scientifique (chimie du solide) par les marges disciplinaires (physique du solide) et les outsiders sociaux au sein de la communauté. La position dominante de Friedel parmi les physiciens interrogés permet d'extrapoler en suggérant une structuration mandarinale comparable en physique du solide. La position centrale du *cluster* Hagenmuller et la connectivité maximale du nœud Hagenmuller s'explique par trois éléments entremêlés. Une domination sociale d'une part : plus de 300 docteurs formés à Bordeaux durant la direction d'Hagenmuller ainsi que des laboratoires héritiers dans tout l'ouest de la France, notamment à Nantes et dans une moindre mesure Rennes. Une contingence historique ensuite : son grand rival, Collongues, ne figure pas sur le réseau car, décédé en 1998, il n'a pas pu être interrogé. Un « effet de sources » enfin : la collecte des témoignages entre 2004 et 2007 ayant été faite de proche en proche à partir du laboratoire Collongues puis Hangenmuller, ces deux écoles de recherche ont été favorisées dans la constitution du corpus.

3.2 Valider le connu et expliquer le surprenant pour des liens localisés

La représentation numérique globale du corpus est venue confirmer certaines conclusions historiques, notamment la structuration sociologique et épistémique de la communauté. Pour raffiner cette approche globale, nous analysons trois liens localisés autour d'un nœud et entre deux nœuds. L'objectif est de montrer que la valeur numérique d'un lien est une réduction pratique mais trompeuse qui cache une grande variété de contenus et de significations.

3.2.1 Le nœud central : Paul Hagenmuller

Nous commençons par étudier un nœud très particulier : le plus gros du corpus, c'est-à-dire le plus connecté aux autres, Paul Hagenmuller. Nous répertorions, dans la table 3.5, la liste des liens les plus forts de ce nœud et les expressions correspondantes. Ceci renseigne sur les thématiques fortes d'un témoin particulier et sur les témoignages les plus significativement liés à lui. Une telle analyse est indispensable pour comprendre le rôle et la place d'un acteur particulier dans le corpus. Le cas Hagenmuller donne deux types d'information.

Des collaborateurs privilégiés. Premièrement, ces liens forts sont souvent dus à des partages avec un ou deux autres témoins d'expressions ayant peu d'occurrences dans le corpus. Par exemple, le lien le plus fort du corpus (poids de 0.606) concerne le « vanadium », un élément chimique que Hagenmuller partage avec J. Galy et M. Pouchard de manière quasi-monopolistique :

5. Les témoignages de P. Hagenmuller et J. Lucas mentionnent l'influence du premier sur le second à Rennes même s'il n'y a pas eu direction de thèse.

ces trois témoignages comptent pour 22 des 23 occurrences du mot dans le corpus. Ceci est encore plus vrai avec l'équipartition monopolistique de « transition métal-isolant » avec D. Jérôme : 3 occurrences chacun, soient les 6 que compte le corpus. S'accaparer une expression-clé avec un ou deux autres témoins fabrique donc des liens de forte intensité.

L'organisation sociale du laboratoire. Deuxièmement, en extrapolant cette première tendance sur un grand nombre de liens, nous comprenons que ce qui fait le caractère central de P. Hagenmuller dans le corpus est son aptitude à savoir utiliser des expressions spécifiques qui ne sont partagées que par quelques spécialistes d'un domaine, de manière aussi précises qu'eux. Ceci lui permet de se lier fortement, et de manière privilégiée, avec beaucoup de témoins. Cette conclusion sémantique peut s'expliquer par l'organisation sociale de son laboratoire à Bordeaux : un immense et riche institut, composé d'équipes de recherche spécialisées sur de nombreux domaines différents, dont le directeur a une vision d'ensemble.

C'est un apport à la fois surprenant et fructueux d'*Haruspex* de suggérer un isomorphisme de l'espace social du laboratoire (connu antérieurement) et de l'espace sémantique de la communauté. Dans cette perspective, un mandarin comme Hagenmuller n'est pas seulement celui dont tout le monde parle et qui forme le plus d'héritiers, ce qui était suggéré par l'analyse historique, mais encore celui qui parle le langage des spécialistes du corpus sans investir lui-même une spécialité.

Mot-clé	Poids	Occ.	in A	in B	Interview associée
vanadium	0.606	23	6	8	Jean Galy
vanadium	0.606	23	6	8	Michel Pouchard
Trombe	0.533	66	4	45	Paul Caro
bronzes de tungstène	0.511	9	4	5	Michel Pouchard
John Goodenough	0.509	19	4	9	Michel Pouchard
Jacques Lucas	0.472	24	4	4	Jean-Luc Adam
fluor	0.441	12	3	6	J. Flahaut - J. Rivet
verres fluorés	0.421	55	2	23	MM_Poulain
Trombe	0.413	66	4	4	J. Flahaut - J. Rivet
transition métal-isolant	0.409	6	3	3	Denis Jerome
verres fluorés	0.408	55	2	19	Jean-Luc Adam
bronzes de vanadium	0.395	22	2	15	Michel Pouchard1
octaèdres	0.375	19	2	12	Michel Pouchard1
théorie des bandes	0.330	17	4	4	Guy Ouvrard

TABLE 3.5 – Extrait (poids > 0.32) des mots-clés les plus significatifs dans leur mise en relation entre Paul Hagenmuller et les autres interviews. La colonne **Occ** indique le nombre d'occurrences total dans le corpus du mot-clé les colonnes **in A** et **in B** indiquent les quantités mesurées dans l'interview d'Hagenmuller et dans l'interview associée.

3.2.2 Un chaînon manquant : Monique Pérez et Jeanine Théry

Le lien entre Monique Pérez y Jorba et Jeanine Théry a été choisi pour cette deuxième analyse locale en raison de son intensité étonnamment faible par rapport à notre présupposé historien. Son intensité atteint à peine le seuil d'affichage alors que ces deux chercheuses de la même génération ont été les deux lieutenantes indéfectibles de R. Collongues pendant quatre décennies. Ainsi, leur position, côte à côte, sur la photographie des premières années du laboratoire Collongues à la fin des années 1950 (figure 3.5) tranche avec leur faible connexion sur la cartographie générale du corpus (figure 3.3). Au contraire, chacune structure un *cluster* de l'école de recherche Collongues : M. Pérez au laboratoire historique ; J. Théry pour le laboratoire héritier de Polytechnique.

Comment interpréter une telle configuration ? L'analyse qualitative des témoignages avait déjà suggéré une rivalité profonde et durable entre les deux lieutenantes du professeur. Une telle rivalité leur a fait suivre durant leur carrière des chemins parallèles évitant autant que possible les croisements épistémiques et, longtemps après durant leur entretien (2004), leur a fait éviter de parler l'une des sujets de l'autre. Ces choix diminuent d'autant la probabilité de thèmes communs, d'événements partagés, de rencontres interpersonnelles malgré l'appartenance à un même laboratoire durant leur carrière. Cette interprétation qualitative peut être complétée par une liste des expressions extraites communes aux deux interviews présentée dans la table 3.6. Les mots-clés communs les plus importants (alumine, réfractaires, Vitry-sur-Seine) sont des expressions courantes de l'école Collongues, ce qui explique la faible discrimination que cela implique pour lier les entretiens de J. Théry et M. Pérez.

3.2.3 Liens de génération et rapports de genre

Le troisième lien concerne les acteurs les plus anciens, repérés par un cerclage noir sur la figure 3.3, qui appartiennent à la génération qui a institutionnalisé la chimie du solide à partir des années 1950. Il montre une différence significative en terme de genre. Les hommes, qu'ils soient de puissants mandarins (J. Friedel, P. Hagenmuller) ou des professeurs moins influents (J. Flahaut, A. Hérold) sont éloignés les uns des autres dans le réseau. Leur ancienneté dans le champ social peut alors être vue comme cause de peuplement de leur environnement par des chercheurs plus jeunes, qu'ils ont notamment dirigés en thèse. Ils se trouvent ainsi éloignés les uns des autres. En revanche, les femmes restent proches les unes des autres qu'elles soient rivales (M.

FIGURE 3.5 – Photographie de Robert Collongues et ses six « maîtresses de conférences », nov. 1959 (Archives personnelles de H. Mondange)

Mot-clé	Poids	Occ	in A	in B
alumine	0.414	30	4	4
matériaux réfractaires	0.197	23	3	3
Daniel Vivien	0.093	28	2	2
Vitry-sur-Seine	0.062	50	3	6
Perez	0.052	11	1	1
ferrites	0.052	10	1	1

TABLE 3.6 – Liste exhaustive des expressions clés liant Monique Pérez et Jeanine Théry. La colonne **Occ** indique le nombre d'occurrences total dans le corpus du mot-clé. Les colonnes **in A** et **in B** indiquent les quantités mesurées dans l'interview de J. Théry et de M. Pérez. Notons leur faible part dans l'ensemble des occurrences de ces expressions. Elles ne partagent rien de spécifique.

Pérez, J. Théry) ou pas (H. Mondange, M. Pérez). Cette fois-ci, l'ancienneté ne structure pas l'environnement social. Ceci est d'autant plus étonnant que si H. Mondange n'a pas occupé de position de pouvoir, M. Pérez et J. Théry furent chefs d'équipe et encadrèrent les thèses de doctorat du laboratoire Collongues. Ceci n'a, semble-t-il, pas suffi à garder les jeunes générations dans une relation de dépendance sociologique ou épistémique.

Le contexte des années 1950 et 1960, en effet, n'était pas facile pour les femmes, qui « n'étaient pas prises au sérieux » comme le rappelle H. Mondange. Les expressions « maîtresses de conférences » pour désigner, à l'époque, les collaboratrices de Collongues (titre de la figure 3.5) et « pères fondateurs » pour désigner, encore aujourd'hui, Hagenmuller et Collongues montrent la position différenciée des hommes et des femmes dans le champ social des sciences de la deuxième moitié du vingtième siècle. Il y aurait donc là matière à approfondissement pour formaliser les articulations entre division du travail, liens générationnels et rapports de genre dans le champ scientifique.

3.3 Approche heuristique des humanités numériques

Jusqu'à présent, nous avons commenté une représentation numérique du corpus (figure 3.3) grâce à des tables explicitant des listes d'expressions et des intensités de lien et à notre connaissance historique. Tel une « porte d'entrée », l'outil numérique « ouvrait » le corpus selon une perspective différente. Nous inversons ici la démarche en générant des représentations quantitatives à partir d'un questionnement *a posteriori* afin de tester les possibilités heuristiques de l'analyse numérique pour les historiens. Le travail historique a montré que la communauté française de chimie du solide s'est construite sur trois éléments forts : une rivalité identitaire, un apport (instrumental et théorique) de la physique, une structuration par les financements industriels Teissier (2014). Nous aborderons par la suite les deux premiers éléments de construction de la communauté, identitaire et disciplinaire, en essayant d'évaluer l'apport de la méthode quantitative dans chacun des deux cas.

3.3.1 Interroger la structuration identitaire d'une communauté disciplinaire

L'émergence d'une communauté scientifique est induite par la construction d'une identité nouvelle, basée sur des organisations sociales et des perceptions psychologiques. Dans le cas de la chimie du solide en France, la mémoire collective est polarisée

autour de l'opposition entre deux mandarins, Hagenmuller et Collongues, qui sont encore considérés aujourd'hui par de nombreux chimistes du solide comme « les deux pères fondateurs » de la discipline. Ce trait constitutif de la mémoire collective peut-il être visualisé ? Nous interrogeons le corpus en demandant à *Haruspex* de dénombrer, pour chaque entretien, les occurrences des expressions « Hagenmuller » et « Collongues » et nous visualisons la répartition obtenue sur la figure 3.6.

Celle-ci met en évidence un équilibre des populations entre les deux mandarins, ce qui confirme l'importance de cette polarité dans la mémoire collective. Citer les deux noms est donc un signe d'appartenance à la communauté. Le noyau dur de chacun des deux camps (qui cite davantage l'un que l'autre) se situe près du mandarin, mais entre les deux : A. Kahn, D. Serreau pour Collongues ; M. Pouchard, G. Ouvrard pour Hagenmuller. Les nœuds proches d'un mandarin mais tournés vers l'extérieur (qui n'en citent qu'un) sont soit des étrangers à la communauté (J.-F. Fauvarque, D. Jérôme), soit des marginaux de la communauté du point de vue de la génération (H. Mondange) ou des thèmes de recherche (J-L. Adam, J. Zarzycki). La ligne verticale centrale permet de visualiser les représentants des autres écoles de recherche (P. Caro, J. Flahaut, A. Rousset, M.-C. Vitorge), ne prenant position ni pour l'un mais citant également les deux. Ainsi, la figure 3.6, issue d'un questionnement, raffine l'analyse globale de la figure 3.3, confirme l'interprétation historique par des données quantitatives et ouvre des perspectives d'interprétation nouvelle sur les raisons de certains positionnements. D'autres dénombrements pourraient être tentés avec d'autres ensembles de noms pour visualiser d'autres configurations.

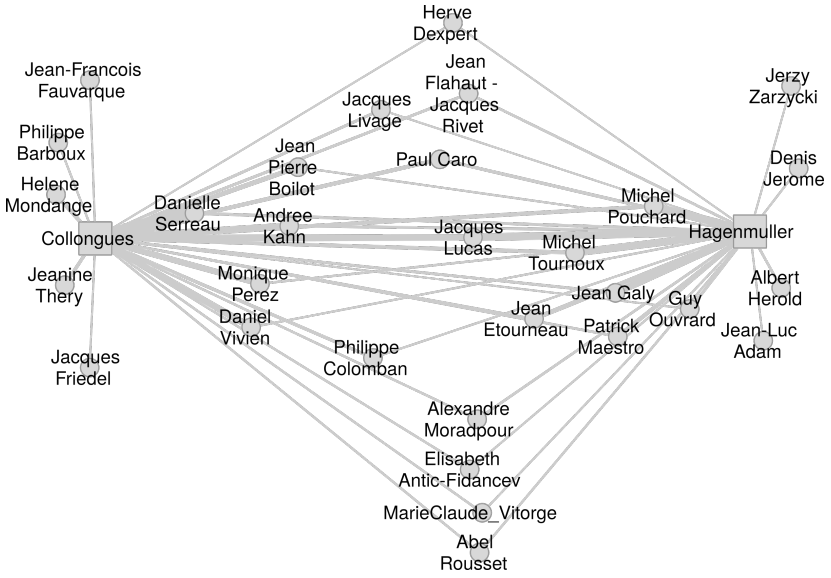


FIGURE 3.6 – Répartition des entretiens en fonction du nombre d'occurrences à Collongues ou à Hagenmuller. Certains ne citent pas l'un, d'autres pas l'autre. La position est relative aux nombres d'occurrences de Hagenmuller ou Collongues représentés en 2 pôles (nœuds de forme carrée). Ces pôles sont des nœuds artificiels. L'entretien avec P. Hagenmuller a été retiré de cette analyse.

3.3.2 Sélectionner les registres sémantiques de la mémoire collective

Une communauté scientifique comme la chimie du solide se définit aussi par des thèmes de recherche (étude de la relation structure-propriété), des outils matériels (diffraction des rayons X) et théoriques (théorie des bandes), des objets d'étude (cristaux inorganiques). Nous construisons une nouvelle représentation numérique du corpus dans sa globalité en sélectionnant uniquement les mots-clés liés au registre épistémique (disciplines, thèmes, outils, objets, etc.). Ceci nécessite une supervision manuelle des 1372 expressions-clés (étape d'enrichissement des expressions). Bien que laborieuse, cette étape ne peut probablement pas être automatisée à cause de potentielles erreurs d'indexation. La figure 3.7 fait ainsi apparaître une sous-structure épistémique du corpus global. Ont été écartées les relations entre nœuds des registres identitaires (noms propres notamment) et institutionnels (organisation des laboratoires, interactions industrielles). La cartographie épistémique rebat les cartes par rapport à la représentation globale du corpus en figure 3.3. Elle ne structure plus le corpus selon la polarisation gauche/droite des deux mandarins (Hagenmuller/Collongues) et dissout le *cluster* périphérique centré sur les industries automobile et électrique. En d'autres termes, en surprenant l'historien trop habitué à son corpus, elle le pousse à modifier ses habitudes de pensée.

Cartographie épistémique. La cartographie épistémique fabrique un réseau plus dense, plus régulier et plus éclaté dans la distribution des nœuds que la cartographie globale. Son interprétation *a posteriori* permet de dégager trois ensembles d'observations. Premièrement, il est possible de retrouver des regroupements par école de recherche, quoique de manière moins saillante. Ainsi, en haut à droite, un *cluster* Collongues est regroupé par l'évocation de l'alumine, composé-phare du laboratoire des années 1960 et 1970 : J.-P. Boilot, P. Colomban, A. Kahn, M. Perez, J. Thery, D. Vivien. Mais, trois membres de l'école ont été redistribués ailleurs : D. Serreau, secrétaire du laboratoire, et P. Barboux, chercheur héritier, ont des intensités de lien trop faibles pour être liés au réseau (en bas au centre). De même, H. Mondange, qui a fait sa carrière à l'ombre de Collongues et de leur maître

à tous deux, Georges Chaudron (chimie métallurgique), se trouve repositionnée dans un *cluster* métallurgie, entre physique et chimie, à gauche de la figure (J.-C. Catonné, J. Friedel, M. Pouchard, L. Priester). Les relations de type école ou discipline sont amoindris devant les objets et thèmes de recherche : « acier », « dislocation » et « fer » pour le *cluster* métallurgie. Dans le même registre, apparaissent aussi les thèmes marginaux de la chimie du solide. Lié au *cluster* métallurgie, on repère le *cluster* des solides faiblement organisés (grains, morphologie, sels, surfaces) par rapport à la norme dominante des cristaux (C. Bonino, J.-C. Catonné, J.-F. Fauvarque, S. Lefrant, A. Rousset, J. Zarzycki). L'exemple le plus emblématique, présentant les liens les plus forts en triangle, à droite (J.-L. Adam, J. Lucas, MM. Poulain), est le laboratoire rennais d'étude des verres à base d'éléments autres que l'oxygène (« fluorés », « chalcogénures », « lanthanides »). Ce triangle est relié à plusieurs chimistes du solide étudiant les cristaux à base des mêmes éléments (P. Caro, J. Flahaut et J. Rivet, G. Ouvrard, M. Tournoux).

Porosité entre industrie et académie. Deuxièmement, si les *clusters* par école de recherche sont démembrés ou affaiblis au profit des regroupements épistémiques (objets, thèmes), les modifications de relation entre sphères académiques et industrielles sont plus surprenantes encore : le *cluster* périphérique (automobile et énergie) est reconfiguré autour de deux « chaînes » distinctes, c'est-à-dire des enchaînements de trois ou quatre nœuds fortement liés par paires. La première chaîne est visible en bas à droite : J.-C. Griesemann (pile à combustible chez Renault) est lié à P. Lucchese (énergie renouvelable au CEA), qui est lié à P. Maestro (développement de pigments chez Rhodia), lui-même associé à P. Caro (luminescence fondamentale au CNRS). La seconde chaîne part de P. Beuzit (manager chez Renault) puis J.-P. Lisse (chimiste des piles à combustible chez Citroën) et se termine avec M. Pouchard (chimie théorique des matériaux électrochimiques à l'université). Ces deux chaînes parcourent toutes deux un chemin géographique de la périphérie vers le cœur de la cartographie et un chemin professionnel du commercial et industriel vers le fondamental. La transformation des *clusters* en chaînes confirme la porosité des sphères industrielles et académiques dans la mise en place de systèmes d'innovation sur les matériaux. Elle suggère, dans le même temps, des chemins pour étudier ces interactions.

Relations épistémiques vs. homogénéité sociale. Abordons dans un troisième temps ce qui constitue la partie la plus centrale de la figure 3.7 et la plus épineuse à interpréter aussi : un *cluster* dense formé par 5 nœuds au centre (P. Caro, J. Galy, P. Hagenmuller, M. Pouchard, D. Jérôme). Il ne s'agit plus d'un effet d'école comme pour la figure 3.3 puisque le nœud P. Hagenmuller est moins dominant qu'auparavant et qu'il est peu ou pas lié à certains de ses héritiers directs (J. Etourneau, M. Tournoux) ou indirects (J. Lucas, G. Ouvrard). De quoi s'agit-il alors ? Trois observations préliminaires avant de répondre : la présence notable d'un chimiste indépendant des deux écoles dominantes (P. Caro), lui-même lié à trois chimistes plus âgés que lui (J. Flahaut, A. Hérold, M. Tournoux) ; la présence non moins notable d'un physicien du solide (D. Jérôme de la même génération que M. Pouchard), lui-même lié à un second physicien (son mentor J. Friedel) et un chimiste de son laboratoire d'Orsay (A. Moradpour) ; enfin, des liens entre les nœuds basés sur des références à des structures cristallines (« bronzes », « cuprates »), des éléments chimiques (« vanadium »), des propriétés physiques (« conducteur », « isolant », « raies », « spectres ») et des concepts théoriques (« spin », « transition métal-isolant »). Le *cluster* central fait donc surgir des relations épistémiques fortes liant des hétérogénéités sociales (physiciens versus chimistes, mandarins versus outsiders). Il rend compte d'un ensemble épistémique plus large qu'une école de recherche mais différent d'une discipline comme la chimie du solide. Ce faisant, l'analyse numérique pourrait donner lieu à des typologies nouvelles en terme d'organisation sociale des sciences.

Conclusion sur la représentation épistémique. Au terme de cette analyse, la représentation épistémique de la figure 3.7 permet de dégager deux perspectives intéressantes. Tout d'abord, au niveau réflexif, elle sert de garde-fou méthodologique : une image inattendue qui suscite la surprise refrène les généralisations trop hâtives. Ensuite, l'image épistémique relativise le rôle de la cellule sociale de la communauté française des chimistes du solide : l'école de recherche et son mandarin, qui structuraient le corpus global sur la figure 3.3. Elle fait donc plus de place aux interactions hétérogènes (université/industrie, physique/chimie, outsiders/mandarins), qui ont pu être sous-estimées dans la mémoire collective et/ou dans l'interprétation historique. Ainsi, sélectionner le registre épistémique dans les mots-clés pourrait permettre de visualiser des épistémologies multiples par école de recherche, par discipline ou sous d'autres formes à déterminer. Dans la perspective d'un élargissement du corpus, les cartographies épistémiques permettraient de s'affranchir des spécificités disciplinaires pour identifier et caractériser plus largement les aspects épistémiques de la recherche sur les matériaux dans le cas de la France (et ensuite au niveau international) et ainsi d'affiner la cartographie de la mémoire collective sur les matériaux.

L'une des pistes prometteuses touche à la définition de régimes de production des savoirs Pestre (1997), qui sont définis par la manière dont les aspects économiques, organisationnels et épistémiques sont liés pour produire une connaissance donnée. Des pistes de réflexion peuvent ainsi être esquissées par l'observation de la représentation. Ainsi, de manière centrale, apparaissent des clusters plus ou moins forts indiquant des complexes industriels bien reconnaissables (automobile, chimie, énergie, métallurgie, télécommunication), dont le rôle est primordial dans l'organisation de la recherche en matériaux. De même, sur les marges (en haut à droite), apparaît un cluster quadrilatère reliant Caro, Flahaut, Pérez et Rousset autour du concept d'atelier, c'est-à-dire un espace de savoir-faire, intermédiaire entre laboratoire et centre R&D. La formation de tels clusters invite à des études plus approfondies des systèmes économico-techniques esquissés, des dynamiques d'innovation ou au contraire des inerties.

Une cartographie des relations entre les acteurs sur un autre plan, celui de l'économie, de l'industrie et de la gestion fait fait ressortir une autre structuration de la communauté.

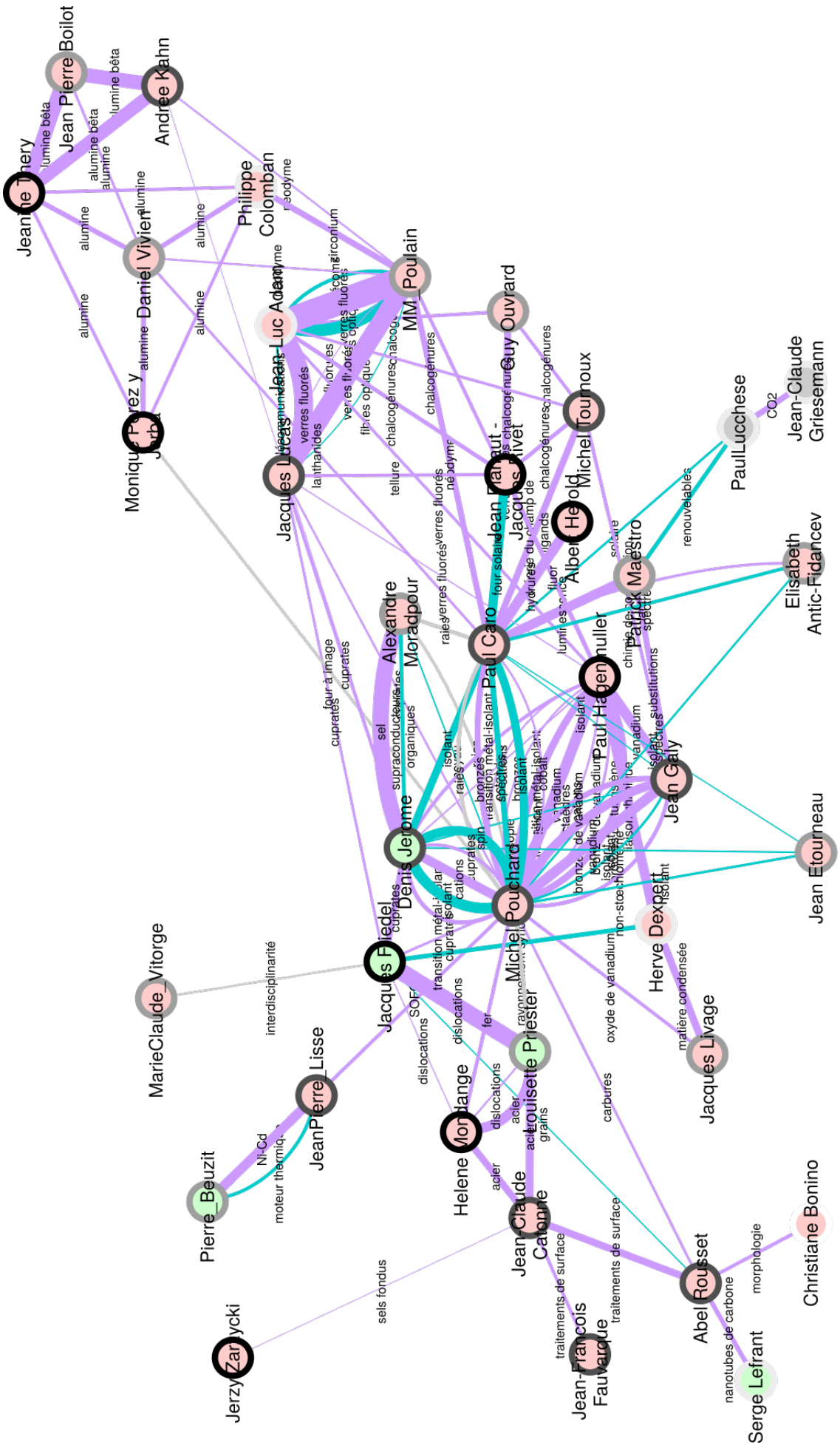


FIGURE 3.7 – Interviewés évoquant des sujets scientifiques : physique, chimie ou indifférencié (sciences). En dessous d'un seuil de pondération (0.32), les liens ne sont pas affichés. Les nœuds roses sont des chimistes, les verts des physiciens, en gris les autres. Les liens bleus sont de physique, les liens roses sont de chimie les jaunes sont indifférenciés (physique/chimie). Les cerclages des nœuds sont fonction de la date de thèse : les plus anciens sont plus foncés.

3.3.3 D'autres vues du mêmes univers : les relations politiques et industrielles

Ayant introduit la forte importance des proximités sociales et économiques dans une cartographique globale (figure 3.3), au détriment de l'expression des liens épistémiques qui se révèlent uniquement derrière un filtre (figure 3.7), nous nous intéressons maintenant aux proximités industrielles entre les entretiens du corpus (figure 3.8). Cette représentation du corpus souligne la

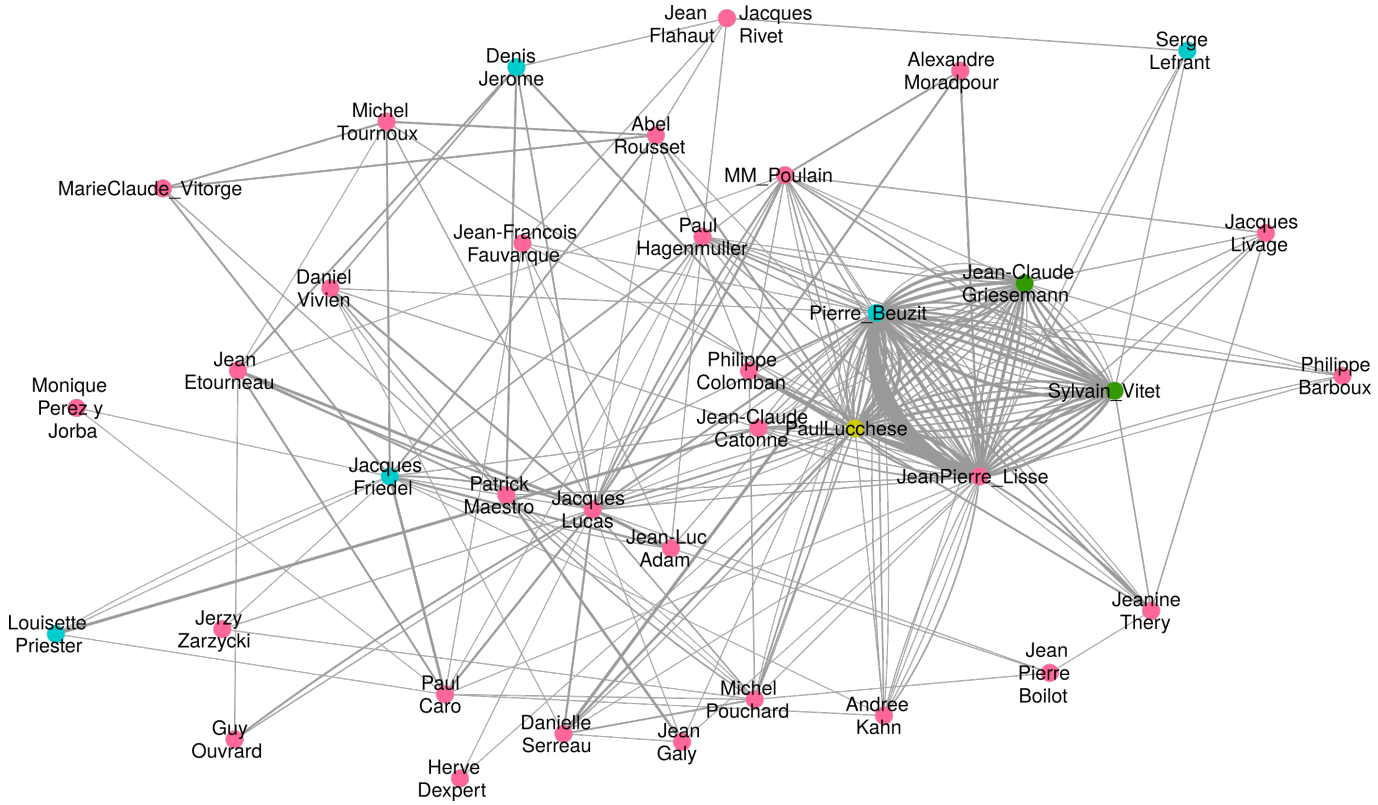


FIGURE 3.8 – Graphe des acteurs liés par des proximités d'ordre industriel. Les nœuds verts ou jaunes sont les industriels et ingénieurs, les roses les chimistes, les bleus les physiciens

structuration d'un cluster déjà identifié : celui des industriels (en bas à droite de la figure 3.3). Parmi les proches de ce cluster sur la figure 3.8 (Jean-Claude Catonne, Philippe Colomban, Jeanine Théry, Philippe Barboux, Jacques Livage), on retrouve certains proches précédemment analysés. Des différences notables sont intéressantes : Philippe Colomban et Jacques Livages sont très distants du groupe des industriels sur la figure 3.3, leur relation aux autres est davantage d'ordre épistémiques. Philippe Barboux, totalement mis à l'écart des relations épistémiques (figure 3.7) s'avère être lié aux industriels d'une part mais également aux chimistes d'autre part ; il conviendrait d'investiguer ces liens. La proximité de Jeanine Théry, avec ce cluster d'industriels ne se traduira que par une vague influence sur le graphe général, davantage polarisée par ses recherches sur l'alumine avec Boilot et Kahn, objet de recherche caractéristique de l'école de Collongues.

Il est alors possible de s'intéresser au détail des liens unissant le noyau d'ingénieurs et d'industriels. La figure 3.9 montre que ces liens sont essentiellement des groupes automobile. Quelques proches collaborateurs de ce cluster central, mais clairement hors de la zone d'influence, évoquent également l'industrie automobile, c'est le cas de Jean-Claude Catonne. Une autre représentation intéressante est celle des liens (toujours d'ordres industriels et politiques) régissant les relations entre les non-industriels, c'est-à-dire à la figure 3.8 sans les 5 nœuds du cluster central identifié précédemment. La figure 3.10 permet d'identifier que les relations de type industrielles entre les non-industriels sont très différentes de celles composant le noyau exclu. Les groupes automobiles ont perdu en influence au profit des grands groupes de chimie : Rhodia, Péchiney, Rhône Poulenc, Saint Gobain ainsi que le monde de la pharmacie, ou des grands acteurs de l'électronique : Philips, Bell, Alcatel, Thomson, ESA. Ce changement de nature et d'intensité laisse néanmoins percevoir la relativement forte dimension industrielle de Jacques Lucas et dans une moindre mesure celle de Jacques Friedel. À l'inverse, Philippe Barboux, précédemment identifié à plusieurs reprises comme proche des milieux industriels se retrouve isolé. Cet isolement traduit une forte dépendance à l'égard du milieu automobile et un fort décalage avec les (quelques) préoccupations industrielles des autres chercheurs académiques.

Cette dernière analyse moins aboutie ici sur le plan historique met en évidence des pistes de recherches très ciblées, n'ayant pas été relevées par les travaux d'analyse qualitative précédents (thèse, article, livre) sur la structuration de la discipline. Les pistes sont la place de Jacques Lucas sur le plan industriel dans la communauté académique, le rôle pivot de Philippe Barboux, fortement dépendant du secteur automobile sur le plan industriel, faiblement intégré à la communauté scientifique sur le plan épistémique et lien crucial entre les communautés académique et industrielle sur un plan encore non-identifié.

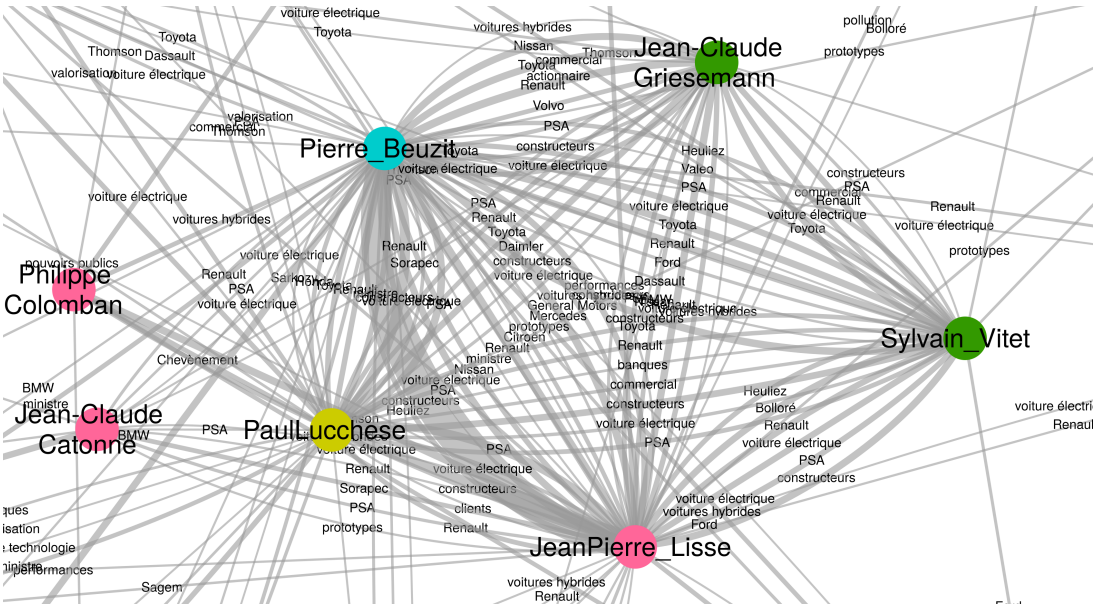


FIGURE 3.9 – Zoom des interactions de type industriel entre les acteurs centraux du graphe figure 3.8

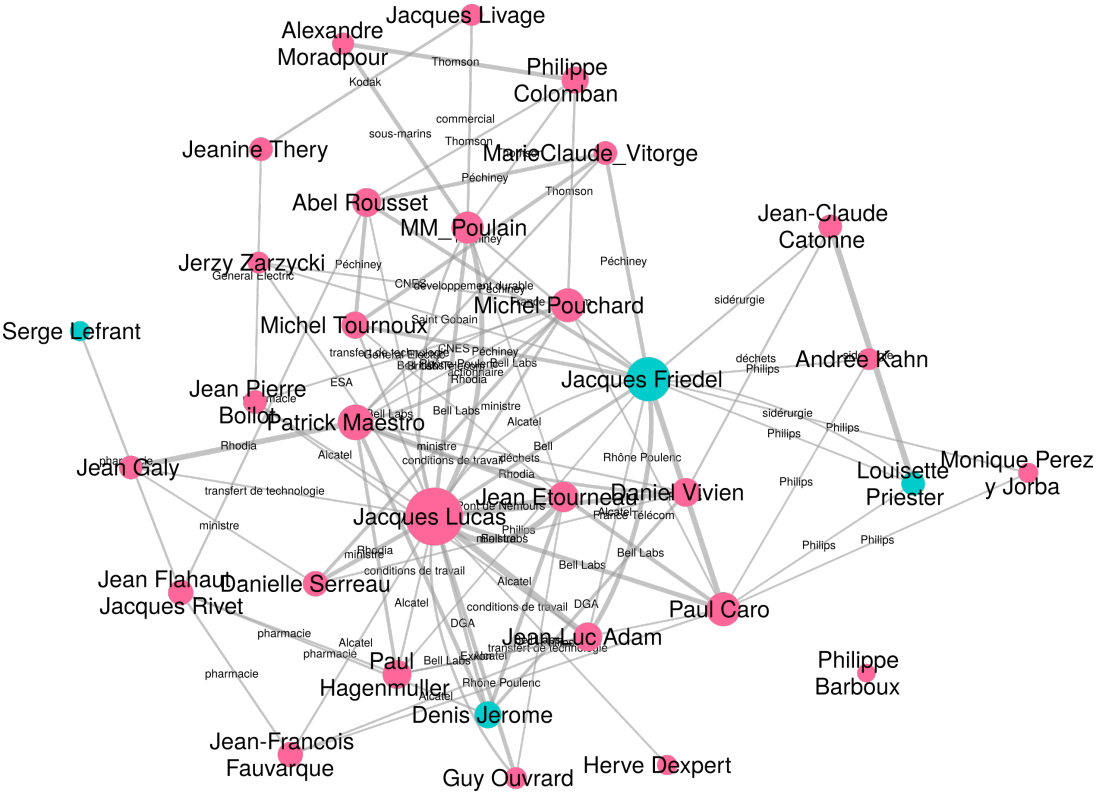


FIGURE 3.10 – Graphe représentant les sujet industriels partagés entre les scientifiques académiques

4 Conclusion et perspectives

La conclusion reprend les trois perspectives présentées dans le résumé : pragmatique, heuristique et réflexive.

Tout d'abord, la dimension pragmatique concerne la fiabilité de la méthode numérique d'analyse du corpus. Cette approche n'a pas donné lieu à des résultats aberrants. Au contraire, la représentation globale du corpus de la figure 3.3, pour laquelle la seule intervention a consisté à fixer des paliers de visualisation, a pu être interprétée de manière conforme à notre connaissance qualitative du corpus et historique des communautés scientifiques correspondantes. Ceci est aussi le cas pour l'analyse spécifique de la sous-structure épistémique de la figure 3.7. Les résultats des analyses de différentes vues du corpus sont cohérentes entre elles, c'est le cas de l'identification du cluster d'industriels et des liens qu'il tisse avec les académiques. Notre méthode numérique semble donc adaptée à l'analyse de corpus textuels de sciences humaines et sociales. Ceci ne signifie pas pour autant que toute représentation puisse trouver une signification, que toute interprétation puisse trouver une visualisation ou que toute interprétation

soit univoque.

Ensuite, notre méthode peut ouvrir des perspectives de recherche aux historiens. L'heuristique se joue au moins à trois niveaux. Premièrement, l'interprétation de relations *surprenantes* permet de braquer le regard sur un angle mort ou d'ouvrir une voix non explorée. La possibilité de dé-corréler les registres de langage, organisant les expressions en catégories (disciplinaires, identitaires ou organisationnelles) est une voie intéressante pour faire apparaître des relations insoupçonnées : *clusters*, chaînes, paires de nœuds et nœuds isolés. Deuxièmement, le tracé de nouvelles représentations suite à un questionnement historique peut confirmer quantitativement des résultats qualitatifs. Le clivage de la mémoire collective entre deux « pères fondateurs » (figure 3.6) a fourni un cas d'étude satisfaisant. Le filtrage des mots-clés par des registres sémantiques est prometteur, notre essai pour le registre épistémique s'étant avéré concluant. Troisièmement, une piste plus porteuse encore nous semble être la comparaison entre la structure globale du corpus (figure 3.3) et l'une de ses sous-structures, notamment épistémique (figure 3.7). Ceci pourrait modifier l'usage et la signification de l'outil numérique *Haruspex* : plus que dans l'interprétation d'une image statique, l'heuristique pourrait se trouver plus fondamentalement dans l'écart entre deux images numériques, c'est-à-dire dans les relations entre différentes structures mémorielles.

Enfin, notre pratique réflexive détaillée en section 5, nous a appris que l'interaction entre le numérique et les humanités est d'autant plus efficace que numériciens et historiens peuvent dialoguer librement et à égalité. De telles interactions interdisciplinaires, constructives et critiques, façonnent d'ailleurs le meilleur garde-fou contre de possibles débordements d'ordre numérique (car à peu près n'importe quoi peut être visualisé) ou historique (car à peu près n'importe quoi peut être expliqué). Cette analyse marque une étape dans le programme de recherche consacré aux archives orales concernant la « recherche sur les matériaux » depuis les années 1940. Il était indispensable, pour développer *Haruspex* et tester sa fiabilité sur les archives orales, de choisir un sous-corpus bien connu : la chimie du solide.

À mesure que l'analyse des archives orales sera élargie à d'autres domaines que la chimie du solide, la connaissance qualitative des sous-corpus s'amoindrira ou sera fragmentée, chaque sous-corpus ayant été constitué par des historiens différents. Or, nous avons compris, avec ce premier cas d'étude, à quel point une appréhension intime des textes mémoriels et des méthodes numériques était déterminante. La cartographie de la mémoire collective sur les matériaux ne se fera donc pas sans mal. Elle ne se fera pas, quoiqu'il en soi, sans un travail collectif, empirique, durable et interdisciplinaire. Loin des standards des humanités numériques et des méthodes de cartographie de *big data*, *Haruspex* esquisse une voie étroite, modeste certes, mais résolument humaine et stimulante.

Chapitre 4

Autres applications : corpus de textes et patrimoine

> « C'est l'hypothèse qu'émet l'atlas : qu'un lien existe entre ce qui apparemment diffère au plus haut point. »
>
> Georges Didi-Huberman
> à propos de l'Atlas Mnémosyne
> d'Aby Warburg

Contents

En plus du corpus d'histoire de la chimie du solide, *Haruspex* a permis d'étudier d'autres corpus de textes. Ces études font le lien avec le monde du patrimoine. Chaque exemple choisi ici met en avant un type d'usage différent.

1 Intégration continue de littérature scientifique aux données du projet *Nantes1900*

Cette étude montre comment *Haruspex* peut créer des relations entre des documents avec des regards croisés sur des objets communs. Il s'agit de documents complexes divisés en parties. Nous nous intéressons ici aux relations inter et intra-documents, ainsi qu'aux relations que ces documents peuvent entretenir avec des sources structurées différemment (base de données de fiches historiques liées).

1.1 Présentation de l'expérience

1.1.1 Le corpus

Il s'agit de « littérature grise » : 2 mémoires de recherches de master2 en histoire, divisés en parties et sous-parties au moins. Ces mémoires présentent une quantité de mots similaires et sont écrit par des auteurs différents. Ces mémoires traitent de sujets connexes. La base de données *Nantes1900* traite également de ces sujets parmi beaucoup d'autres. Cette expérience n'a pas fait l'objet d'analyse historique.

Les 2 mémoires. Le premier mémoire étudie l'histoire d'une parcelle industrielle sur l'île de Nantes entre son élaboration (1860) et son état actuel (2014). L'approche revendiquée est celle d'une archéologie industrielle. Axé sur l'histoire des techniques, ce mémoire s'intéresse aux évolutions architecturales, sociales et technologiques d'une usine. La parcelle a connu différentes usages : fonderie, construction mécanique, chauffages, pompes, etc. Nous désignerons cette source de « parcelle Voruz », du nom du premier industriel à s'y installer.
Le second mémoire étudie l'histoire d'une entreprise de construction navale tentaculaire qui a occupé la parcelle pendant une grande partie de son histoire (1909-1978). Cette entreprise a également occupé de nombreuses autres parcelles pendant la période considérée. Une activité unique (moteurs, construction mécanique) concernait la parcelle. Nous désignerons cette source par « ACB », état principal de l'acronyme de l'entreprise étudiée.

La base de données *Nantes1900*. Le projet *Nantes1900* est un projet à l'initiative du musée d'histoire de la ville de Nantes, pour valoriser un objet monumental de sa collection : la maquette du port industriel de la ville de Nantes au début du siècle (1903-1915 environ). Ce projet a su mobiliser de nombreux conservateurs, chercheurs et étudiant pour — entre autre — constituer un ensemble de « fiches » (équivalent de nos *pages*) qui documentent un objet de musée complexe. De nombreux lieux industriels sont indiqués et documentés par un texte de médiation concis. D'autres textes de médiations ne concernent aucun lieu mais documentent un élément abstrait de l'histoire du port industriel (par exemple « techniques de construction navale »). Ces fiches sont liées entre elles lorsqu'elles partagent un mot-clé. Cet étiquetage par mot-clé a été réalisé à la main, à partir d'une liste obtenue par consensus.

1.1.2 Objectifs et méthodes

Objectif. Le premier objectif de cette étude est de mettre en évidence les intersections de ces mémoires. Un objectif secondaire consiste à montrer la prévalence du contenu sur le style dans l'analyse par *Haruspex* ; cette prévalence se traduirait par quelques parties fortement liées indifféremment de l'auteur.

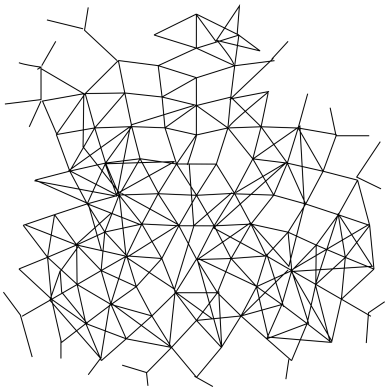
Le second objectif est de montrer qu'il est possible d'augmenter un réseau de *pages* établi, celui de la base de données *Nantes1900* (qui compte 381 *pages* dans notre version) par l'apport (potentiellement continu) de nouvelles productions scientifiques via *Haruspex*. La figure figure 4.1 illustre l'idée de s'accrocher à certains nœuds d'un réseau pré-existant pour le densifier localement. Dans notre cas il s'agirait de densifier le réseau de *Nantes1900* qui traite de l'ensemble du port par les 2 mémoires qui traitent conjointement d'un ensemble restreint de ce port, sur une période de temps plus large.

Méthode. Nous traitons d'abord les 2 textes de recherche à part pour identifier leurs potentielles interactions, jonctions. Puis nous ajoutons le corpus de *Nantes1900*, qui a une structure très différente (taille des textes et type d'écriture). *Haruspex* est réutilisé sur l'ensemble (les 3 ensembles). Nous n'utilisons donc pas les mots-clés établis manuellement qui restreignent l'amplitude de l'étiquetage et donc les possibilités de liaisons avec d'autres corpus. Nous identifions alors comment ces 3 ensembles interagissent. Nous cherchons à détecter d'éventuelles erreurs.

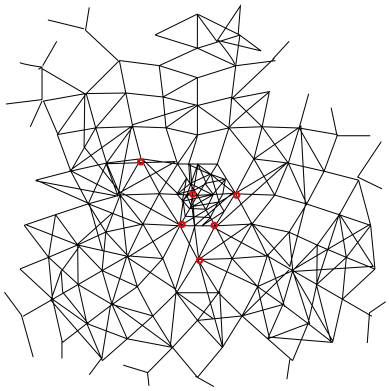
Nous ne modérons pas les résultats de *Haruspex* pour se placer dans les conditions d'une intégration continue automatique de nouvelles *pages*, fiches historiques liées aux objets de la maquette du port de Nantes du musée d'histoire de la ville de Nantes.

1.2 Résultats

Les résultats sont principalement des graphes qui montrent comment les différentes sources interagissent et se complètent.



(a) Exemple de réseau d'information original (ex : Nantes 1900)



(b) Augmentation du réseau original par des contenus liés, densifiant localement le réseau

FIGURE 4.1 – Densification locale d'un réseau existant par agrégation continue de production scientifique.

	ACB (cyan, préfixé 0)	parcelle Voruz (jaune, préfixé 1)	*Nantes1900* (violet, préfixé 2)
nb de *pages*	90	82	381
nb. de mots moy.	558	282	212
écart type	565	174	214

TABLE 4.1 – Caractéristiques des 2 mémoires et de *Nantes1900*

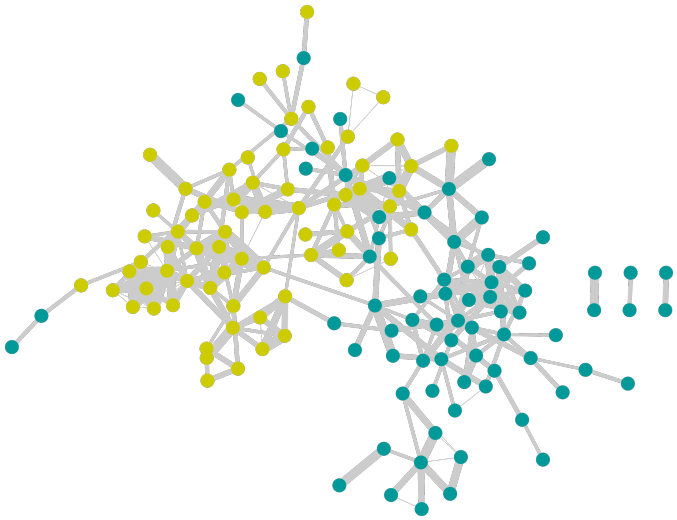


FIGURE 4.2 – Représentation des connexions les plus fortes entre les *pages* des 2 mémoires

1.2.1 Sur les 2 mémoires

Globalement. Haruspex découpe les 2 mémoires en *pages* de taille aussi homogène que possible en faisant varier le niveau de découpe (paragraphe ou sous-partie), en évitant les contenus trop courts et en découpant les contenus trop longs et tentant de produire un nombre similaire de *pages*. Le tableau 4.1 montre que le mémoire sur la parcelle Voruz contient des parties plus courtes et de longueur plus homogènes. Il est également plus court. À titre de comparaisons nous exposons les caractéristiques de *Nantes1900*, non découpé par *Haruspex*. Une première analyse visuelle de la figure 4.2 montre qu'une section d'un mémoire a plutôt tendance à être liée à une autre section du même mémoire. Une analyse plus fine mais toujours globale (figure 4.3) filtre uniquement les plus fortes liaisons (liens uniques, voir section 5.2.2) du graphe. D'abord cette analyse confirme le résultat précédent, les liaisons les plus fortes sont intra-mémoire. Ce premier résultat possède une valeur en soi (qui n'intéresse pas l'historien) : il s'agit d'un clustering efficace, on sépare assez simplement les 2 mémoires avec les résultats de *Haruspex*.

La figure 4.3 met en évidence un point d'intersection majeur entre les mémoires : la naissance de société anonyme (les ACB) qui occupa la parcelle Voruz et qui fait l'objet du mémoire sur les ACB. Ce résultat n'est pas surprenant et confirme l'intuition. Cette jonction majeure est composée d'une sur-représentation des occurrences de « Atelier et Chantiers de Bretagne » et « ACB » dans le mémoire sur la parcelle Voruz (termes très courants dans l'autre mémoire). De nombreuses occurrences des termes « société anonyme », « naissance », « Brosse et Fouché », « société anonyme des Ateliers et Chantiers de Bretagne » distinguent cette *page* du mémoire sur les ACB.

Les jonctions entre les 2 mémoires. En réalité la jonction majeure (et triviale) mise en évidence par la figure 4.3 occulte plus de 1700 jonctions mineures entre les 2 mémoires. Il serait vain de vouloir toutes les visualiser. Nous pouvons néanmoins visualiser les plus importantes de ces liaisons, toujours sur la base de la pondération proposée en section 5 pour les liens uniques. La figure 4.4 montre les jonctions repérées entre les mémoires. Ces jonctions sont faibles pour la plupart, le graphe est très peu dense. Ce filtre montre également des liens forts mais inintéressants pour l'historien des techniques, par exemple : un lien entre

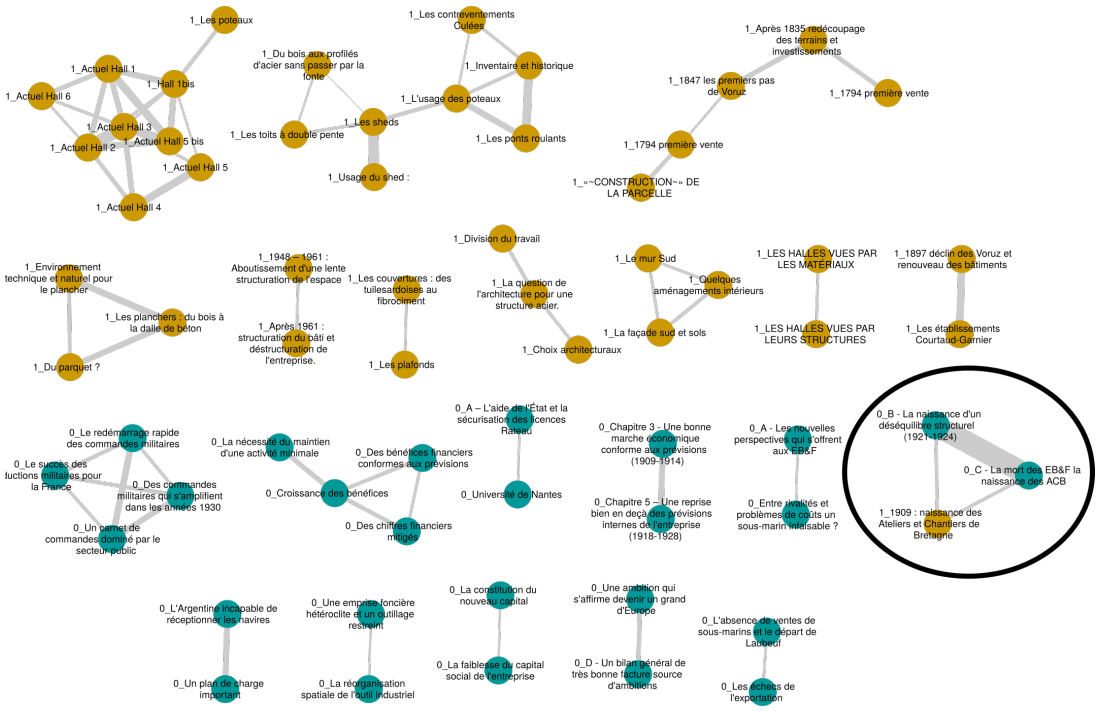


FIGURE 4.3 – Liens uniques les plus structurants du graphe. En jaune le mémoire de la parcelle Voruz, en bleu le mémoire sur les ACB. Entourée une zone de jonction majeure entre les mémoires.

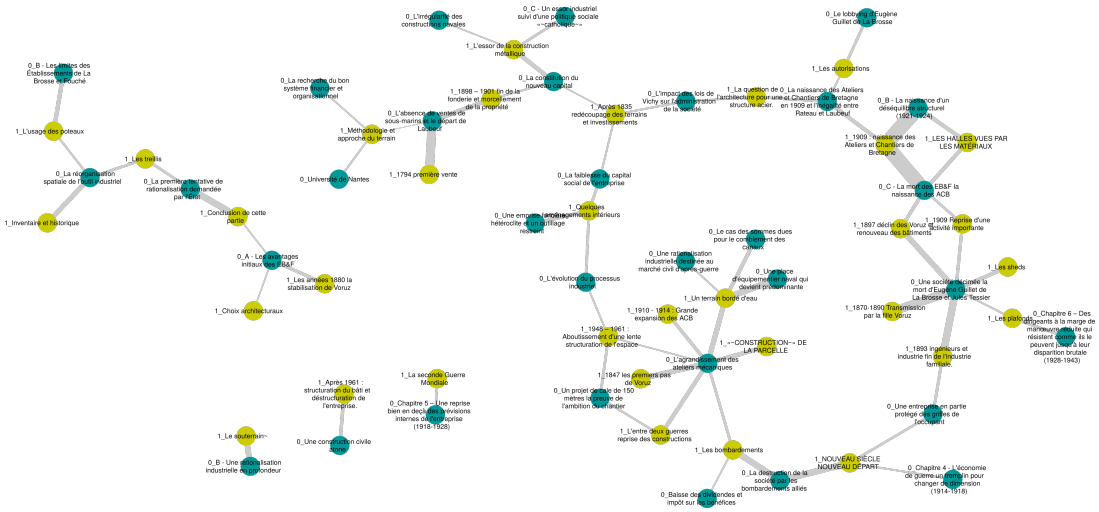


FIGURE 4.4 – Uniquement les liens entre les 2 mémoires les mieux pondérés.

les *pages* « absence de vente de sous-marins » et « 1794 : première vente » articulés autour de la terminologie de la vente.

Néanmoins on identifie une plus forte densité en bas à droite, cette densité se divise en 2 noyaux, le premier à droite est lié à la gestion familiale des entreprises du XIX^{e} siècle ; le second noyau en bas est consacré à la construction, aux bombardements et à la reconstruction des bâtiments. Un zoom sur la partie construction et bombardement (figure 4.5 permet de saisir l'intrication des parties. Le mémoire sur la parcelle Voruz s'intéresse à la construction et consacre une longue étude aux ateliers mécaniques des ACB qui ont occupé la parcelle, la partie du mémoire sur les ACB intitulée « L'agrandissement des ateliers mécaniques » agit comme un pôle de jonction.

Le choix de ce filtre et les représentations qui en découlent sont contestables, car 2 *pages* issues du même mémoire ne peuvent pas établir de lien entre elles. Il serait intéressant de disséquer la composition des jonctions de plusieurs sections de mémoires différents sur des thématiques communes, en supprimant l'obligation de se lier à une section de l'autre mémoire.

Composition de jonctions. Une jonction est composée de nombreux liens, par exemple chaque mémoire dédie une section à l'agrandissement et une autre aux bombardements. La figure précédente (figure 4.5) montre que ces éléments sont liés. Comment sont liés ces textes *a priori* antagonistes ? La notion de sémantique (bombardement vs. agrandissement) l'emporte-t-elle sur les distinctions entre 2 auteurs ? La figure 4.6 répond en partie à cette question. Les bombardements et l'agrandissement forment 2 ensembles distincts, fortement liés en interne (à droite et à gauche de la figure 4.6), reliés par des notions d'urbanisme (site, carte,

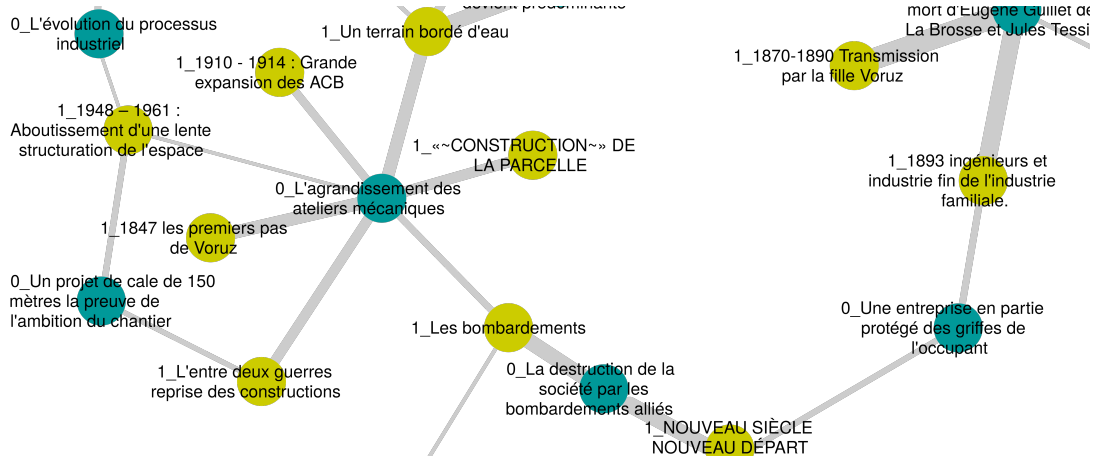


FIGURE 4.5 – Zoom sur la partie la plus dense du graphe 4.4.

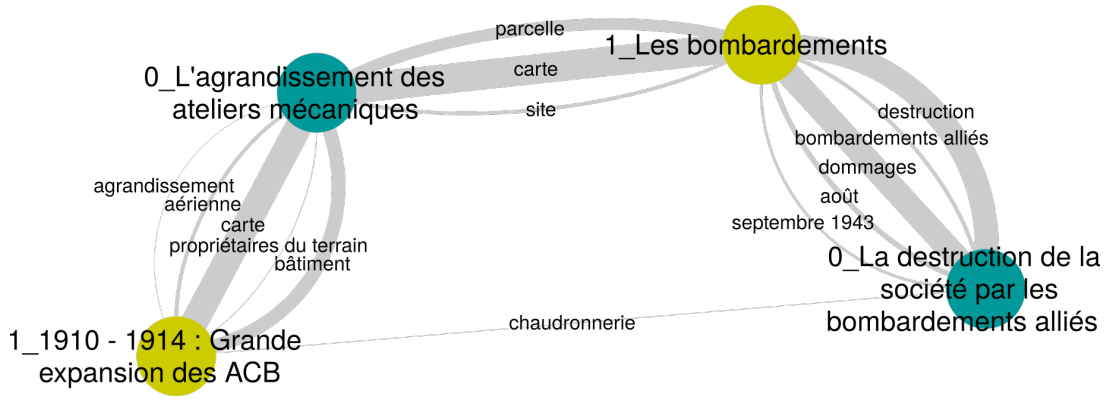


FIGURE 4.6 – Composition des liaisons d'une jonction d'antagonistes.

parcelle). Dans ce cas l'empreinte de l'auteur disparaît totalement. Ce résultat est prometteur pour lier un réseau à un ensemble plus grand, écris dans un style totalement différent.

1.2.2 Les 2 mémoires en lien avec *Nantes1900*

Nous ajoutons ici le corpus de *Nantes1900* en considérant que chaque fiche est un *page*. Nous réalisons une extraction des expressions-clés et un calcul des liens sur l'ensemble des fiches (des 3 sources).

Premier résultats, échecs et correction. Étant donné les thématiques très diverses abordées par *Nantes1900*, les premiers résultats avec les paramètres habituels (voir section 5.2.1) discriminaient fortement les thématiques abordées par les mémoires au profit des thématiques plus spécifiques à des petits îlots (cluster) de *pages* : les engrais, le bois, une rue, un pont, une biscuiterie, etc. En effet, *Haruspex* valorise les particularités, les relations spécifiques et discriminantes. C'est pourquoi les thématiques abordées par les mémoires étaient exclues. Elles sont partagées par un grand nombre de fiches (un tiers des fiches environ), et donc jugées génériques et inintéressantes pour l'analyse. Nous avons dû adapter une fonction de pondération qui interdit les liens trop spécifiques (pour moins de 3 *pages*) et autorise les liens plus génériques (jusqu'à 30 *pages*).

Globalement. Dans les conditions établies précédemment, nous obtenons un large réseau pour *Nantes1900*, qui englobe les mémoires. En effet, les thématiques sont plus diversifiées et le nombre de fiches plus important. La figure 4.7 ne montre que *Nantes1900* fonctionne par hubs (périphériques sur la figure 4.7) et que certaines de ses *pages* viennent améliorer la jonction entre les mémoires. La distinction entre les 2 mémoires est toujours assez nette. Il existe quelques rattachements périphériques (en cyan à droite, en jaune en haut) qu'il serait intéressant d'étudier dans une perspective de compréhension du patrimoine et d'analyse historique, comme le chapitre 3. Mais ce n'est pas l'objet ici.

Comme précédemment, nous nous penchons sur les modes de jonction entre ces différentes sources.

Les jonctions. Les jonctions entre les sources sont principalement articulées sur la partie centrale du graphe de la figure 4.7. Nous limitons donc notre analyse à cette zone. La figure 4.8 montre la composition du noyau dense. Sans surprise les *pages* traitant des Ateliers et Chantiers de Bretagne est centrale. Les résultats ne présentent aucune aberration apparente : les pages concernées sont toutes effectivement à la jonction des 3 sources : Ateliers et Chantiers de la Loire, Jean-Simon Voruz, Joseph

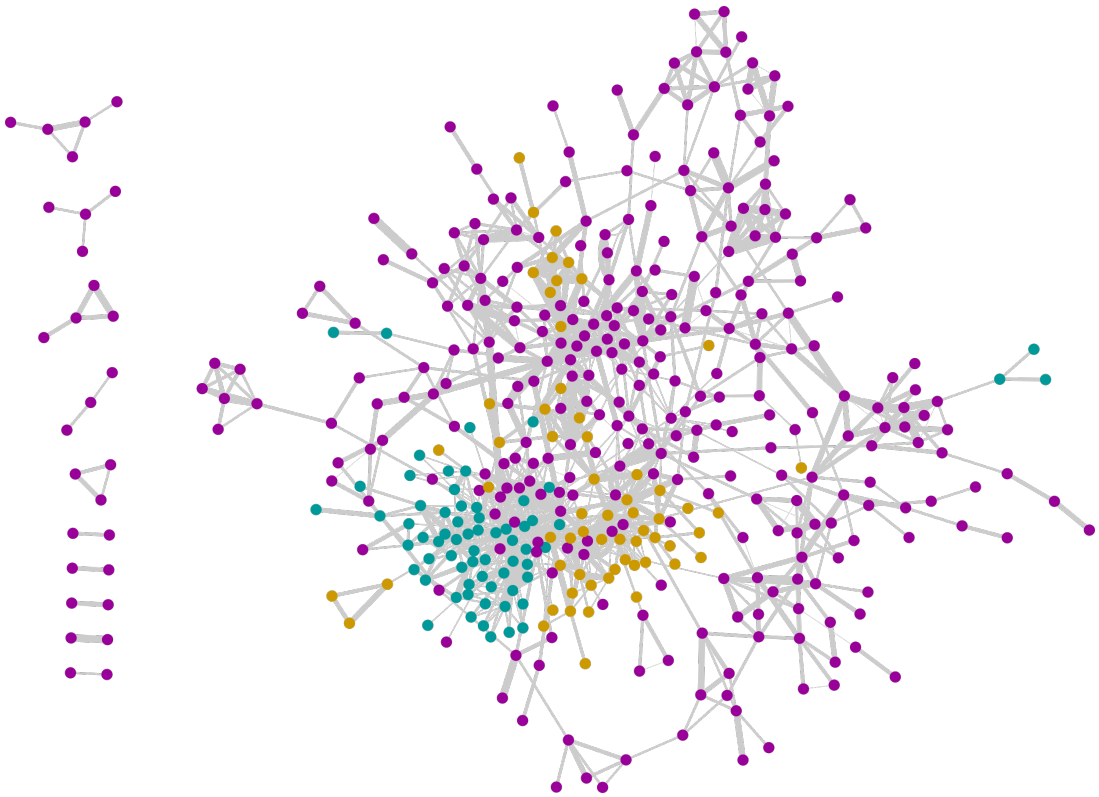


FIGURE 4.7 – Interaction entre *Nantes1900* (violet), les ACB (cyan) et la parcelle Voruz (jaune).

Paris, Canal Jollet, Dubigeon, Construction navale, outil industriel, sheds, etc. Nous établissons 3 remarques supplémentaires

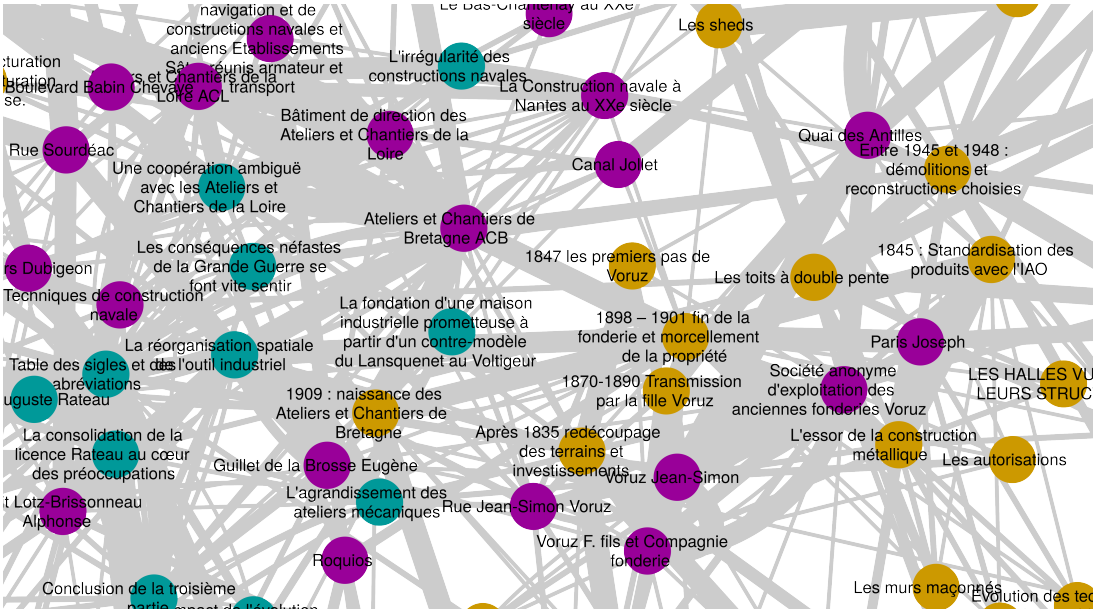


FIGURE 4.8 – Zoom sur le centre principal des jonctions du graphe 4.7

concernant la contextualisation des données :

— Le mémoire sur la parcelle Voruz contextualise fortement les ACB avec le monde de Voruz et les techniques de construction (construction métalliques, IAO, mur maçonnées, autorisations, premiers pas de Voruz, etc.).
— Le mémoire sur les ACB contextualise les ACB avec des éléments de la vie de l'entreprise : licence Rateau, conséquences de la guerre, coopérations industrielles, etc.
— Le « dézoom » sur des données périphériques (impossible à afficher ici) fait apparaître d'autres éléments de *Nantes1900* contextualisant les mémoires : Blasse Alfred, chantier naval, grands voiliers, Bas-Chantenay au XX^{e} siècle, Roquios, etc.

Nous pouvons parler de « contextualisation réciproque » des données. En effet, nous observons ici qualitativement une complémentarité entre les apports.

Exemple de contenu de jonction. Nous nous intéressons maintenant au contenu de la jonction principale. Pour cela nous « déployons » les liens de types *lien-clé* qui sous-tendent les liens uniques précédemment exposés. Pour des raisons de lisibilité nous n'affichons que les liens les mieux pondérés. La figure 4.9 montre le proche voisinage des Ateliers et Chantiers de Bretagne et les expressions partagées entre les fiches, origines des liens. Nous remarquons l'importance des expressions complexes (composées

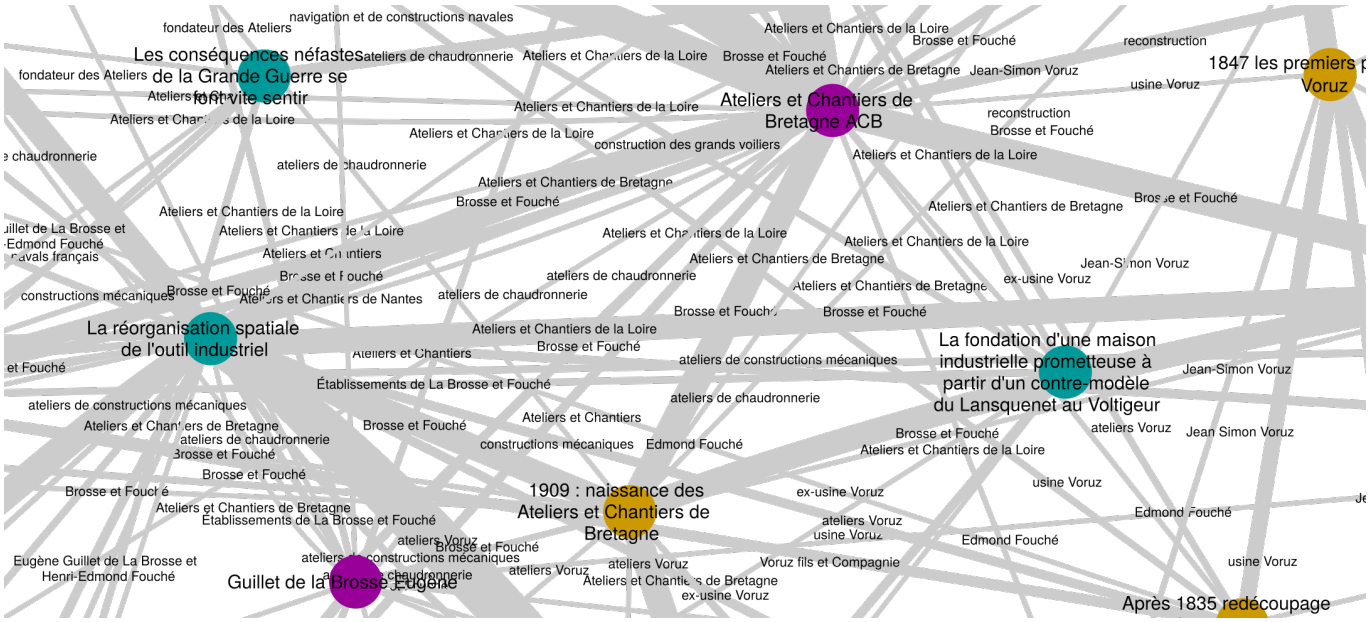


FIGURE 4.9 – Composition des liens structurant le noyau des jonctions entre les 3 sources

de plusieurs mots) dans la génération de ces liens.

1.3 Conclusion

À propos des jonctions. Nous montrons par cette courte étude, très peu historicisée, la double possibilité de séparer des contenus différents sans nombre cible (clustering) et de trouver des jonctions entre ces contenus différents et d'explorer ces jonctions. Ces jonctions sont à large échelle fortement structurées par l'auteur du texte, mais à plus petite échelle, sur des thématiques précises, les liens peuvent concerner des relations sémantiques fortes, parfois totalement émancipées de l'empreinte de l'auteur.

À propos de l'étiquetage. Nous constatons également que l'étiquetage manuel et l'étiquetage automatique des contenus de *Nantes1900* ne produisent pas les mêmes résultats. L'étiquetage manuel des centaines de *pages* de *Nantes1900* propose des liens assez limités au regard des résultats de *Haruspex*. En conservant l'étiquetage manuel, de nombreuses *pages* de *Nantes1900* n'auraient pas été liées aux autres sources. La différence entre les graphes manuels (voir travaux de Hervy (2014)) et automatiques sur les résultats de *Nantes1900* seuls montre un écart de densité et une forte variation du diamètre du graphe. Une étude sur ces variations pourrait être intéressante. Cette étude devrait être menée au croisement des disciplines de la muséologie et de la médiation numérique. Par exemple : quel est le diamètre idéal d'un graphe pour une application muséale ? La pondération permet-elle de filtrer le graphe pour influencer son diamètre et sa densité en conservant sa cohérence ? Ces variations répondent-elles à des besoins d'accès de différents niveaux ?

Perspectives d'application. Une application majeure de ce résultat consiste à agréger en continu des productions scientifiques (articles, thèses, mémoires) ayant une thématique commune pour constituer un graphe de contenus liés. Les liens devraient donc être — autant que possible — émancipés de l'empreinte de l'auteur. Ce graphe peut servir de documentation numérique à un patrimoine.
Par exemple dans le cas du projet Nantes 1900, un ensemble de fiches de documentation concernant le port industriel de Nantes au début du XX^{e} siècle est établi, le réseau existant peut être densifié localement par des mémoires, comme montré dans l'exemple précédent.

J'insiste ici sur la finalité d'une telle densification : elle s'adresse à un public avisé, curieux d'explorer les *pages* de *Nantes1900* et de trouver des détails issus de différentes sources. Typiquement un historien ou un chercheur. Les *pages* issues des mémoires ne sont pas formatées par la médiation et l'autonomie du contenu d'une page n'est pas assurée. Néanmoins il est toujours possible de re-contextualiser une *page* dans son environnement d'origine (le mémoire), puisque la structure initiale (du mémoire) est enregistrée. Il est donc possible de récupérer la section d'avant et d'après par exemple. Pour améliorer ce dernier point il faudrait conserver la possibilité de requêter une partie quelconque du mémoire, liée dans le texte (c'est-à-dire suivre les liens explicitement décris par l'auteur).

Attribute	CIRP
lang.	en
file format	pdf
internal struct.	no
dates	2008 - 2015
docs	109
data quality	≈
tot. words	400k
av. words/docs	3700
stdev.	431

TABLE 4.2 – Caractéristiques du corpus *CIRP Annals-Manufacturing Technology*

Perspectives d'augmentation du corpus. Cette courte expérience utilise 2 mémoires de Master, il existe de nombreuses études et sources textuelles au sujet de l'histoire du port industriel de Nantes au début du XX^{e} siècle. Pour une expérience de plus grande ampleur, il faudrait intégrer l'histoire des engrais chimiques dans l'estuaire de la Loire (Martin et Philippe, 2015; Martin, 2015), le pont transbordeur de Nantes, les raffineries de sucre (Robineau, 2011; Biette, 2013), les savonneries (Dutertre, 2005), les biscuiteries, etc. D'autres sources transversales pourraient également rejoindre le corpus, des sources d'autres éléments contextuels hors du port de Nantes pourrait aussi contextualiser certaines études, par exemple l'ouvrage fondamental de Rochcongar (2003) sur le sujet du patrimoine industriel et des grandes entreprises de l'ouest, ou encore des sources sur le pont transbordeur de Rochefort, ou de Bilbao, etc.

2 Étude des proceedings du CIRP

2.1 Introduction

L'étude des articles de l'assemblée générale du CIRP (*CIRP Annals-Manufacturing Technology*) réalisée à l'occasion d'une publication à la conférence CIRP design (Quantin *et al.*, 2016) montre les possibilités d'application à d'autres cas d'étude que l'histoire et le patrimoine. Il permet également de tester *Haruspex* sur un corpus en anglais et montre quelques possibilités d'études diachroniques rudimentaires (en 2015, *Haruspex* n'est pas totalement développé).

2.1.1 Le corpus

Le corpus (cf table 4.2) est constitué de 109 articles écris en anglais. Les thématiques de la conférence sont assez larges mais restent toutes liées à l'ingénierie et à la fabrication industrielle (*manufacturing*). Les documents sont au format pdf. La qualité de l'extraction du flux de texte est moyenne : les tableaux et certains autres éléments sont mal pris en compte. Le contenu des documents est stable : 3700 mots en moyenne avec un écart type de 431 mots seulement, pour un total de 3700 mots dans le corpus. Ces caractéristiques de forme du corpus sont plutôt propices à *Haruspex* : corpus de taille réduite, composé de documents non structurés et de longueur homogène, au contenu spécialisé.

2.1.2 L'invariant d'*Haruspex*

L'usage de *Haruspex* dans plusieurs contextes montre une sensibilité à lier entre eux les textes de même auteur. Ainsi le corpus de chimie du solide (chapitre 3) a du être amputé d'entretiens répétés avec le même solidiste, ici dans les articles, on retrouve parmi les documents les plus liés – et de loin !– des articles écrits par le même auteur à plusieurs années d'intervalle. Dans le monde de la recherche, la production scientifique la plus proche d'une production donnée est souvent la production du même auteur à une autre époque. La figure 4.10 montre cette cohérence très forte : il s'agit simplement des résultats des liens uniques les plus forts du corpus, point d'entrée classique d'une analyse avec *Haruspex*. De la figure 4.10, 2 paires font exception : ce sont des articles liés bien qu'ils soient d'auteurs différents :

— à droite de la figure : Astoul 2014 avec Mermoz 2011 qui ont respectivement écrit : « *New methodology to reduce the transmission error of the spiral bevel gears* » et « *A new methodology to optimize spiral bevel gear topography* ». Le lien fait totalement sens, les articles semblent traiter de problèmes très similaires.
— au centre de la figure : Paralika 2011 avec Salonitis 2014 qui ont respectivement écrit : « *Product modularity and assembly systems : An automotive case study* » et « *Modular design for increasing assembly automation* ». Le lien fait totalement sens, il semblerait que ces articles traitent de sujets proches.

À mieux regarder la première paire (Mermoz - Astoul), il s'avère que le second auteur de l'article d'Astoul est Mermoz et inversement. Cette paire n'échappe donc pas à la règle de forte cohérence avec soi-même dans le temps. Une interprétation optimiste que je propose ici serait que le duo a totalement co-écrit, au point de fusionner en un nouvel auteur double identifiable par *Haruspex*.

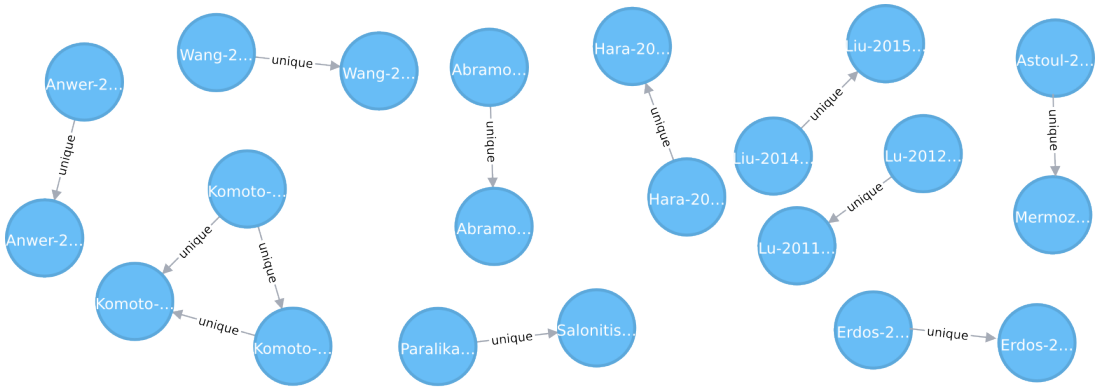


FIGURE 4.10 – Les articles les plus liés sont écris par les même auteurs

Seule la paire (Paralika - Salonitis) fait exception à la règle. Il aurait certainement été possible de trouver d'autres paires d'articles à interroger en abaissant les seuils. Cependant en abaissant les seuils le taux de paires d'articles du même auteur chute.

2.2 Quelques résultats sur des thématiques

2.2.1 Étude de thématiques

En conservant ces auteurs invariablement proches de leurs propres articles dans le temps, nous nous interessons maintenant à des thématiques identifiées avec *Haruspex*. La figure 4.11 trace différents exemples de thématiques entretenant des relations variées. La figure 4.11b est un exemple assez fréquent dans *Haruspex*, nous retrouvons un groupe dense et un groupe moins dense entretenant des relations asymétriques. Seul 1 membre de la communauté « virtuel » est totalement détaché du la « mécanique ». À l'inverse, seuls quelques membres de la communauté « mécanique » ont un lien avec le « virtuel », mais sont définitivement rangés du côté de la mécanique.

Enfin la dernière figure (4.11c) montre que certains membres du domaine de la mécanique (Perry, Huang, Hong, etc.) sont rattachés au domaine du virtuel. Le biomedical n'a aucun penchant pour le virtuel. L'article de HA(2009) qui passait inaperçu entre la « mécanique » et le « virtuel », devient le pivot de presque toutes les relations entre « médical » et « virtuel ». HA a écrit : « *Virtual prototyping enhanced by a haptic interface* ». Je n'ai pas lu cet article, mais il évoque explicitement le virtuel, la mécanique et le médical dans le corps de texte.

La communauté « medical » est relativement forte dans *Haruspex* car peu d'articles la mentionne, elle est discriminante d'une partie du corpus, mais le nombre d'occurrences des mots du domaine médical sont assez faibles (en comparaison avec la mécanique par exemple). Ce type de fonctionnement est typique de *Haruspex* : capable de faire ressortir des thématiques transverses qu'une étude des occurrences directes n'auraient pas mentionnées.

2.2.2 Étude diachronique de 2 thématiques

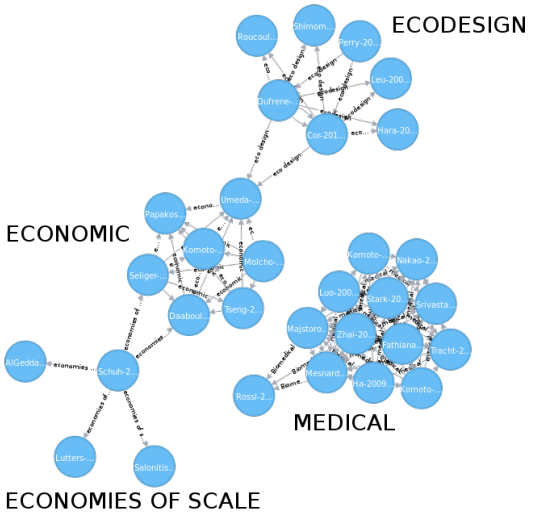
Cette étude diachronique de la représentation et des liens qui unissent 2 thématiques de la conférence au fil des années ne fait pas l'objet d'un choix particulier. D'autres thématiques auraient pu être étudiées. On remarque ainsi que « *life-cycle* » est absent avant 2012, et « *CAD* » est très faible. Dès 2012, « *life-cycle* » apparaît lié à « *CAD* ». Les communautés se séparent et se lient à nouveau en 2014 et 2015.

3 Conclusion

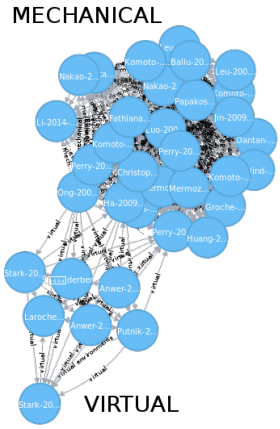
Analyse de CIRP entre 2008 et 2015 Il faudrait développer cette analyse et investiguer davantage. Un spécialiste de la communauté avec des questions précises aurait pu amener de Il n'y a pas d'intérêt particulier à utiliser *Haruspex* pour ce type d'étude simple sinon pour percevoir l'intensité de la relation entre les articles concernés.

4 Application pour la valorisation du patrimoine : le cas des Salons Mauduit à Nantes

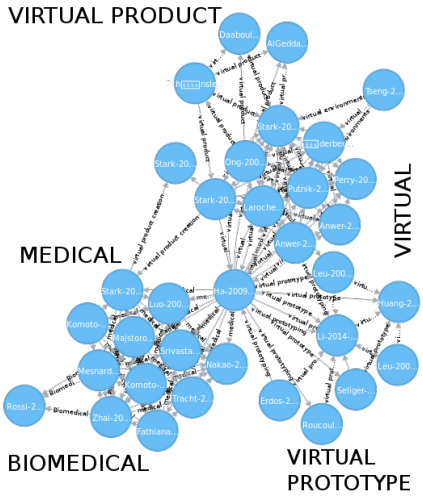
L'étanchéité de la circulation des données entre le monde de l'histoire et de celui du patrimoine, notamment lorsqu'il s'agit de science et technique n'est pas volontaire. Il existe un manque de documentation accessible au public avisé ou historien pour les collections des musées d'histoire, notamment orientées sur les sciences, les techniques et l'industrie. Pourtant dans la plupart des cas, une littérature scientifique (productions d'historiens) existe à propos des objets. Le Musée du Conservatoire des Arts et



(a) « éco » et « médical » sont des thématiques disjointes



(b) les thématiques « mécanique » et « virtuel » sont ponctuellement liées



(c) « virtuel » et « medical » sont des thématiques liées

FIGURE 4.11 – Différentes thématiques abordées par CIRP, et les relations qu'elles entretiennent.

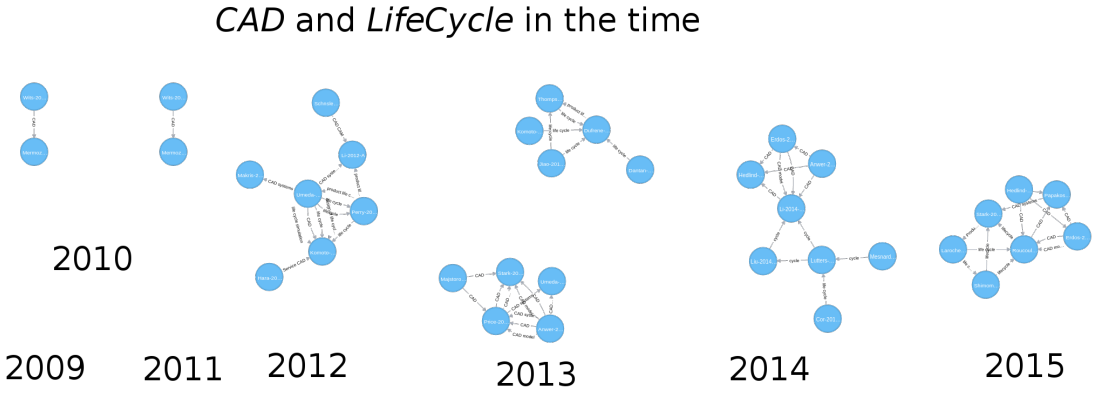


FIGURE 4.12 – Articles mentionnant CAD et life-Cycle dans le temps

Métiers est concerné par cette problématique : la documentation de ses collections est faiblement accessible au visiteur souhaitant dépasser le texte de vulgarisation du cartel.

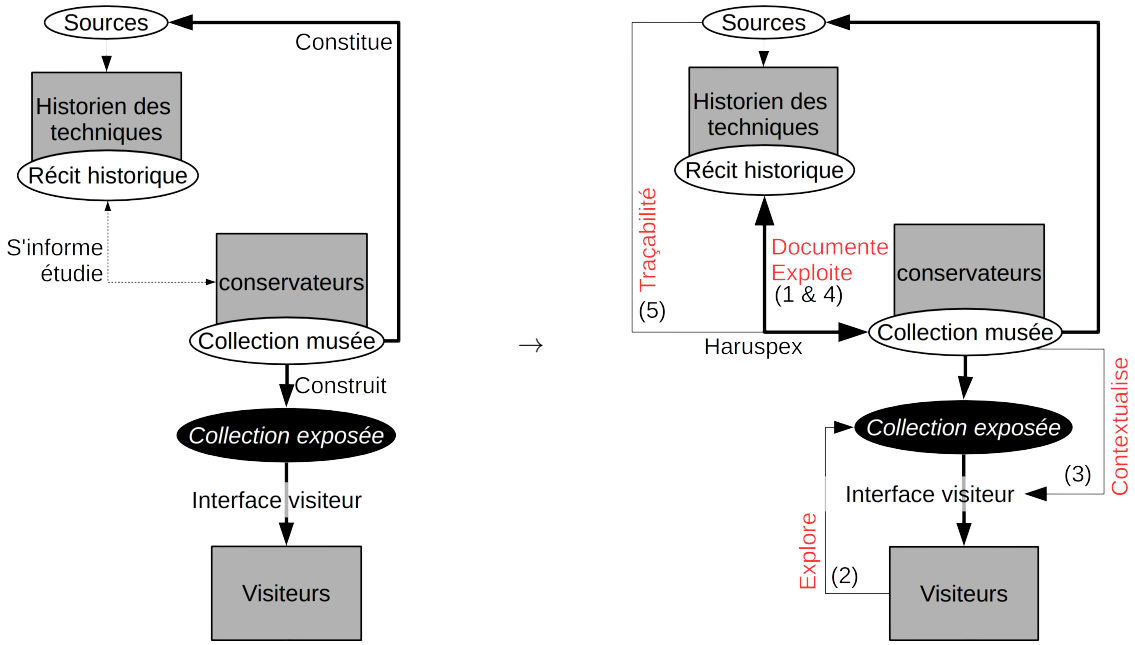
Cette section est en grande partie issue d'un article publié au cahier d'histoire du CNAM (Quantin, 2016).

4.1 Ambitions

Nous présentons ici 5 ambitions qui prolongent les ouvertures réalisées par le projet *Nantes1900* Hervy *et al.* (2014). Ces ambitions sont représentées par leur identifiant (numéro) sur la figure 4.13b.

(1) L'histoire comme support continu. La première ambition, colonne vertébrale de la proposition, est de considérer l'étude historique comme support continu de l'objet patrimonial et non pas seulement comme déclencheur du processus. Le patrimoine est une construction collective perpétuellement revisitée, et non plus « une fin en soi ». L'idée de support continu est incrémentée d'une dynamique de remise en question permanente, ou plutôt de réévaluation de l'objet. Il s'agit de créer une relation dynamique entre recherches historiques (récits) et objets de patrimoine dans le temps long. Ce processus vise aussi à valoriser les connaissances historiques hors des articles.

(2) S'approprier le patrimoine. La seconde ambition définit le patrimoine comme bien collectif à se réapproprier. Opposé à l'idée d'un patrimoine transcendantal, l'idée de réappropriation se place du côté de l'usage. Elle établit la valeur d'un patrimoine à partir de la reconnaissance dont celui-ci bénéficie, de sa capacité à être « continuité entre nous et l'ailleurs d'où il vient » (Davallon, 2000). Dans une conception topologique du témoignage, « passé et présent se superposent dans le présent de telle sorte que ce dernier en vient à former en quelque sorte un pli » (Davallon, 2000) ; il s'agit donc de créer une relation entre visiteur et objet de patrimoine.



(a) Principe actuel de fonctionnement

(b) Proposition avec enrichissements mutuels

FIGURE 4.13 – Représentation des différentes ambitions altérant le fonctionnement classique.

(3) Exploitabilité des collections. La troisième ambition vise l'exploitabilité des collections du musée comme contexte pour les collections exposées. Les musées possèdent des collections difficilement exploitables, dont seule une infime partie (moins de 5%) est accessible. Pourtant l'inventaire de ces collections existe.

La quatrième ambition (4), prolongeant la précédente, consiste à valoriser les collections du musée hors les murs (Malraux, 1965). Ce processus vise à donner accès (numériquement) aux collections, qui peuvent alors servir de sources. L'historien et le public avisé doivent pouvoir trouver des contenus riches dans une interface de médiation numérique.

(5) L'article comme API. Enfin la dernière ambition conçoit « l'article comme API » (le Deuff, 2014). Malgré un évident décalage entre l'idée d'API telle qu'implémentée en informatique et le principe évoqué par l'auteur, l'idée de requêter un article semble intéressante. Pour cela, il faut en structurer les contenus.

4.2 Expériences précédentes

4.2.1 *Nantes1900*

Le projet *Nantes1900* documente et contextualise une maquette des collections à l'aide d'archives et de sources historiques via un dispositif numérique. Les liens entre les notices (sources historiques) permet de naviguer sémantiquement ou géographi-

quement à partir d'un point géographique (point d'intérêt sur la maquette). Cette navigation permet d'accéder à des informations non-localisables (conditions de travail entre les guerres par exemple). La figure 4.14 [1] montrant l'interface s'approche de l'inten-

FIGURE 4.14 – Interface de navigation de *Nantes1900* (montage car mauvaise luminosité sur photos, Devocité 2013)

tion du schéma figure 5.

Perspectives. Un travail manuel fastidieux et chronophage (plusieurs années) construit le lien entre les résultats de recherche historique et le dispositif. Le dispositif final est éloigné des sources, son exploitation est compliquée pour l'historien. Par ailleurs si la proximité spatiale semble être quantifiable par la maquette, la proximité sémantique est binaire.

4.2.2 Exploration historique

Le projet de CITE BORIS LAM développé à l'École Centrale, consiste à démontrer le potentiel d'une interface de navigation dans des données multi- dimensionnelles. Il permet d'accéder à des informations non-localisables, ou atemporelles selon le principe de proximité. La figure 4.15 montre ce fonctionnement à l'aide de curseur dans une interface prototype ne contenant que des cubes. Le cube rouge est sélectionné, les cubes jaunes ainsi que d'autres éléments non-3D (événement) sont liés sémantiquement, accessibles par le menu. Ce prototype montre l'intérêt d'une navigation multi-dimensionnelle de proche en proche (calcul de proximités).

Perspectives. Il faut instancier ce prototype et trouver un moyen de calculer des distances fiables entre les items non spatiaux. Ces distances pourraient être adaptées au parcours de l'utilisateur. *Haruspex* permettrait de développer cette perspective du travail de CITER BORIS LAM. Nous montrons par la suite une approche conceptuelle pour y parvenir.

4.3 Approche conceptuelle

4.3.1 OLAP et Haruspex

Nous introduisons ici une vision OLAP des données (voir section 2.1.2) du patrimoine à partir des résultats d'*Haruspex*. Les dimensions retenues sont le temps, l'espace et la sémantique. Nous représentons donc nos données comme des points dans un cube à 3 dimensions (voir figure 4.16). Les proximités spatiales, temporelles et sémantiques sont issues d'Haruspex : calcul de proximité entre les documents sur différentes dimensions. On retrouve également l'accès multi-échelle caractéristique d'une structuration OLAP (*drill*). L'apport principal d'*Haruspex* ici consiste à produire des proximités sémantiques entre les textes d'une documentation du patrimoine. En effet les distances spatiales et géographiques ont déjà été résolues par les expériences précédentes. Pour l'iconographie, la proximité est calculée à partir du texte de description associé. Dans la perspective d'exploiter la production historique, ceci n'est jamais un problème, puisque l'iconographie est systématiquement commentée et analysée.

Les résultats d'*Haruspex* ne sont pas des distances (voir section 4.1), toutes les vues de cubes conceptuels présentées sont donc construites autour d'un item. En effet les proximités entre items sont réciproques mais ne respectent pas l'inégalité triangulaire.

1. montage car mauvaise luminosité sur photos

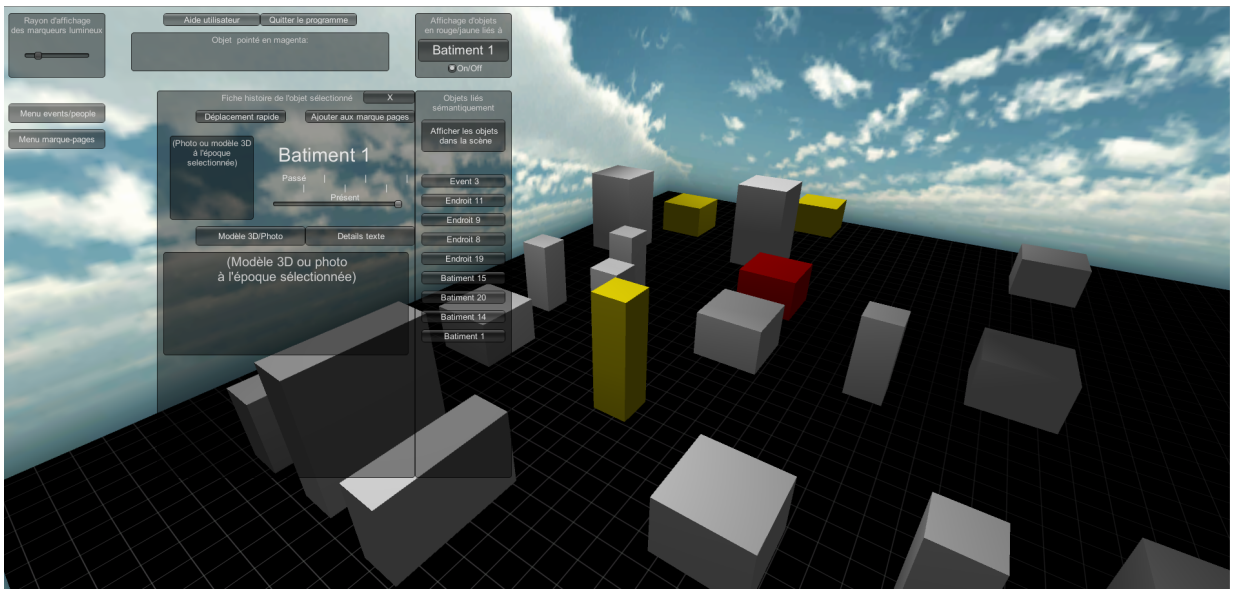


FIGURE 4.15 – Interface de navigation prototype de « navigation historique »

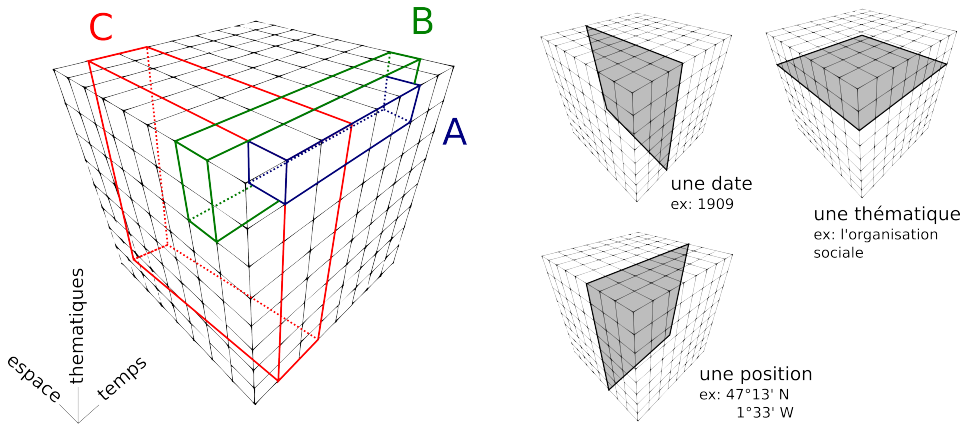


FIGURE 4.16 – Représentation OLAP de plusieurs ensembles de données : en bleu (A) l'analyse mécanique du pont transbordeur de Nantes ; en vert (B) analyse historique d'une usine de l'île de Nantes ; en rouge (C) les données de *Nantes1900*. Ces éléments présentent des intersections.

L'origine commune à tous les items est donc arbitraire, les proximités sont relatives. Il en va de même de la projection des distances géographiques (2D ou 3D) sur une seule dimension.

Les opérations de *drill-up* (voir section 2.1.2) sur la dimension sémantique fonctionnent de manière discrète avec plusieurs niveaux : les liens uniques (voir section 5.2.2), les liens de thématiques (voir section 2.1.5), les liens de classes, et les *liens-clés*. Chaque type de lien est pondéré et étiqueté, permettant de créer des filtres. Cette approche correspond au mantra « overview first, zoom and filter, then details on demand » (Shneiderman, 1996).

La forme base de données graphe telle que produite par *Haruspex* se prête bien au stockage de ce type d'information.

4.3.2 Visite guidée mais libre

L'existence et la structuration des données n'est qu'un potentiel inutile s'il n'est pas doté d'accès. Nous traitons ici des accès à la documentation du patrimoine telle que précédemment établie. La « visite guidée mais libre » est une approche conceptuelle de l'exploration de données multidimensionnelles du patrimoine. Appuyé sur notre construction précédente, nous proposons de construire un « fil rouge » : une liste ordonnée d'items.

Ce fil rouge correspond à la visite guidée. Il est défini à l'avance par un humain, le médiateur de l'objet. Inspirée des visites guidées réelles (par exemple d'un monument historique) 2 items peuvent avoir un lien sémantique fort et pourtant demander un saut dans le temps, un déplacement voire une représentation. Par exemple, la visite de l'aciérie Schneider du Creusot, peut montrer la maquette du CNAM représentant l'Aciérie Martin de Saint-Etienne, (modèle de J. Boudin, 1912). Nous représentons cette visite guidée parmi les éléments de la documentation du patrimoine liés par un fil rouge (voir figure 4.17). Différents parcours thématiques (différentes visites guidées) peuvent concerner le même objet. La partie « libre » de la visite guidée est également inspirée de visites réelles : en chaque point de la visite, il est possible d'explorer les environs : tel autre élément des

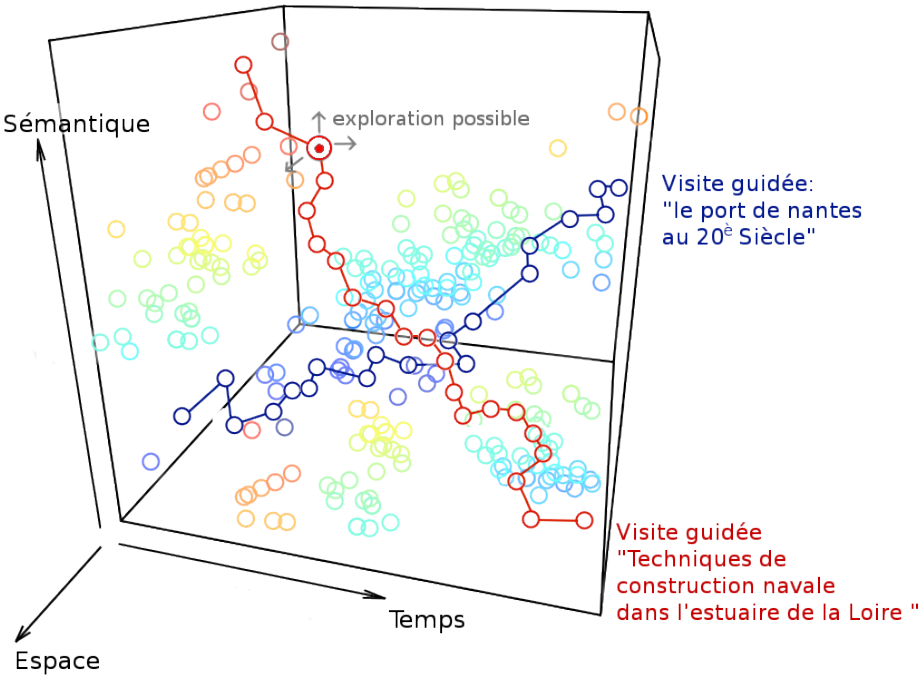


FIGURE 4.17 – Représentation en cube OLAP de 2 visites guidées parmi la documentation (les items) d'un même patrimoine.

collections ou de la documentation très proche sémantiquement, un usage ultérieur du même lieu, etc.

Pour reprendre des exemples précédents, dans une visite imaginaire d'un laboratoire de chimie du solide, l'évocation d'une application industrielle sur le fil rouge (par exemple la pile à combustible) amènerait la possibilité d'explorer le graphe des industriels qui amènerait sans doute à converger vers le cluster des ingénieurs et industriels du secteur automobile.

Ce type de visites permet donc de guider le visiteur parmi la masse de données, tout en stimulant sa curiosité par le potentiel de sérendipité qu'offre le graphe total.

4.3.3 Adaptation aux traces

La « visite guidée mais libre » est personnalisée à chaque utilisateur. La visite guidée est strictement fixe, définie à l'avance par un humain. Cette fonctionnalité est créée à partir des proximités calculées par *Haruspex* et des actions de l'utilisateur. Elle affecte la partie « libre » de la visite et vise à améliorer le potentiel de sérendipité des propositions.

Il s'agit donc d'exploiter les traces de l'utilisateur hors de la visite guidée pour lui proposer d'autres items de la documentation. Ce fonctionnement est inspiré de celui d'un ANN, et vise à apprendre le comportement de l'utilisateur. À un instant de la visite, si le visiteur choisit de sortir de la visite guidée pour explorer un item parmi les propositions, alors sur les 3 dimensions, tous les voisins de l'item choisi reçoivent un bonus ; les items non choisis reçoivent un malus. Ce système de bonus et malus est indépendant du graphe produit par *Haruspex* et ne vise qu'à modifier l'ordre de propositions faites à l'utilisateur.

Par exemple, lors de la visite d'un laboratoire de chimie du solide, si le visiteur quitte le fil rouge pour explorer la grève du laboratoire en 1953 (retarder l'âge de la retraite), alors tout item proche de 1953 ou étiqueté par la thématique « sociale » ou proche des lieux mentionnés reçoit un bonus ; à l'inverse les autres propositions non explorées reçoivent un malus.

4.4 Cas d'étude

Le cas d'étude proposé est une expérience qui n'a pas pu être menée à terme faute de moyen. Le projet, parmi d'autres projets concurrents, a été élu pour un financement de 6 mois afin de participer à un concours. N'ayant pas été retenu au concours le projet est resté en *stand-by*.

4.4.1 Présentation du sujet

Les Salons Mauduit sont un espace de réception à Nantes construit au début du XX^{e} siècle. Plusieurs destructions et incendies laissent des salles partiellement rénovées, constituées d'un vaste décor de plâtre dans un hangar, et de plusieurs dépendances (cuisines). Le lieu n'est plus aux normes, en 2015 il est détruit et reconstruit à l'apparence identique en sous-sol. L'intérêt que lui porte la municipalité (ville de Nantes) en tant que patrimoine local est multiple :

— la décoration et l'architecture du décor des années 30 typiques du style paquebot art-déco est intéressante [2].

2. quelques bas-reliefs seront déposés pour être réintégrés dans la reconstruction

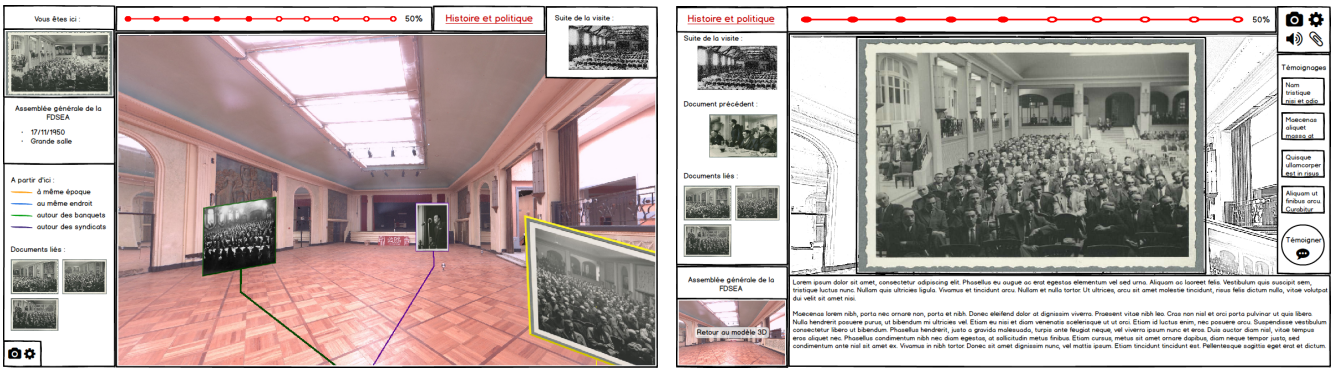
— le lieu a accueilli de prestigieuses réceptions au début du siècle (bal présidentiel) et des reconversions d'usage de guerre (hôpital), témoins de l'histoire de la ville.
— la plupart des Nantais ont participé à — au moins – un événement aux Salons. De nombreux et divers événements populaires et étudiants y étaient organisés à partir des années 1980. Il s'agit d'un lieu de mémoire collective nantaise.

4.4.2 Projet

Les principes. Les 4 piliers du projet qui ont guidé son développement sont : (1) la réalité virtuelle pour faire revivre le lieu original, (2) la « visite guidée mais libre » pour naviguer parmi des archives en se « déconnectant » de la 3D de la réalité virtuelle, (3) le *crowd-sourcing* (dépôt d'archives privées et de témoignages) pour impliquer les Nantais dans la patrimonialisation du lieu qu'ils sont nombreux à avoir connu, (4) la mise à jour permanente du contenu historique par les témoignages et collectes d'archives, calcul par *Haruspex* du réseau d'information.

Le phasage. Les phases du projet se sont déroulées comme suit. Dans la pratique le recouvrement entre les phases était non négligeable et certaines phases ont plutôt été menées en tâche de fond (continues) en raison des faibles délais. Néanmoins la liste suivante indique les activités et la décomposition du projet, il semble inutile de réaliser un diagramme de Gantt (ou autre) a posteriori.

1. Choix des technologies et transfert de connaissances (ingénierie documentaire et informatique avec histoire de l'art et patrimoine) : notions de données et métadonnées, notion d'histoire de l'art
2. Recollement d'archives (par les historiens)
3. Scan 3D (lasergrammétrie) du site avant destruction
4. Mise en place d'un système d'édition de métadonnées pour les archives
5. Déploiement d'un CMS (Omeka, voir section 2.3.2) pour édition collective des données, simplicité de stockage et communication [3]
6. Stratégie et formulaire en ligne de *crowdsourcing* pour collecter des archives privées
7. Création des visites guidées parmi les archives : sélection des items et écriture de textes.
8. Construction de la visite guidée mais libre : Structure de base de données centrée sur les documents d'archives, ETL depuis la base du CMS, création des vues et requêtes OLAP ; chaînage des items de la visite et ajout des commentaires de médiation.
9. Travail du nuage de points du site pour un rendu fluide et réaliste (maillage *low-poly* et *normal map*)
10. Maquette d'interface de visite en réalité virtuelle (figure 4.18), design d'interaction



(a) Navigation en réalité virtuelle dans les Salons et interaction lieu-archive

(b) Vue d'une archive et interaction avec son environnement multi-dimensionnel

FIGURE 4.18 – Maquettes d'interface de visite des Salons Mauduit en réalité virtuelle mixte (immersion 3D et « visite guidée mais libre »)

Place de *Haruspex*. Les 6 mois de projets n'ont pas laissé le temps de mettre en place une collecte et une analyse des documents par *Haruspex* pour établir un solide réseau sémantique. Les corpus de textes d'analyses historiques, et d'OCR des contenus n'ont pas vu le jour. Alors les 198 archives ont fait l'objet d'étiquetage thématique manuel (vocabulaire contrôlé) pour la démonstration de principe. *Haruspex* aurait été déployé dans une seconde phase du projet, défendant l'idée d'un patrimoine local, citoyen et dynamique : en permanence mis à jour par des données historiques nouvelles.

3. http://mauduit.univ-nantes.fr/

4.4.3 Résultats

Il n'y a jamais eu de développement de prototype fonctionnel de navigation en réalité virtuelle parmi les archives. Les trois principales briques du système sont restées décorrélées : (1) les archives et autres contenus historiques associés à des proximités initiales ; (2) la mise à jour permanente des contenus par *Haruspex* ; (3) la navigation en réalité virtuelle dans le site.

Il apparaît néanmoins que la création manuelle d'un système de navigation sémantique entre documents, est très laborieuse, même pour de petites quantités de documents (ici 198 archives). Ce résultat en négatif confirme ceux établis par l'expérience *Nantes1900*. Par ailleurs les résultats d'un étiquetage manuel sont très pauvres et ne permettre pas d'accès multi-échelle. Le

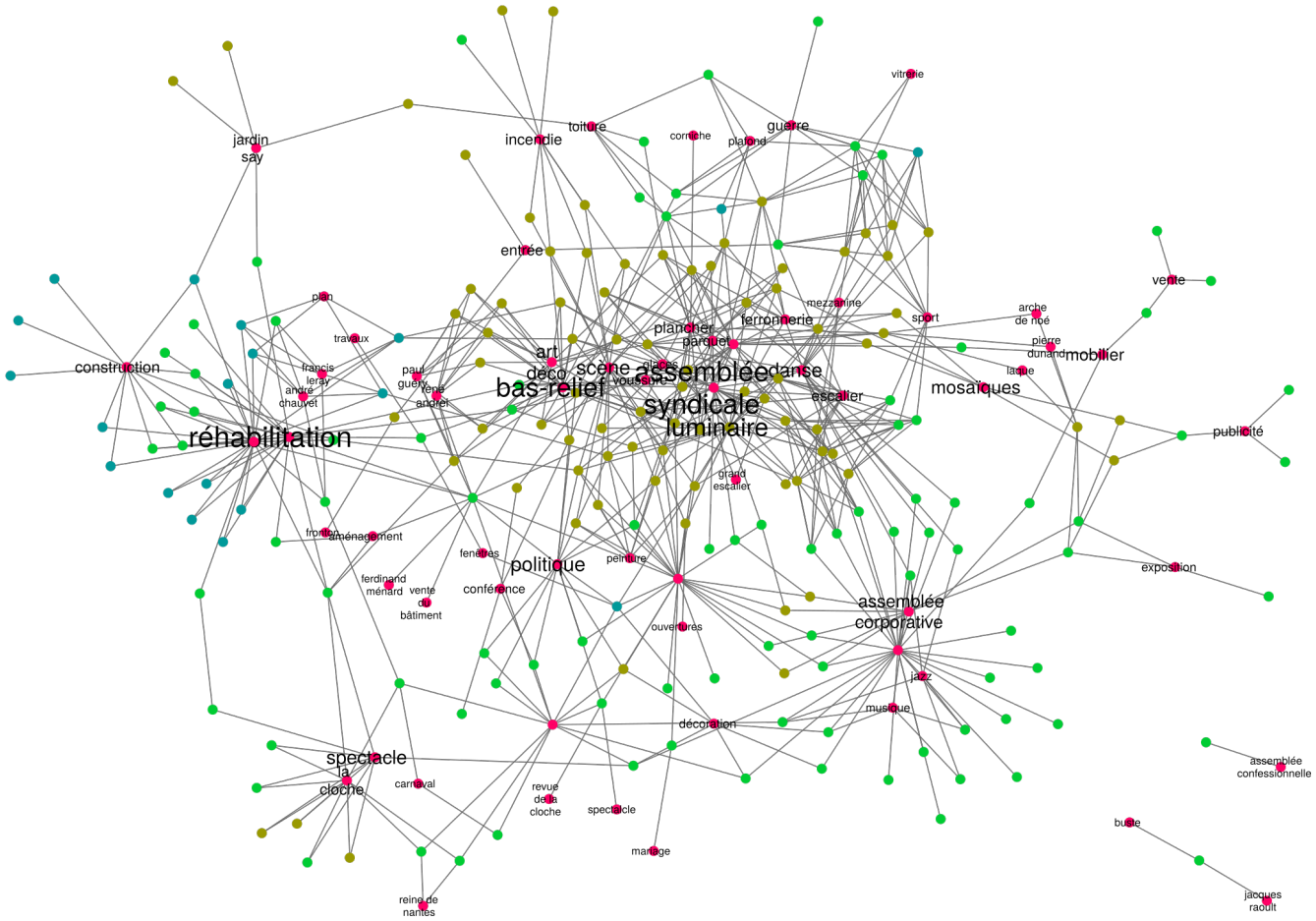


FIGURE 4.19 – Graphe de l'ensemble des archives sur les salons Mauduit : en rose les thèmes (centroïdes), en jaune les iconographies, en bleu les plans, en vert les textes

graphe issu de relations produites manuellement présente plusieurs inconvénients : il utilise massivement certains thèmes et en délaisse d'autres. Certains thèmes sont même inutiles (utilisés 1 seule fois). Les documents sont rarement liés à 2 thèmes.

Une analyse du graphe 4.19 (Table 4.3) nous montre qu'il est étalé (diamètre élevé par rapport au nombre de nœuds), il est moyennement centralisé, mais fortement hétérogène. Il s'agit donc d'un graphe avec des nœuds centraux déconnectés les uns des autres. Le nombre moyen de voisins est plutôt faible. Le coefficient de clustering d'un nœud (figure de droite) correspond au nombre de relations qu'entretiennent entre eux les voisins du nœud (divisé par le nombre maximum de connexions possibles). Les valeurs très basses (voire nulles) de ce coefficient confirment les analyses précédentes : le graphe fonctionne par *hubs* autonomes : chaque hub est un thème, ces thèmes sont presque déconnectés et de tailles très variables.

4.5 Conclusion sur l'usage de *Haruspex* pour la documentation du patrimoine

Des versions visite in-situ des Salons Mauduit (application mobile) ont également été envisagées. L'introduction de jeux à partir des éléments d'archives a été proposée. Aucune de ces propositions n'a donné lieu à un développement, même prototypique.

Néanmoins, cette expérience vient compléter et approfondit les résultats de la section 1. Elle montre le potentiel de *Haruspex* pour la documentation d'un objet patrimonial. Elle établit quelques grands principes : « intégration continue d'informations historiques » dans une application de valorisation et de documentation du patrimoine, « visite guidée mais libre » pour s'adapter aux besoins des utilisateurs tout en proposant des contenus très riches.

Cette expérience sur les Salons Mauduit, à première vue en décalage avec l'application directe de *Haruspex* telle que le chapitre 3 en fait la démonstration, révèle finalement un approche inhabituelle à la fois des sources historiques et du patrimoine.

propriété	valeur
diamètre	11
nœuds	260
centralité	0.087
hétérogénéité	0.951
nb. de voisins moy.	4.55

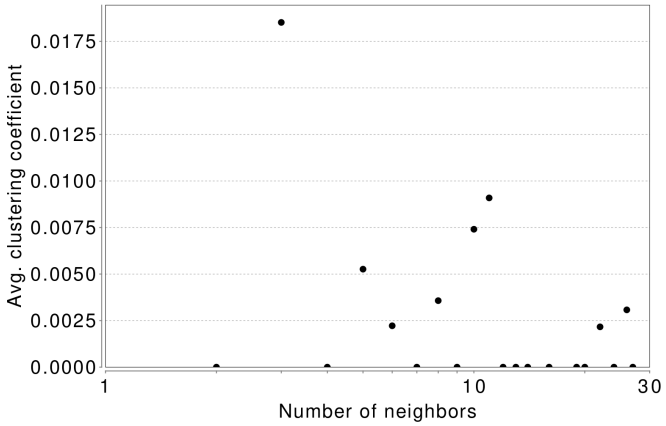


TABLE 4.3 – Caractéristiques du graphe de la figure 4.19

En effet les sources historiques sont alors exploitées (découpées, indexées, calculées, associées, etc.) au bénéfice de la documentation d'un patrimoine. En corollaire les sources sont valorisées dans un usage de documentation auprès du public avisé. On peut parler de valorisation mutuelle des sources historiques et du patrimoine.

5 Conclusion

Les résultats obtenus avec *Haruspex* permettent des analyses poussées d'un corpus de textes, ils permettent également de nouer des relations entre production historique et patrimoine.

Valorisation mutuelle et visite guidée mais libre. Nous avons établi plusieurs relations mutuelles. Premièrement, dans la section 1 nous avançons l'idée de contextualisations mutuelles des objets de la littérature scientifique en histoire. *Haruspex* serait un moyen de parvenir à réaliser cette contextualisation mutuelle. Nous démontrons également qu'il est possible d'enrichir des réseaux d'information existants par ces nouvelles sources.
Secondement, dans la section 4 nous avançons l'idée de valorisation mutuelle de la littérature scientifique en histoire et des objets du patrimoine. Cette valorisation mutuelle est réalisée par l'intégration continue d'informations historiques dans la documentation d'un objet patrimonial. Un mécanisme de visite guidée mais libre permet d'accéder à cette documentation au choix.

La concurrence des technologies de l'information pour la documentation du patrimoine est forte. Avec les concepts avancés ici (valorisation mutuelle et visite guidée mais libre) *Haruspex* marche sur les plate-bandes des ontologies (type CIDOC-CRM). Ces dernières sont totalement qualifiées pour documenter finement un patrimoine. Il serait intéressant d'établir un comparatif : durée de mise en place, opacité de l'information, finesse des relations, etc. La valeur ajoutée de *Haruspex* se déplace alors du potentiel d'analyse vers l'intégration automatique.

Patrimoine dynamique. *Haruspex* permet une intégration continue et automatique de nouvelles informations issues de productions spécialisées. Cette nouvelle documentation du patrimoine non figée, en perpétuelle évolution et remise en question va de pair avec une conception émergente du patrimoine scientifique et technique telle que M. Cotte la pense. En effet le patrimoine scientifique et technique semble être un cas particulier au sein des patrimoines. Une des particularités qui le distingue des autres (patrimoines culturels) est la multiplicité de ses états de références (voir section 1.2), ses cycles de vie nombreux et intriqués, et parfois même une continuité d'usage. Il s'agit d'un patrimoine dynamique qui appelle une documentation dynamique également.

Ici encore, de nombreuses autres technologies existent et font une forte concurrence potentielle à *Haruspex*. En effet les projets comme NELL (voir section 3.3) sont déployés à large échelle. Il serait intéressant d'envisager des projets plus ciblés, des essais de documentation d'un nombre restreint d'objets par peuplement d'ontologies à partir de sources spécialisées. La limite de cette critique est la nécessaire prédétermination des classes (verrou 2) qui empêche l'émergence de thèmes nouveaux et inattendus, capables de bousculer le positionnement de l'existant. Sur ce point les graphes faiblement structurés garderont l'avantage.

Vers le monde de l'industrie. La diversité des exemples montre que *Haruspex* est à la fois indépendant de tout domaine d'étude et conçu pour analyser les spécificités de chaque domaine indépendamment. Cette capacité pousse à prolonger la réflexion de l'intégration continue de connaissance dans le cadre du patrimoine. *Haruspex* serait-il capable d'intégrer de manière continue des connaissances issues de sources diverses et non structurées au cycle de vie d'un objet ? En absence de cas d'étude, la piste sera laissée ouverte. Ce n'est pas directement l'objet de cette thèse.

Chapitre 5

Épistémologie d'une méthode numérique

> "C'est l'homme qui met des valeurs dans les choses, afin de se conserver, -c'est lui qui créa le sens des choses, un sens humain ! C'est pourquoi il s'appelle homme c'est-à-dire, celui qui évalue. Évaluer c'est créer : écoutez donc, vous qui êtes créateurs ! C'est leur évaluation qui fait des trésors et des joyaux de toutes choses évaluées. C'est par l'évaluation que se fixe la valeur [...] . Les valeurs changent lorsque le créateur se transforme. Celui qui doit créer détruit toujours."
>
> Nietzsche (1885)

Contents

Nous proposons une réflexion épistémologique sur les humanités numériques à partir du cas pratique développé dans le chapitre 3. Nous analysons les interactions entre les sphères numériques et historiques en deux temps, la première plus tournée vers les sciences numériques, la seconde sur le rôle que pourraient jouer les outils numériques pour les humanités, notamment l'histoire des sciences et des techniques.

1 Réflexions sur l'application d'*Haruspex* à l'étude des archives orales

Le cas pratique permet de proposer une étude des influences hommes/machines à l'œuvre dans *Haruspex* puis de schématiser son fonctionnement selon une typologie des inférences.

1.1 Influences réciproques hommes/machines

Au cours de la construction d'*Haruspex* et de son application spécifique à un corpus d'archives orales nous identifions des influences réciproques. Conformément au schéma en figure 1c, ces influences relèvent de l'interdisciplinarité. Ces influences sont réparties selon 2 types d'interactions :

type A : L'humain influence le comportement des algorithmes à l'œuvre dans *Haruspex*
type B : Les résultats produits par l'algorithme influencent l'interprétation des textes et modifient les connaissances de l'humain

La constitution du corpus (type A) L'algorithme d'extraction de mots-clés des documents du corpus reposant sur une approche statistique, le choix de constitution du corpus influence de manière évidente les résultats de l'analyse. La constitution du corpus, par le choix des interviewés, des questions, de la durée des entretiens, est une opération relevant de la compétence du chercheur en sciences humaines.

La constitution du corpus (type B) Dans le cas d'étude décrit ici, le corpus a été constitué avant la conception de la méthode numérique. On remarque que la connaissance de la méthode numérique influence l'historien dans la constitution de son corpus (voir partie 1.2). Au delà de ces ajustements, si le corpus devait être constitué en connaissance de la méthode d'analyse, l'historien procéderait-il à une sélection plus large ou au contraire plus réduite des entretiens ? Utiliserait-il *Haruspex* en accompagnement pour la constitution du corpus et la conduite d'entretiens supplémentaires ? Garderait-il au contraire l'analyse numérique pour un usage *a posteriori* ?

L'objet d'étude analysé (type A) L'objet d'étude porte sur l'analyse d'une communauté scientifique. Ce choix, explicite, influence à la fois la constitution du corpus (item précédent), mais également la sélection des mots-clés non-pertinents, les modifications/enrichissements apportés aux données traitées (les métadonnées descriptive d'un interviewé comme la date de thèse, le genre, le laboratoire, etc.), les requêtes d'interrogation des résultats, et les modes de représentation des données.

Les algorithmes et leurs paramètres (type B) Conçu sans *a priori* sur les corpus potentiels à traiter, *Haruspex* contient de nombreux paramètres [1]. Ces paramètres sont de natures très diverses : seuils de fréquence à la fois sur la sélection de termes ou de liens entre entretiens, fonctions mathématiques et facteurs associés, nombre d'itérations lors de l'extraction de mots-clés. L'influence de ces paramètres sur la globalité de la chaîne de traitement numérique est difficilement quantifiable mais indéniable.

Les algorithmes et leurs paramètres (type A) La signification des paramètres n'est pas toujours univoque. Cependant, il est intéressant de constater que la conception même des algorithmes, de leur choix ou du choix des paramètres, est influencée par le périmètre global d'application d'*Haruspex* (analyse de corpus d'historiens, pour un historien spécialiste du domaine, documents choisis et étudiés, dans l'optique d'une exploration de leurs similarités). Le fait de cibler des expressions complexes spécifiques et discriminantes au sein du corpus impacte directement le choix des algorithmes, à la fois pour l'extraction automatique de mots-clés, mais également pour le choix des fonctions mathématiques servant à la pondération et au filtrage des liens entre documents. Plus qu'une influence, l'interdisciplinarité conduit au développement de nouvelles fonctionnalités et affine la méthode.

La sélection des mots-clés non pertinents Étape critique bien que non nécessaire dans l'absolu, la question du nettoyage des données implique des choix évidents. En effet, une fois l'étape d'extraction de mots-clés terminés et la priorité étant donnée à l'identification d'expressions spécifiques, aucune limitation supplémentaire sur le volume de termes extraits n'est imposée. Parmi les centaines ou milliers de mots-clés extraits, deux modes de nettoyage des mots-clés non pertinents influencent le processus algorithmique d'*Haruspex* :

1. **(type A)** un mode de nettoyage semi-automatique permettant d'identifier des groupes de mots jugés non pertinents par le spécialiste. Ainsi les actions de l'utilisateur influencent l'algorithme de filtrage.

1. il est possible de traiter un corpus sur le même mode (historique) quelle que soit sa thématique : histoire de la danse ou de la géographie de terrain par exemple

2. **(type B)** une intervention manuelle au cas par cas, qui n'est encore une fois pas rédhibitoire, mais qui influence la base de données générée. La base de termes servant à créer des liens de proximité entre documents s'en trouve assainie. Moins de liens non-pertinents seront présents dans la masse de liens générés. L'analyse historique que l'on peut faire des données obtenue dépend donc de cette étape.

La base de données générée (type B) La base de données finale dans le processus de traitement d'*Haruspex* contient à la fois les documents représentés sous forme de nœuds et des liens pondérés entre ces documents. La génération automatique des liens repose sur des choix de paramétrage qui relèvent des mêmes considérations que celles évoquées à l'item *Les algorithmes et leurs paramètres (type B)*.

Les représentations graphiques proposées Les représentations graphiques sont une forme intéressante d'influence réciproque entre les sphères humaines et numériques. En effet, les représentations proposées dans le chapitre 3 sont des vues filtrées (c'est-à-dire des troncatures, des coupes) de la base de données générée, constituée à la fois des entretiens, de leurs métadonnées, et des liens entre ces entretiens basés sur les mots-clés pondérés. Ce filtre est une traduction dans un langage informatique (langage de requête) d'une question posée par le chercheur. Cette étape est donc une discussion entre plusieurs disciplines, médiatisée par une machine. Il faut en effet interpréter, voire adapter, une question de recherche en fonction des possibilités de la machine et des données disponibles, qui sont le fruit d'une suite d'autres choix humains. Ces représentations ont alors plusieurs finalités : comprendre un phénomène particulier (voir section 3.2), mais également illustrer un propos auprès de personnes tierces (médiation, publication). Le choix du mode de représentation des données (données tabulaires, graphe, cartographie, graphes statistiques) a une influence sur la compréhension des données, alors même qu'elle n'est qu'une vue partielle de l'ensemble des données. C'est par la combinaison des représentations, issues des questions posées, que de nouvelles questions émergent et qu'elles peuvent servir la démarche de recherche. Enfin, les représentations sont parfois elles-mêmes la résultante de choix implicites : par exemple, la représentation sous forme de graphe est induite par des algorithmes de spatialisation qui eux-mêmes supposent des choix de conception. La compréhension des conséquences de ces choix est nécessaire pour assurer une compréhension globale des représentations construites.

1.2 Typologie des inférences dans *Haruspex*

En contre-pied aux propositions de méthodes a priori, nous proposons de retracer la méthode a posteriori, à partir de l'expérience du corpus sur la chimie du solide.

Nous retraçons l'utilisation d'Haruspex en huit étapes successives. Les quatre premières étapes correspondent à celles qui ont été détaillées comme processus linéaire théorique (section 1.2) : traitement du corpus, extraction d'expressions, enrichissement des expressions et création de liens. La figure 5.1 représente ces étapes de manière plus complète en indiquant les décisions humaines (flèches pleines) et les rétroactions machine (flèche pointillée) sous forme de boucles.

Les étapes 1 à 5 Les boucles de retour comptent pour beaucoup dans la robustesse et la finesse d'analyse d'Haruspex. Elles permettent de contrôler certains résultats intermédiaires et de contraindre le processus à des résultats jugés satisfaisants selon une perspective humaine. La première partie de la figure 5.1 (étapes 1 à 5) comporte deux boucles imbriquées (2-3-3bis et 3-4-5) et une rétroaction (3bis vers 2). La rétroaction indique un apprentissage de la machine grâce à la modération de l'utilisateur humain qui influencera une extraction ultérieure. La première boucle concerne l'extraction d'expressions et la seconde la création de liens. La seconde partie de la figure est plus simple à comprendre car elle comporte une seule boucle (6-7-8).

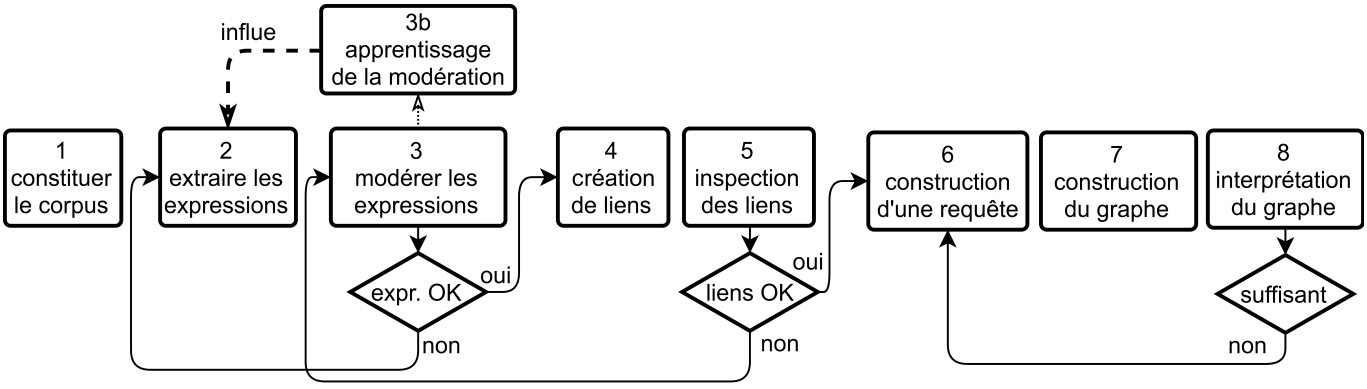


FIGURE 5.1 – Typologie des opérations suivies

Ce qui amorce une boucle est une décision humaine (symbolisée par un losange), motivée par la détection d'une *anomalie* dans les résultats produits. Qu'entend-t-on par là ? Une anomalie représente un *écart de valeur* entre l'évaluation algorithmique et

l'évaluation humaine. Ceci nécessite l'intervention d'un spécialiste du corpus. Le repérage d'un écart de valeur entre les sphères numérique et historique permet de *rectifier* les étapes antérieures avant de poursuivre.

Donnons quelques exemples d'anomalies. Au niveau de l'extraction de mots-clés (losange sous étape 3), la machine attribue une grande valeur (fort poids discriminant) à une expression comme « heure de cours » alors qu'elle a, selon le jugement historien, une faible valeur dans le corpus. La suppression d'un certain nombre d'expressions de ce type est suivie par une nouvelle extraction de mots-clés par Haruspex, qui donne une liste plus *signifiante* selon le jugement historien. Au niveau de la création de liens (losange sous étape 5), une anomalie peut être liée à une liaison locale, trop bien noté, à partir d'un mot-clé sans grande signification : « cycle de vie » par exemple nous semblait trop bien noté par la machine alors que ce concept peut être exprimé de nombreuses façons et qu'il n'est pas central dans « recherche sur les matériaux ». Cette seconde modération offre la possibilité de rectifier les causes d'erreur (mots-clés) plutôt que les effets absurdes (liens).

Les étapes 6 à 8 Au niveau de l'interprétation de graphes (losange sous étape 8), plutôt que d'anomalie, nous parlerons d'heuristique interprétative. En effet, l'interprétation de la figure 3.7 nous a conduit à fabriquer les figures 3.6 et 3.3. Cette boucle peut être parcourue un nombre indéfini de fois durant le processus de recherche. C'est aussi ce genre de considérations qui nous a permis d'écarter certains interviews qui faussaient la lecture du corpus : nous disposions de deux entretiens supplémentaires avec M. Pouchard et P. Caro que nous avons exclus du corpus car ils constituaient les deux interviews d'une même personne formaient des paires trop fortement liées. On pourrait ainsi figurer une nouvelle boucle partant du troisième losange vers la première étape.

L'analyse typologique du fonctionnement d'Haruspex (figure 5.1) montre deux caractéristiques fondamentales de notre méthode : sa non-linéarité, ce qui rend indispensable des rencontres répétées et des discussions contradictoires entre informaticiens et historiens ; ses possibilités de *rectifications numériques* à partir de *jugements de valeurs historiques*. Ce dernier point montre clairement l'incommensurabilité des sphères numériques et historiques dans *l'évaluation de la valeur des relations*.

1.3 Schéma de fonctionnement de *Haruspex*

Contrairement à certains algorithmes, *Haruspex* nécessite l'intervention d'un spécialiste du corpus étudié. Cette typologie de 8 opérations (fig. 5.1) linéaire peut redéployée suivant deux autres variables : en abscisse, la part relative de travail humain et travail machine ; en ordonnée, la part relative de fabrication de données et de savoirs (figure 5.2). Une telle représentation rend compte d'une conception continue et symétrique reliant données numériques et savoirs historiques. La figure 5.2 réécrit la

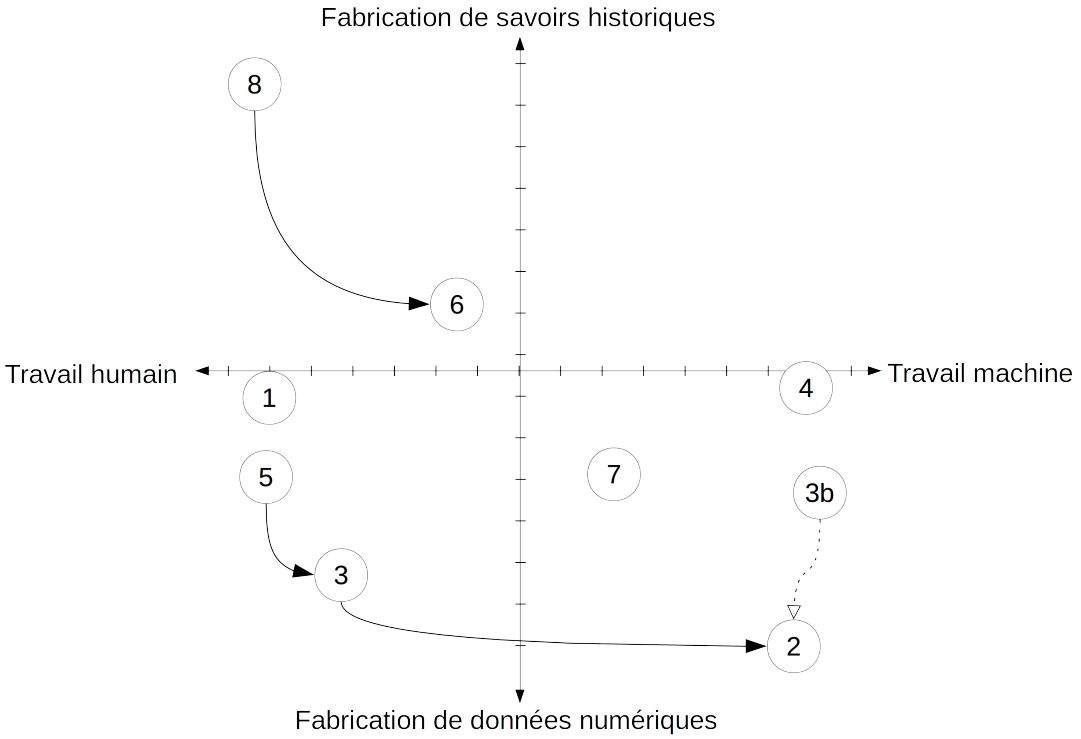


FIGURE 5.2 – Diagramme de la typologie des opérations suivant 2 dimensions, l'auteur et le produit du travail.

typologie et permet d'identifier certaines particularités d'Haruspex. Dans le paradigme hiérarchique de la pyramide DIKW [2] (voir section 2.1), nous concevons le passage des données aux savoirs selon un *continuum*. Les flèches de cette figure sont les issue de la typologie précédemment établie (figure 5.1). Nous énonçons ci-dessous les particularités d'*Haruspex* que met en valeur la représentation de la figure 5.2.

2. DIKW : Data Information Knowledge Wisdom

Place de la machine dans la méthode Le quart supérieur droit de la figure (5.2) est vierge. La conception d'*Haruspex* cantonne la machine aux inférences « bas-niveau » produisant de nouvelles données, éventuellement de nouvelles informations, mais ne livre pas de nouvelles connaissances ni n'établit de nouveaux savoirs. Les inférences « haut-niveau », orientées savoirs, restent le fruit d'une analyse par l'historien. En ceci, *Haruspex* offre une alternative aux approches de modélisation totale et *a priori* d'un domaine, aux inférences automatiques sur les données qui restreignent l'analyse du corpus à une approche de type « *distant reading* », en donnant la part belle aux possibilités de « *close-reading* » (voir section 1.1.1).

Interactions D'incessants allers-retours entre les parties droite et gauche de la figure (de 1 à 2, puis de 2 à 3, puis 3 à 4, etc.) sont symptomatiques des fortes interactions entre l'homme et la machine. Ces mouvements horizontaux sur le graphique sont combinés avec des mouvements verticaux (de 1 à 2 ou 5 à 6 puis de 6 à 7 par exemples) : ainsi des étapes de raisonnement ponctuent les étapes de récolement de données. Cette combinaison s'opère dans une tendance globale ascendante jusqu'à l'interprétation d'une représentation quantitative et graphique du corpus par l'historien (par exemple la figure 3.7).

Rectifications et heuristique Les 3 boucles présentées précédemment (section 1.2) ont la même forme : elles sont issues d'un travail humain et dirigées vers une étape précédente (indice inférieur), de travail davantage "machine" et "fabrication de données". Pour les deux premières boucles, l'humain demande à la machine de rectifier les données avant de procéder à l'étape suivante vers la fabrication de savoirs. Pour la troisième (grande) boucle, il s'agit de construction heuristique. Cette boucle génère de nouvelles représentations du corpus à chaque question de l'historien via la construction d'une requête. Dans notre cas, elle opère entre la figure 3.3 et les tables 3.5 et 3.6 d'une part et entre la figure 3.3 et la figure 3.6 d'autre part.

2 Vers une approche interdisciplinaire des humanités et des sciences numériques

Cette dernière section interroge le rôle épistémologique que pourrait avoir l'analyse numérique pour les sciences humaines, et plus particulièrement l'histoire des sciences et des techniques.

2.1 Expliquer et comprendre

Notre cas d'étude souligne le problème épistémologique fondamental de l'incommensurabilité des données numériques et des discours humains, marqués par l'incomplétude de leurs significations et pour lesquels il n'existe pas d'atome de connaissance Veyne (1971). On retrouve la différence introduite par Wilhelm Dilthey entre les sciences de la nature et les sciences de l'homme et de la société : « Nous expliquons la nature, nous comprenons la vie psychique ». L'explication des sciences naturelles tient surtout à la formulation d'un déterminisme de type causal, rendue possible par la quantification des objets naturels, notamment grâce aux outils mathématiques, numériques et expérimentaux. Les sciences humaines et sociales sont aussi orientées vers la recherche d'explications de type causale, sans possibilité d'expérimentation toutefois, ainsi que vers la formulation d'une compréhension des acteurs humains.

Cette intentionnalité humaine introduit des « effets de sens » (Launay, 1986) dans les sciences humaines et sociales qui les rendent irréductibles aux seuls déterminismes explicatifs. L'appréhension des humanités par les sciences numériques ne doit donc pas être comprise comme la réduction du champ social au champ numérique mais plutôt comme l'introduction d'un instrument supplémentaire (mais jamais suffisant) pour explorer le champ social.

2.2 Rôle instrumental des outils numériques

Un outil de mesures expérimentales. Masterman (1962) présente l'outil numérique comme « un télescope pour l'esprit » (*a telescope for the mind*) dans l'étude des textes. Plus profondément, nous défendons, avec Alfred N. Whitehead au sujet des sciences instrumentales, que l'outil numérique, en stimulant l'imagination des chercheurs, « montre les choses selon des combinaisons inhabituelles », ce qui conduit à « une transformation » de l'objet d'étude. [3]. Ainsi, chacune des représentations numériques a transformé notre perspective d'étude du corpus à la base de nouvelles interprétations.

Fiabilité des mesure. La question est maintenant de savoir si notre méthodologie fabrique une cartographie fiable ou une juxtaposition d'artefacts *ad hoc*. Trois critères de fiabilité soutiennent notre démarche : (1) la *pluridisciplinarité*, par laquelle la construction collective (interdisciplinaire) de représentations et d'interprétations fait écho à la singularité des regards individuels et disciplinaires ; (2) la *dualité* des productions numériques, qui sont simultanément signifiantes (image d'un objet d'étude) pour les historiens et signifiées (objet d'étude en soi) pour les numériciens ; et (3) la *sensibilité* d'*Haruspex*, dont les résultats évoluent à mesure que le corpus est modifié.

3. "The reason we are on a higher imaginative level is not because we have a finer imagination, but because we have better instruments. [...] a fresh instrument serves the same purpose as foreign travel ; it shows things in unusual combinations. The gain is more than a mere addition ; it is a transformation" cité par Ihde (Ihde, 2009)

Entre cartographie et mesure expérimentale. Au triple garde-fou méthodologique, il convient d'ajouter une souplesse conceptuelle et instrumentale. Nous concevons et produisons une cartographie dynamique du corpus. Sur ce dernier point nous nous accordons avec Bertin (1983) qui déclare « On ne "dessine" plus un graphique une fois pour toutes. On le "construit" et on le reconstruit (on le manipule) jusqu'au moment où toutes les relations qu'il recèle ont été perçues » . Mais à la différence de l'auteur, nous cherchons à multiplier les sens de la représentation en reconstruisant le graphique. Cette plurisémie de la carte rapproche nos observations du monde instrumental et l'éloigne des problématiques de *sémiologie graphique* [4]. En effet, chaque graphe présenté doit être interprété. Puisqu'il n'y a pas d'intention de contenu à priori, il est possible qu'aucune interprétation soit satisfaisante ou intéressante. En ce sens, la construction des graphes s'apparente à une sciences instrumentales avec des résultats d'expérience et d'observation, plutôt qu'à l'élaboration de schéma (ou carte) avec une intention, un message pré-défini. *Haruspex* contient la totalité des liens formés entre tous les nœuds par toutes les expressions spécifiques. Seul un sous-ensemble de cette totalité numérique est visualisé par une requête. La multiplication des requêtes n'a, jusqu'ici, pas donné de résultats aberrants.

2.3 Administration de la preuve

Cette dernière section pose plus de questions qu'elle n'apporte de réponse sur l'administration de la preuve en humanités numériques. Le principal problème est de définir la ou les fonctions épistémologiques des outils numériques pour les humanités. *Haruspex* fabrique une représentation numérique du corpus avec un bon degré de précision et de fiabilité sémantique. Il dessine une cartographie de la mémoire collective en traçant des inhomogénéités sociales et épistémiques. Ce faisant, il joue, pour les sciences de l'homme, un rôle instrumental comparable aux mesures expérimentales pour les sciences de la nature.

Pourtant aucun équivalent des méticuleuses calibrations des appareils de mesures des sciences de la nature n'existe pour les sciences de l'homme. Certes, des tests de la qualité de l'extraction (F-mesure) et des corpus étalons existent en traitement automatique de la langue (TAL) [5] ainsi que des thésaurus et d'autres taxonomies lexicales. La calibration d'outils d'analyse, en revanche, fait défaut. Ceci rend indispensable la connaissance préalable d'un corpus. L'administration de la preuve en humanités numériques articule interprétations, représentations et quantifications de corpus. Les commentaires critiques de tiers et les développements de méthodes alternatives par d'autres équipes constituent des rouages-clés de cette machinerie argumentative complexe sur laquelle s'appuient les humanités numériques. Face à la « quantophrénine » Sorokin (1956), l'école des Annales, montre l'heuristique des emprunts transdisciplinaires.

4. terme employé par Bertin pour désigner l'univocité de la représentation

5. Les étalons disponibles pour la NERC (section 3.2.1) qu'ils soient anglophones (MUC, ACE, DUC) ou francophones (ESTER, ester2) sont des corpus annotés permettant d'évaluer la performance d'un outil.

Chapitre 6

Conclusion

La proposition scientifique *Haruspex* se situe dans plusieurs catégories énoncées en état de l'art : traitement automatique du langage naturel (TAL), création de graphes ; structure de données et design de pratiques de navigation parmi les données.

Retour sur les verrous scientifiques

Les verrous énoncés sont tous abordés et résolus à des degrés différents. Certains verrous sont plus conceptuels, d'autres plus opérationnels ; tous sont assimilables à des contraintes circonvenant cette thèse. La résolution des verrous individuellement peut éventuellement avoir fait l'objet d'études plus ou moins poussées (cf. chapitre 1 État de l'art). La combinaison de ces différents verrous compose le problème.

Premier verrou. Le verrou intitulé « une vision dynamique du patrimoine » est partiellement résolu par la section 4. Il s'agit de donner les moyens à un système de documentation patrimoniale d'intégrer en continu de nouvelles informations. Ce verrou conçoit le patrimoine comme une production vivante et contemporaine, à l'opposé d'un élément figé issu du passé, imposé au présent. La résolution est indirecte : nous avons utilisé *Haruspex* comme un *moyen* d'y parvenir en étendant conceptuellement son champ d'action. Cette résolution est articulée autour d'une application (Salons Mauduit, section 4). La résolution partielle car les concepts de navigations parmi les données (« visite guidée mais libre ») n'ont été mis à l'épreuve d'aucune réalité.

Deuxième verrou. Le verrou intitulé « non-restriction du domaine d'étude » est totalement résolu par la modélisation des données de bas niveau en graphe libre (graphe multiple flou) en section 5. La légèreté des contraintes du modèle de données établi permet l'émergence de pistes de recherche sans *a priori* : aucun carcan pré-établi (série de classes et relations) ne contraint la direction que prendra le résultat et son analyse. L'interaction avec l'historien, alors maître de l'interprétation des données, est nécessaire et maximisée. Le chapitre 3 (Application : Cartographie d'un corpus en Histoire de la Chimie du Solide) démontre ce fonctionnement.

Troisième verrou. Le verrou intitulé « unicité et unité des corpus » est totalement résolu par la section 3 et de manière plus large par l'ensemble de la proposition *Haruspex*. En effet, les deux écueils énoncés sont évités. Premièrement aucun corpus extérieur (corpus étiqueté par exemple) n'est utilisé dans *Haruspex* pour le traitement d'un ensemble de textes. Secondement aucune sous-partie du corpus n'est utilisée comme échantillon représentatif pour le traitement du corpus entier. En pratique ce verrou évite les méthodes d'apprentissage supervisé. Dans ce cas, on peut considérer que les données obtenues sont exemptes de tout biais, à partir du moment où le corpus est établi. Ceci assure l'*intégrité* de l'analyse du texte, condition importante pour engager une étude croisée avec un historien.

Quatrième verrou. Le verrou intitulé « logique floue » est résolu par la section 5. Les données prennent la forme d'un graphe à arcs pondérés. Les propriétés mathématiques de la logique floue sont sous-exploitées par *Haruspex* et de manière générale par les cas d'étude présentés. Néanmoins ce verrou est levé et la restitution des données permet de retranscrire la variabilité des états non binaires. L'analyse nuancée de l'historien trouve écho dans ces pondérations d'arcs.

Cinquième verrou. Le verrou intitulé « représentation des mots » est résolu par la section 5.1.3. Il s'agit de dépasser la pure sémiotique pour tendre artificiellement vers sa sémantique. En quantifiant des similitudes entre mots (plutôt que de compter les formes exactes) les calculs basés sur ces valeurs d'occurrences sont plus représentatifs du sens contenu.

Sixième verrou. Le verrou intitulé « approche multi-échelle » est résolu par la section 5 d'une part (voir logique floue) et par la section 2.2 d'autre part, respectivement au niveau sub-corpus et super-corpus. La résolution de ce verrou en 2 étapes décompose le problème en 2 sous-ensembles imbriqués : le niveau super copus est en amont du process et n'est pas dynamique (pas de retour possible passé cette étape) ; le niveau subcorpus en aval, à son tour propose dynamiquement plusieurs niveaux d'accès

discrets d'abord (lien unique, lien pondéré) puis continu (fonction de pondération). La mise en application en section 1 confirme la possibilité de fournir différents niveau de détails (vues) d'un graphe multi-corpus.

Septième verrou. Le verrou intitulé « approche multi-critères » est partiellement résolu par la section 5.2 pour la partie technique et par la section 4 pour la partie conceptuelle. Il s'agit de proposer plusieurs critères d'accès (passerelle de similitudes) entre *pages* (élément unitaire,de connaissance) et de les combiner : critères temporels, géométriques et spatiaux, critères non représentables comme la proximité sémantique (voir figure 5).

Bilans

Du point de vue des sciences du numérique, les performances en TAL donnent un avantage à *Haruspex* qui améliore un algorithme existant (ANA) et rivalise avec des développements contemporains basés sur d'autres algorithmes avec des objectifs similaires (cf. 6).

Du point de vue de l'usage, la phase de création de graphe est originale et se combine avec la structure libre des graphes pour permettre une flexibilité d'usages et de domaines d'applications (cf. chapitre 4). Les tests ont jusqu'alors essentiellement été restreints à l'analyse de textes à dimension historique et intègrent faiblement la dimension patrimoniale.

Des résultats empiriquement intéressant pour l'historien (cf. chapitre 3, Application : Cartographie d'un corpus en Histoire de la Chimie du Solide) valident la méthode et son implémentation. La complexité des inférences de l'historien et la finesse des interactions quantitatif / qualitatif confirment une double hypothèse importante : il semble illusoire de vouloir permettre à quiconque de devenir historien d'un domaine sans connaissance préalable du corpus ; il semble impossible remplacer l'historien pour l'analyse historique.

Nous avons choisi d'adapter l'outil aux besoins et aux fonctionnement de l'historien, néanmoins on peut imaginer une évolution des pratiques de l'historien ayant une connaissance préalable de l'outil d'analyse numérique. Ainsi les corpus étudiés (chapitre 3 et 4) explicitement produits pour l'analyse numérique diffèrent de ceux qui pré-existaient à l'outil.

D'un point de vue théorique, il a été envisagé de déployer la méthode sur des applications à caractère patrimonial. Le développement conceptuel de ce déploiement (approche OLAP et « visite guidée mais libre ») est avancé.

Ces travaux de recherche ont permis de démontrer que les liens, entre les historiens ou acteurs du patrimoine d'une part et la recherche en sciences du numérique d'autre part, restent faibles malgré le développement massif des humanités numériques (cf. section 1.1.1). Les fantasmes perdurent (bibliothèque universelle, analyse sémantique, inférences historiques automatiques, etc.) et les représentations « vendeuses » (rendu 3D) tiennent le devant de la scène, tandis que les outils numériques ont une valeur ajoutée faible pour la recherche en histoire, pour l'historien et pour l'analyse du patrimoine.

Les choix radicaux de modélisation de données (graphe libre « no-schema ») pour les besoins de l'analyse historique et pour la levée des verrous de cette thèse ne doivent pas occulter l'importance des structures de données pour le patrimoine. Des expériences connexes ont été entreprises durant cette thèse, principalement autour de la problématique de l'archivage des données patrimoniales 3D. Ces expériences ont fait appel à des structures plus classiques et contraignantes (SQL).

Perspectives

La proposition permet d'améliorer certains processus de gestion de données textuelles existants (voir section 1.1.1), dans le domaine de la recherche de brevets, de bibliographie, de mise en relation de document techniques de référence (construction, médecine, etc) soit en les intégrant, soit en les remplaçant. D'autres domaines d'application que l'histoire et le patrimoine pourraient utiliser *Haruspex*, avec tout ou partie de la chaîne de traitement : TAL, graphe, navigation.

La forme de graphe proposée lie les documents entre eux par différents types de liens : sémantique à différentes échelles, temporels, géographiques. La production de graphes inverses (lieux associés par des documents) a été prototypé mais n'a jamais permis d'obtenir des résultats intéressants à analyser.

En référence au critères du patrimoine mondial UNESCO (cf. section 1.2.1), si la notion d'intégrité des données est traitée par cette thèse, la notion d'authenticité n'est pas abordée. En effet, la continuité d'usages est forte entre le corpus rassemblé par l'historien, sa connaissance de ce corpus et son analyse des données numériques.

La perspective majeure de cette thèse consiste à renforcer le lien avec le patrimoine en multipliant les expériences et en mettant à l'épreuve le cadre conceptuel proposé en section 4 (Application au patrimoine : les Salons Mauduit). Cela implique d'extrapoler le champ de l'analyse textuelle à des problématiques d'ordre géométrique (3D), d'ergonomie (navigation) et de muséologie (médiation des connaissances).

Appendices

Annexe A

Tableaux comparatif des métadonnées

Les métadonnées dans les tableaux suivants (liste d'un seul grand tableau) sont toutes utilisées dans le domaine de l'histoire et du patrimoine, certaines sont conçues pour ce domaine, d'autres sont issues d'un autre domaine.

Sans être exhaustive cette liste comprend un grand nombre de schémas du monde du patrimoine, recensés au cours de cette thèse.

Les comparaisons sont effectuées sur la base de plusieurs critères (colonnes) listés ici :

- Type :structure, valeurs ou contenu.
  - *Structure* concerne les schémas de métadonnées regroupant les questions, et l'organisation des ces questions entre elles, par exemple "titre", "type", "dimensions" ("hauteur", "largeur", "profondeur").
  - *Valeurs* concerne les schémas de métadonnées regroupant les réponses que l'on peut donner aux questions. Ce sont donc des listes de mots plus ou moins organisés (thésaurus). exemple : la question "type" pourra prendre une de ces valeurs "peinture", "3D nativement numérique", "sculpture"
  - *Contenu* concerne les schéma abstrait. Ils guident la mise en œuvre des schémas de structure et de contenu. Par exemple : pour un item de "type" : "3D nativement numérique" il faut aussi impérativement mentionner le nombre de polygones.
- Olds : indique d'anciennes versions du schéma
- Institution : indique l'institution porteuse du schéma.
- Patrimoine : indique si (oui : 1) le schéma a été développé pour le monde du patrimoine
- Finalité : indique les usages du schéma parmi : bibliothèque, musée, partage d'information, formalisation (et raisonnement sur les données), archivage.
- Classe : indique le nombre de classes (Entités, catégories d'item) contenu dans le schéma. Ce paramètre est plus spécifique aux ontologies.
- Propriété : indique le nombre de propriétés (descripteurs) prévues dans le schéma.
- Basé : indique si le schéma est basé sur un schéma précédent : c'est à dire s'il en reprend les classes ou propriétés.
- Nature : indique si le schéma est une ontologie, une organisation hiérarchique (taxonomie, thésaurus), ou un ensemble de descripteurs plat.
- Utilisation : indique si le schéma est actuellement parmi les plus utilisés.
- MAJ : indique la dernière date de mise à jour trouvée.

nom	type	olds	institution	patrimoine	finalité					classes	propriétés	base	orga.	utilisation	MAJ	commentaire
					Biblio-thèque	musée	Partage	Forma-lisation	Archi-vage							
ADS - Archaeology Data Service	structure		ADS (UK)	1	0	0	0	1	1	0	53	DC	hierarchical			
AGLS : Locator Service Metadata Standard	structure		Nat. Arch. Australia	0	0	0	1	0	0	0	44	DC	hierarchical			
ARCO (Augmented Representation of Cultural Objects)	structure		University of Bath	1	0	1	0	0	0				?	-1	###	obsolète
ATT	values		Getty	1	0	0	0	0	0				thésaurus			
BIBFRAME	structure		Lib. of Cong	0	1	0	1	1	0	68	653	MARC	ontology	1	###	remplace MARC orienté linkedOpenData
bibtex	structure			0	0	0	0	0	0	14	27	DC	hierarchical			
Carare2	structure	carare	3Dicons	1	0	1	1	0	0	?	?	carare	ontology		###	remplacé par EDM ? En parallèle ?
CCO: Cataloguing Cultural Objects	content		Getty	1	0	1	0	0	1	0	116	DC	hierarchical			
CDWA	structure		Getty	1	0	1	0	0	0	0	532		taxonomy			
CDWA Lite	structure		Getty	1	0	1	0	0	0	0	22		hierarchical			
CIDOC-CRM	structure			1	0	1	0	1	0	90	149		ontology			
CIDOC-CRM extended	structure			1	0	1	0	1	0	NC	NC		ontology			
CINES	structure			0	0	0	0	0	1	0	34	DC	hierarchical			

nom	type	olds	institution	patrimoine	finalité					classes	propriétés	base	orga.	utilisation	MAJ	commentaire
					Biblio-thèque	musée	Partage	Forma-lisation	Archi-vage							
CONA: Cultural Object Name Authority	values			1	0	1	0	0	0	NC	NC		hierarchical	1		CDWA+ CCO implementation for authority indexes
Core Data Index to Historic Buildings and Monuments	values		ICOM	1	0	0	0	0	0	NC	NC		?		###	obsolète, impasse
Core Data Standard for Archaeological Sites and Monuments	values		ICOM	1	0	0	0		0	NC	NC		?		###	obsolète, impasse
CRMDIG : a model for provenance metadata	structure		ICS	1	0	0	0	0	1	16	41	CIDOC-CRM	ontology		###	documente la provenance des objets virtuels (3D numérique)
DDI : Data Documentation Initiative (life Cycle)	structure			0	0	0	0		0	16	118		ontology			
Dublin Core Extended (dcterms)	structure			1	1	1	0		0	34	55	DC	taxonomy	1		
Dublin Core minimal (dce)	structure			1	1	1	0		1	0	15		hierarchical (flat)	1		
EAC : Encoded Archival Context	structure		Lib. of Cong	1	0	0	0		1	10	29	EAD	ontology			

nom	type	olds	institution	patrimoine	finalité					classes	propriétés	base	orga.	utilisation	MAJ	commentaire
					Biblio-thèque	musée	Partage	Forma-lisation	Archi-vage							
EAD : Encoded Archival Description	structure		Lib. of Cong	1	0	0	0		1	0	146		taxonomy			
EBU-Core : radio and télévision content	structure			0	1	0	0		0				hierarchical			
Europeana Data Model	structure	Carare MIDAS		1	0	1	1		0	11	35		ontology	1		
EXIF	structure			0	0	0	0		1	0	482		hierarchical		###	
FOAF	structure			0	0	0	0		0	13	62		ontology			
GEM: Gateway to Educational Materials	structure		DublinCore	0	0	0	0	1	1	0	23	DC		0		décrit des ressources pédagogiques via extended DublinCore
Geo: WGS84 (LOV)	structure/values			0	0	0	0		0	2	5		hierarchical (flat)			
IAFA: metadata for internet description (ftp, gopher)	?			0	0	0	0		?	NC	NC		?	-1	###	obsolète, impasse. Note ancetre de l'URI ?

nom	type	olds	institution	patrimoine	finalité					classes	propriétés	base	orga.	utilisation	MAJ	commentaire
					Biblio-thèque	musée	Partage	Forma-lisation	Archi-vage							
IFC4 ISO 16739 (BIM: Building information model)	structure		buildingsSMART (ex International Alliance for Interoperability)	0	0	0	0	1	0	###	2234		ontology	1	###	http://www.buildingsmart-tech.org/ifc/IFC4/final/html/index.htm
IPTC IIM IMN	structure	[largely overtaken by XMP]	International Press Telecommunications Council (IPTC)	0	0	0	1		0	4	120		hierarchical		###	obsolète, impasse
ISO 19115: Geospatial Metadata	structure		FGDC	0	0	0	0		0	40	300 ++		ontology	1	###	
ISO 8601 (time description)	content			0	0	0	0	1	1	2	0		-			
JATS: Journal Article Tag suite	structure		NISO: National Information Standard Organisation	0	0	0	0		0							

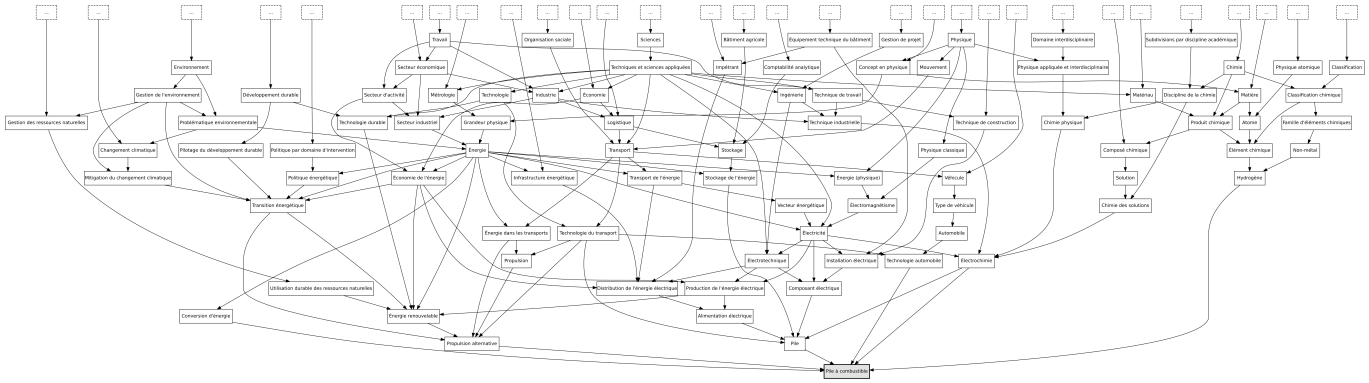
nom	type	olds	institution	patrimoine	finalité					classes	propriétés	base	orga.	utilisation	MAJ	commentaire
					Biblio-thèque	musée	Partage	Forma-lisation	Archi-vage							
LIDO (Lightweight Information Describing Object)	structure			1	0	1	0		0	7	~300		ontology		###	
MAB: Machinele Austauschformat fur Bibliotheken	structure		Deutsche National bibliothek	1	1	0	0		0	NC	NC				###	obsolète, impasse
MADS : Metadata Authority Description Schema – MARC21 compatible XML format	structure		Lib. of Cong	1	1	0	0		0	22	56					fiche d'autorité
MARC (Machine Readable Cataloging) – MARC 21	structure		Lib. of Cong	1	1	0	0		1	4	150					
MEI: Music Encoding Initiative	structure			1	0	0	0		0							
METS: Metadata Encoding and Transmission Standard	structure		Lib. of Cong	0	1	0	0	1	1	40			hierarchical		###	

Annexe B

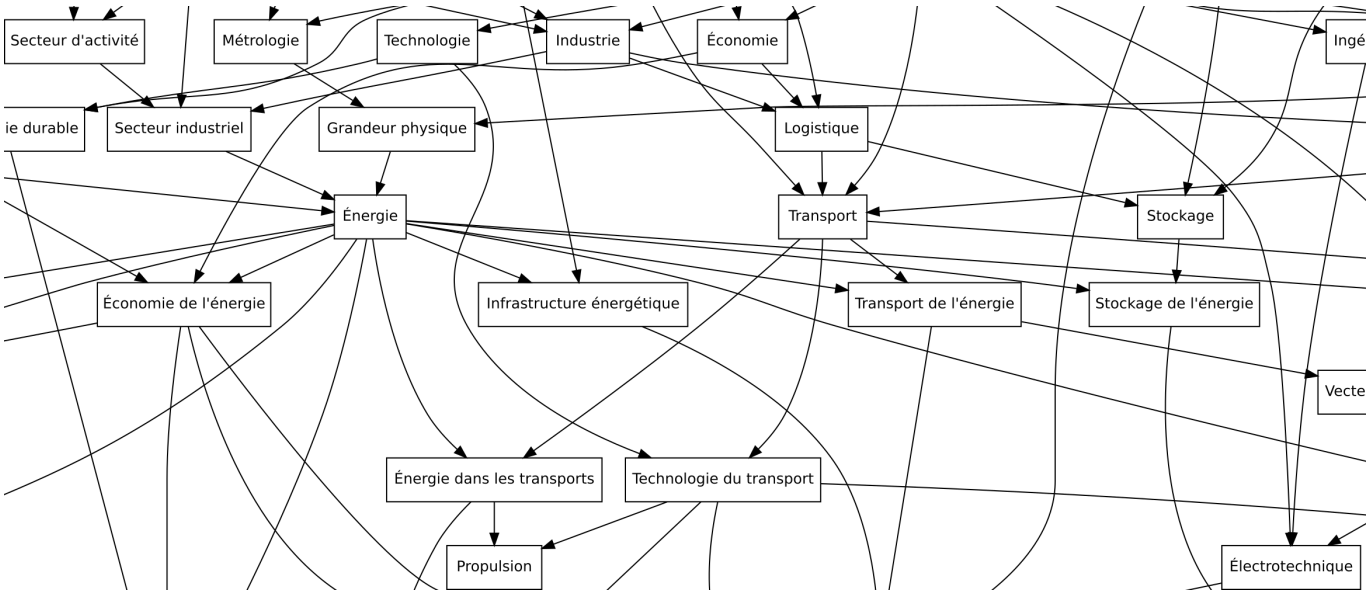
Portails et catégories Wikipédia

Wikipédia (MediaWiki) permet des structurations équivalentes dans toutes les langues. Mais dans les faits, les pratiques sont différentes. Les principaux outils pour structurer les contenus sont les catégories et les portails. Nous présentons ici ces 2 fonctionnements, ainsi qu'une étude de cas sur un corpus.

Les catégories forment un réseau hiérarchisé de connaissances adaptable et précis. En conséquence, ce réseau est très foisonnant. L'exemple pris au hasard (issus des exemples suivants et du chapitre "Proposition scientifique : Haruspex" (2) traitant des sorties de *ANA+*) des liens entre les catégories parentes de la catégorie « pile à combustible » illustre cette idée (figure B.1).



(a) vue d'ensemble



(b) zoom

FIGURE B.1 – Représentation graphique des liens entre catégories parentes de la catégorie « pile à combustible ».

Les portails sont organisés de manière moins foisonnante, ils sont moins nombreux et contiennent en moyenne davantage d'articles. Le taille est plus homogène (voir table B.1). S'il y a certainement beaucoup d'information à exploiter dans les catégories, les portails semblent plus faciles d'accès.

Portail marqué dans x articles	nombre de portails
$x \leq 10$	2
$10 \geq x \leq 99$	83
$100 \geq x \leq 10^3$	788
$10^3 \geq x \leq 10^4$	641
$10^4 \geq x \leq 10^5$	134
$x \geq 10^5$	15

TABLE B.1 – Répartition des 1663 portails Wikipédia Français en fonction du nombre d'articles marqués

Les pratiques ne sont pas identiques en anglais et en français : les anglais utilisent beaucoup les catégories, tandis que les français utilisent les catégories et les portails. L'usage du portail est systématique (ou presque) sur Wikipédia Français, et rare en anglais.

Ainsi, dans *Haruspex*, pour un corpus anglais on remonte l'arbre des catégories jusqu'à obtenir un ensemble de catégories parentes de plus de 1000 articles. Par exemple, en français avec le cas de l'article « pile à combustible » de la catégorie éponyme (16 art.), un seul niveau suffirait. Ce niveau récupérerait les catégories *électrochimie* (234 art.), *Technologie automobile* (653 art.) , *Pile* (inclus dans électrochimie), *Propulsion alternative* (800 art.), *conversion d'énergie* (326 art.) et *hydrogène* (>1000 art.).

Forme la plus occurrente	**Catégories (fr)**	**Catégories (en)**
Thèse de troisième cycle	Doctorat en France	Academic degrees of the USA Doctoral degrees
Pile à combustible	Pile à combustible	English invention Fuel cells Energy conversion Hydrogen economy
Microscopie électronique à transmission	Science des matériaux Histologie Microscope électronique	Electron beam Scientifique techniques Electron microscopie

TABLE B.2 – Exemple de différences de catégories entre Wikipédia français et anglais.

Forme la plus occurrente	**Portails (fr)**	**Portails (en)**
Thèse de troisième cycle	France Éducation	Academic degrees
Pile à combustible	Chimie Énergie Électricité et électronique	-
Microscopie électronique à transmission	Physique Sciences des matériaux Électricité et électronique	-

TABLE B.3 – Exemple de différences de portails entre Wikipédia français et anglais.

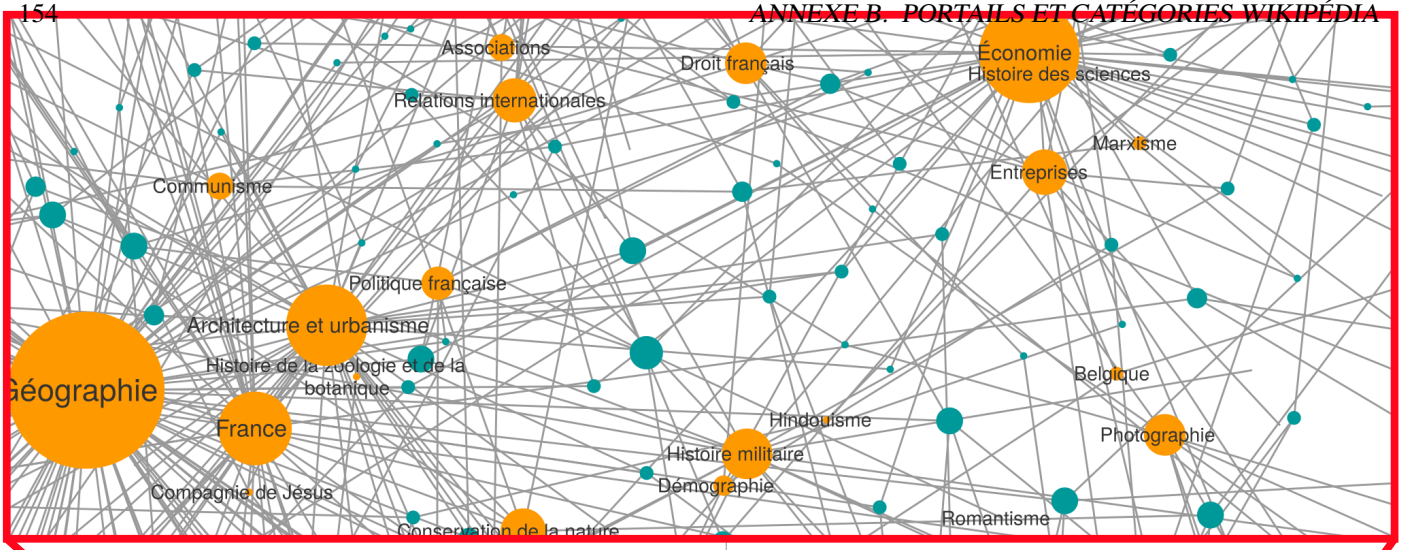
Il ne semble pas exister de hiérarchie entre les portails de Wikipédia. Mais cela n'est nécessaire pour nos besoins, en effet cette information de portail est déjà suffisamment « méta » pour renseigner et regrouper plusieurs pages. En effet les 1663 portails concernent 1 899 486 articles Wikipédia en français (le 17 août 2017). Un seuil constant d'environ 1000 articles sans portail existe depuis 2010. Il serait intéressant de calculer le nombre d'article marqués par 1, 2 ou 3 et plus portails. Sachant qu'un article est affilié à plus d'1 portail en moyenne, on comprend que le nombre d'articles moyen par portail soit supérieur à 2000.

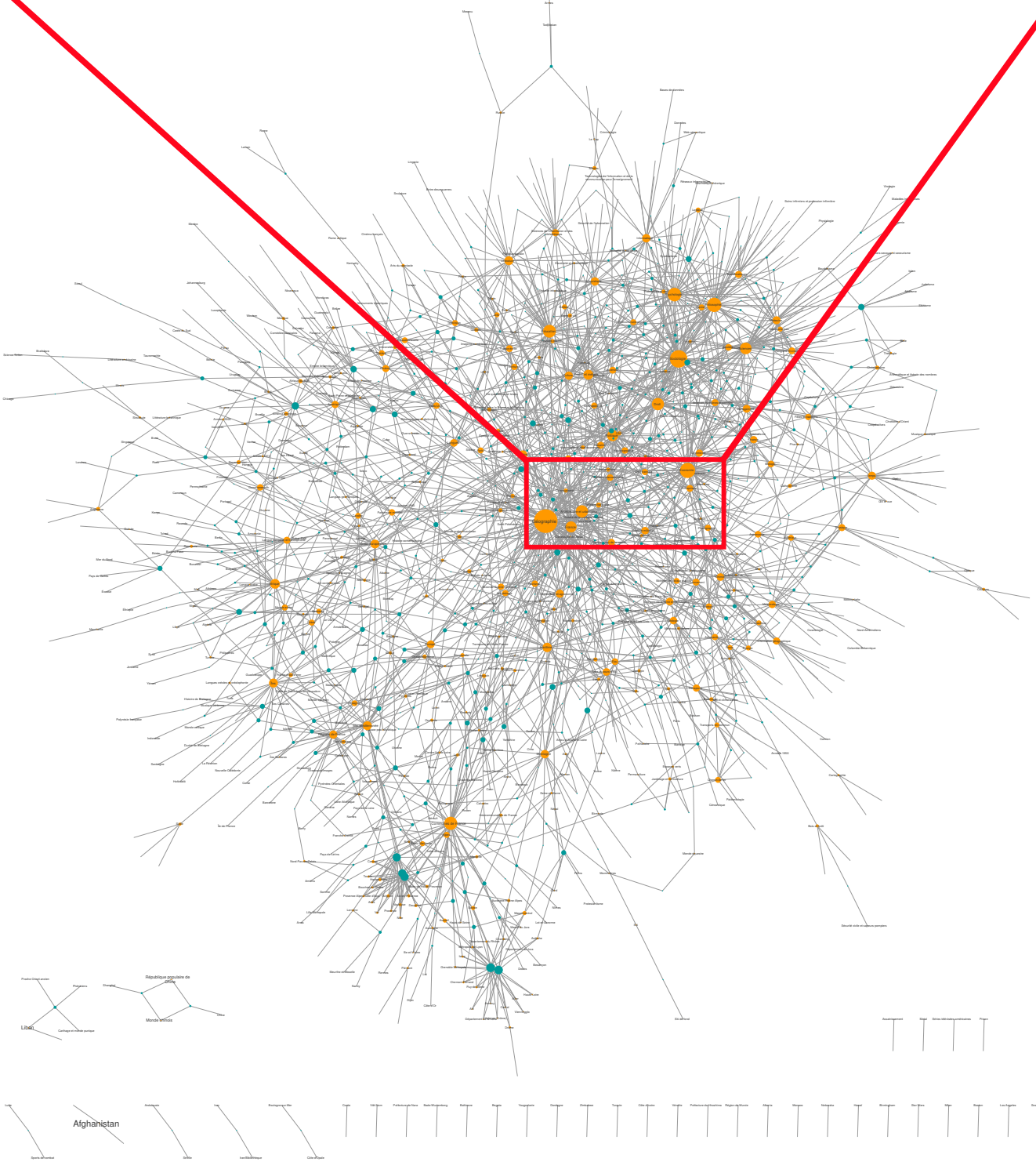
1 Analyse en lien avec un corpus

Dans le cas de corpus spécifique à la géographie française de l'après-guerre en relation avec la notion de « terrain du géographe » : Le tableau B.4 présente des analyses des catégories Wikipédia sur ce corpus :

Les portails vecteurs de liens. Il s'avère que le nombre de portails contenant 1 seule *page* (correspondant à un article Wikipédia) est élevé (il y en a 240 sur 598 soit plus d' $\frac{1}{3}$). Ces portails sont inutiles : ils ne créent aucun lien parmi nos expressions : ils ne

permettent pas de construire des catégories regroupant plusieurs *pages*. Mais puisque les pages peuvent contenir plusieurs portails Wikipédia, les *pages* concernées ne sont pas nécessairement isolée. En effet, pour qu'une page soit isolée, il faudrait qu'elle contienne 1 seul portail et que ce portail ne concerne que cette page. Ce phénomène est rare. D'autre phénomènes plus complexe peuvent isoler des sous-groupes : plusieurs *pages* contiennent un portail commun, mais aucune de ces pages ne contient d'autre portails en lien avec le reste des articles (voir "Les exclus du réseau" (1)). En page 154, nous traçons le graphe des liens entre pages et portails, pour expliquer la capacité des portails à agir comme des classes pour les expressions extraites (et correspondant à des articles Wikipédia).


Associations
Droit français
Économie
Histoire des sciences
Relations internationales
Marxisme
Entreprises
Communisme
Politique française
Architecture et urbanisme
Géographie
Histoire de la zoologie et de la botanique
Belgique
France
Hindouisme
Photographie
Histoire militaire
Démographie
Compagnie de Jésus
Romantisme
Conservation de la nature


Afghanistan

(a) Statistiques générales

nombre d'expressions extraites	1987
expression avec article Wiki	1063
expression avec article Wiki et portails	1063
nombre de portails différents	598
nombre moyen de portails par article	2.53
écart type du nb moyen de p. par art.	1.85

(b) Répartition des portails

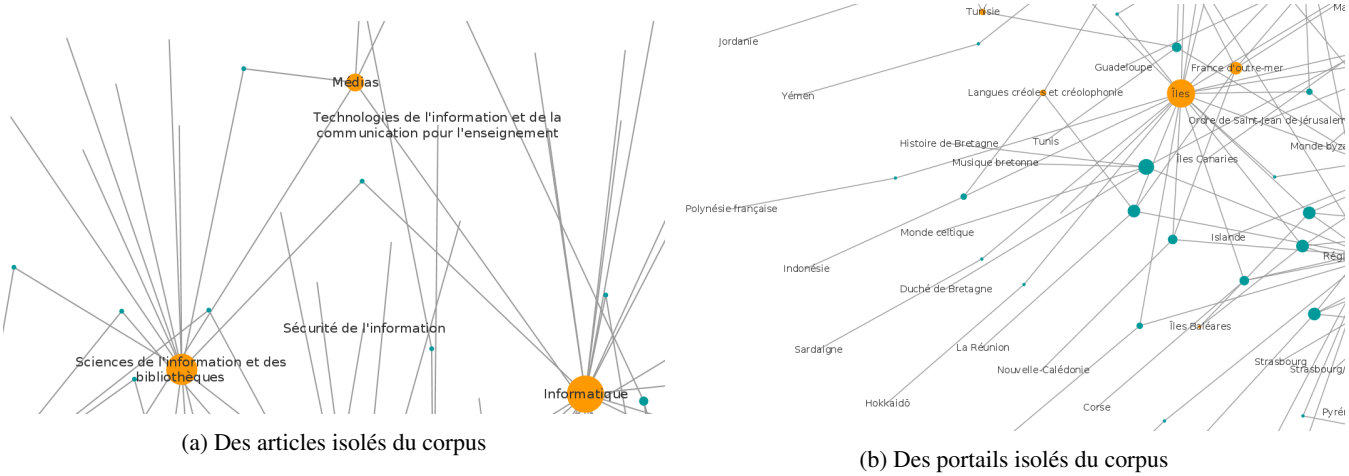
portails concernant 1 seul article Wiki	240
portails concernant 2 articles Wiki	107
portails concernant plus de 3 articles	251
portails concernant plus de 10 articles	25
portails concernant plus de 50 articles	7

TABLE B.4 – Analyse statistique des portails d'un corpus d'histoire de la géographie en France

Le réseau central. Le graphe total (page 154) est construit à la manière du graphe de *Nantes1900* (voir l'algorithme 1, section 4.1.4). Les nœuds sont soit des portails, soit des articles. Chaque lien correspond à un portail marqué dans un article. Il n'y a donc que des liens entre article et portail, jamais entre portails ou entre articles directement. C'est une relation *many to many* : un article peut avoir plusieurs portails, un portails peut avoir plusieurs articles. Ce graphe est produit en appliquant une force centrifuge aux nœuds. Ainsi les nœuds les moins liés sont les plus extérieurs. Les portails sont représentés en jaune, les articles en bleu. Les titres articles ne sont pas représentés pour la lisibilité. La taille d'un nœud (portail ou article) est proportionnelle à son degré de connectivité (voir section 4 pour des détails sur les mesures possibles). La mesure de degré est choisie car elle est adaptée à notre intérêt pour la capacité de connectivité des portails.

Sur ce graphe on remarque qu'un réseau central concerne la plupart des pages liées entre elles. Il existe donc bien des regroupements de pages autour de portails. Certains portails sont centraux et regroupent de très nombreux articles. Parmi ces portails on trouve en tête dans l'ordre (sans surprise) : « géographie » (99), « sociologie » (67), « économie » (55), « psychologie » (51), « philosophie » (51), « éducation » (43), etc. Le portail « communes de France » (57) fait exception : il est marginal. En effet les articles qui y sont liés sont faiblement reliés au reste du réseau. Cela nous amène à la périphérie du réseau.

La périphérie du réseau. La couronne périphérique du graphe (page 154) est composé de portails inutiles (parmi les 240 portails liés à une page seulement) et de pages liées à un portail assez marginal seulement. On remarque que les **portails isolés** sont souvent des portails trop précis par rapport à l'étude menée, par exemple les portails « monde celtique », « musique bretonne », « Histoire de Bretagne », « duché de Bretagne » ne réfèrent qu'à un seul article. À l'inverse les **articles isolés** sont des notions précises qui réfèrent à un portail générique commun mais hors du spectre du corpus. C'est le cas des portails « informatique » et « sciences de l'information » qui regroupent de nombreux articles effectivement évoqués dans le corpus, mais sans lien direct avec les portails centraux du corpus.



(a) Des articles isolés du corpus

(b) Des portails isolés du corpus

FIGURE B.2 – Typologie de la couronne périphérique du graphe

Les exclus du réseau. Certaines *pages* n'ont qu'un seul portail qui n'est marqué que dans cette *page* (ce sont les duos isolés en bas du graphe page 154). Ces portails reçoivent une mauvaise évaluation de *indicateur de confiance* (voir section 4.2). Ce sont souvent des portails hors-sujet (comme « Star Wars », ou « scoutisme »), parfois des portails mal-formés (« Iran/Bibliothèque »). Il y a des cas particulier d'articles marqués par des portails en communs mais non reliés au reste du réseau. C'est le cas d'article sur des thématiques chinoises qui sembles déconnectées des autres portails ou articles (figure B.3)).

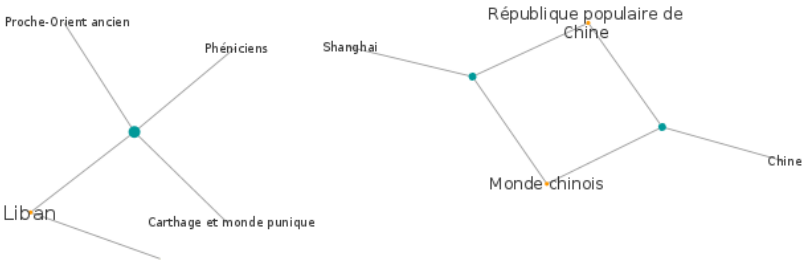


FIGURE B.3 – Exemples de sous-réseaux déconnectés du réseau principal

2 Conclusion

Pour les corpus Français, les portails permettent de relier les pages entre elles sur un mode compréhensible pour l'humain. Certains portails sont plus centraux que d'autres. L'amplitude limitée du corpus donne à l'ensemble une cohérence qui mériterait presque une analyse complémentaire par l'historien. Le réseau dessiné ici est une des couches de *Haruspex*. En effet les expressions (skip-gram) extraites embarquent leur(s) portail(s). Il est alors possible d'accéder aux données (requêter) en utilisant ces portails et les relations qu'ils entretiennent.

Aucun analyse similaire n'a été menée pour les catégories en anglais. L'approche serait différente, car les catégories établissent des liens directs et typés (hiérarchisés) entre elles. Ce n'est pas le cas des portails.

Annexe C

Exemples de fonctionnement de *ANA+*

Les différentes étapes sont décrites par la figure C.1 (rappel). Nous présentons ici les activités A22, A23 et A24.

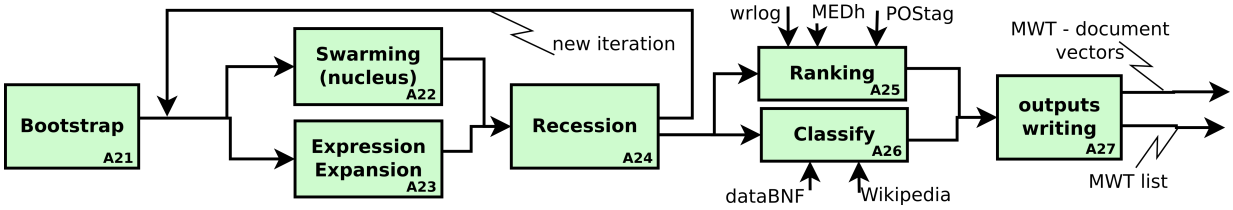


FIGURE C.1 – Schéma SADT décrivant les activité contenu dans l'activité A2 : Extraction d'expressions-clés

Les lemmes mentionnés dans le `bootstrap` sont les `candidats` de la première itération. Par convention nous soulignons les candidats dans les exemples. Rappel : les candidats sont amenées à être modifiés par *Haruspex* au cours du processus : création, augmentation de le longueur de la chaîne de caractères, désactivation...

1 A22 : Essaimage

Il s'agit de repérer les lemmes systématiquement associés à d'autres `candidats` éventuellement via un mot de liaison. Les mots de liaisons sont {de, du, à, en, des}. Le `candidat` associé au lemme peut varier. Après validation du lemme, toutes les occurrences du lemme deviennent à leur tour des `candidats`, même isolées les occurrences qui ne côtoyait pas de candidat.

(1). Bootstrap. si l'on a les candidats {microscope, molécule, champs}

(2). Fenêtres autour de candidats existants. Autour des `candidats` existants, des fenêtres sont construites. Les fenêtres sont ensuite regroupées par lemme contenu. Une fenêtre peut être utilisée plusieurs fois, pour les différents lemmes qu'elle contient. Nous montrons ici les fenêtres valides pour le lemme « fluorescence » (la validité d'une fenêtre est définie en "Essaimage" (3.2.1)) :

1. microscope de fluorescence [...]
2. microscope optique à fluorescence [...]
3. fluorescence des molécules [...]
4. champs à fluorescence [...]
5. fluorescence d'une molécule [...]
6. microscope à contraste de fluorescence [...]
7. champs de fluorescence [...]

(3). Seuil et validation du nouveau candidat potentiel. Nous prenons le calcul défini en section 3.2.1 :

$$\text{Essaimage}(l_i) = \left(\frac{\alpha}{S}\frac{F_v(l_i)}{C(l_i)} + \frac{\beta}{S}\frac{F_v(l_i)}{L(l_i)}\right) \times \frac{F_v(l_i)}{P(l_i)}$$

avec

— $F_v(l_i)$ le nombre de fenêtres valides d'un lemme. Dans notre cas, l'exemple porte sur le lemme « fluorescence » (le lemme « étude » pourrait faire une autre exemple). $F_v(l_i) = 7$

— $C(l_i)$ le nombre de `candidats` différents. Dans notre cas : $C(l_i) = 3$

— $L(l_i)$ le nombre de types de `linkwords` différents, l'absence de `linkwords` est considéré comme un type. Les `linkwords` {de, du, des} sont équivalent ; on donc $L(l_i) = 2$
— $P(l_i)$ le cumul des position occupées dans les fenêtres : 1 pour la position la plus proche du `candidat`, 2 pour la seconde position. Dans notre cas : $P(l_i) = 1 + 2 + 1 + 1 + 1 + 2 + 1 = 9$
— $S = \alpha + \beta$
— α et β à fixer. Des valeurs indicatives sont $\alpha = 10$, $\beta = 3$.

Dans notre cas :

$$\text{Essaimage}(fluorescence) = \left(\frac{10}{13} \times \frac{7}{3} + \frac{3}{13} \times \frac{7}{2}\right) \times \frac{7}{9} = 2.42$$

alors, si les seuils le permettent (ici 2 par exemple), nous aurons un nouveau candidat : « fluorescence ».
Note : ce seuil de 2 est bas, il correspondrait à un corpus peu volumineux.

(4). Essaimage du nouveau candidat. Ce nouveau `candidat` « fluorescence » sera repéré partout dans le texte (4 nouvelles occurrences repérées sans `candidat` associé) :

1. Le microscope de fluorescence [...]
2. microscope optique à fluorescence [...]
3. fluorescence de la chlorophylle [...]
4. utilise la fluorescence des molécules [...]
5. fluorescence de la chlorophylle [...]
6. pour l'étude des champs à fluorescence [...]
7. fluorescence de résonance [...]
8. fluorescence d'une molécule [...]
9. fluorescence de la chlorophylle [...]
10. microscope à contraste de fluorescence [...]
11. champs de fluorescence [...]

(5). Nouvelle boucle. Dans *ANA+* plusieurs boucle d'essaimage sont réalisées avant chaque développement. On peut alors recommencer à essaimer avec ce nouveau set de candidats
{microscope, molécules, fluorescence, champs}.

Ce qui amènerait (peut-être) à trouver « résonance » dans notre exemple.

2 A23 : Développement (expansion et expression)

Il s'agit de repérer des motifs récurrents pour agrandir les candidats itérativement. Le nouveau candidat formé regroupe alors plusieurs composants.

— Pour les expansions, il existe 2 motifs différents :
  1. candidat$_1$ non-candidat$_2$
  2. non-candidat$_3$ candidat$_4$
— Pour les expressions, il n'existe qu'un seul motif :
  candidat$_1$ candidat$_2$

Dans tous les cas un lemme peut être intercalé dans la fenêtre, il s'agit alors d'un skip-gram. Il faut alors prêter attention aux potentielles expansions incluses dans une expansion ou expression skip-gram.

2.1 Expansion

(1). Candidats existants à l'état 1 Suite à la première phase "A22 : Essaimage" (1), les candidats suivants sont identifiés {microscope, fluorescence, molécule, champs}

(2). Les fenêtres Les fenêtres de taille 2 sont construites autours des `candidats` issus de "A22 : Essaimage" (1). L'exemple ici présente les fenêtres autours du `candidat` « fluorescence ».

La notation suivante est utilisée : Les mots vides (`emptywords`) sont ~~barrés~~ ; un pipe rouge (|) marque le bord gauche de la fenêtre de gauche et un bleu (|) à droite. Le bord des fenêtres est fortement influencé par la troncature (un autre `candidat` ou un `stopwords`). Ces troncatures sont marquées par un double pipe (||). Pour l'exemple on conserve l'ensemble de la fenêtre (partie tronquées).

On construit alors dans le texte les fenêtres suivantes :

1. Europe. ~~Le~~ microscope|| ~~de~~ fluorescence ~~est~~ ~~un~~ outil optique | [...]
2. ~~L'~~usage ~~du~~ microscope|| optique ~~à~~ fluorescence permet ~~des~~ mesures | [...]

3. Luciférase. || ~~Les~~ métriques ~~de~~ fluorescence mesurables ~~pour~~ ~~la~~ chlorophylle | [...]
4. | capteur utilise ~~la~~ fluorescence ~~des~~ || molécules radiochimiques [...]
5. |~~L'~~étude systématique ~~de~~ ~~la~~ fluorescence ~~de~~ ~~la~~ chlorophylle contaminée | [...]
6. préparation ~~pour~~ ~~les~~ champs|| à fluorescence induite ~~en~~ laboratoire | [...]
7. transfection. ||~~L'~~utilisation ~~de~~ ~~la~~ fluorescence ~~de~~ résonance ~~se~~ généralise | [...]
8. ~~Lorsqu'elle~~ ~~est~~ | autonome ~~la~~ propriété ~~de~~ fluorescence ~~d'une~~ ||molécule ~~est~~ ~~dite~~ primaire [...]
9. | identification ~~de~~ zone ~~de~~ fluorescence ~~de~~ ~~la~~ chlorophylle répartie | [...]
10. microscope|| à contraste ~~de~~ fluorescence implique ~~les~~ fluo-chromes | [...]
11. ~~L'~~excitation ~~des~~ champs || ~~de~~ fluorescence induite ~~pour~~ ~~la~~ résonance| [...]

(3). Validation et formation d'un nouveau candidat complexe Pour chaque candidat, on compte les occurrences de chaque lemme présent dans les fenêtres. Ici avec un **seuil de 3** occurrences minimum, l'expansion : fluorescence de la chlorophylle peut être formée. Cette expansion est repérée dans les fenêtres n°3, n°5 et n°9. Dont une fois en skip-gram (fenêtre n°3). On parle de comportement *greedy* (glouton). Il faut pour cela que le motif « fluorescence mesurables » ne soit pas repéré 2 autres fois dans le texte. Le seuil d'expansion (3) aurait sinon été dépassé et l'expansion aurait été prioritaire.

2.2 Expressions

Le mécanisme de formation des expressions est globalement identique aux expansions. Les modifications sont (1) les motifs recherchés sont `candidat- candidat` et (2) le seuil est plus bas.

Par exemple, on construit les fenêtres valides suivantes autour du `candidat` « fluorescence ». Les fenêtres ne sont pas construites après les expansion, mais directement à partir des résultats de "A22 : Essaimage" (1). Une fenêtre valide doit contenir 1 autre `candidat` et est tronquée immédiatement après :

1. microscope ~~de~~ fluorescence [...]
2. microscope optique ~~à~~ fluorescence [...]
3. *non valide : 1 seul* `candidat`
4. fluorescence ~~des~~ molécules [...]
5. *non valide : 1 seul* `candidat`
6. champs ~~à~~ fluorescence [...]
7. *non valide : 1 seul* `candidat`
8. fluorescence ~~d'une~~ molécule
9. *non valide : 1 seul* `candidat`
10. microscope ~~à~~ contraste ~~de~~ fluorescence [...]
11. champs ~~de~~ fluorescence

L'expression « microscope fluorescence » occurre 3 fois, dont 2 fois sous forme d'un skip-gram (fenêtre n°2 et n°9). Il faut donc vérifier que les lemmes contenus dans les skip-gram ne peuvent pas former d'expansion. Ici avec 2 skip-gram il n'y a pas de risque (sous le seuil de 3 pour les expansions). De plus les expressions « fluorescence molécule » et « champs de fluorescence » occurrent 2 fois chacun, jamais sous forme de skip-gram.

Avec un **seuil à 2**, les expressions sont valides. S'il y a plusieurs écritures pour ce nouveau `candidat`, la forme retenue sera la forme la plus occurrente après lemmatisation « microscope à fluorescence », « champs de fluorescence » et fluorescence de molécule.

2.3 Résultats des développements

Après expression et expansion nous obtenons alors les résultats suivant. Une nouvelle notation est introduite : un mot inclus dans un skip-gram est grisé (il ne sera plus considéré). Pour l'exemple, nous ne représentons que les fenêtres pré-cités (toujours centrées sur « fluorescence »).

1. Europe. ~~Le~~ microscope à fluorescence ~~est~~ ~~un~~ outil optique [...]
2. ~~L'~~usage ~~du~~ microscope optique à fluorescence permet ~~des~~ mesures [...]
3. Luciférase. ~~Les~~ métriques ~~de~~ fluorescence mesurables de chlorophylle [...]
4. capteur utilise ~~la~~ fluorescence des molécules radiochimiques [...]
5. ~~L'~~étude systématique ~~de~~ ~~la~~ fluorescence de la chlorophylle contaminée [...]
6. préparation ~~pour~~ ~~les~~ champs à fluorescence induite ~~en~~ laboratoire [...]
7. transfection. ~~L'~~utilisation ~~de~~ ~~la~~ fluorescence ~~de~~ résonance ~~se~~ généralise [...]
8. ~~Lorsqu'elle~~ ~~est~~ autonome ~~la~~ propriété ~~de~~ fluorescence d'une molécule ~~est~~ ~~dite~~ primaire [...]
9. identification ~~de~~ zone ~~de~~ fluorescence de la chlorophylle répartie [...]
10. microscope à contraste de fluorescence implique ~~les~~ fluo-chromes [...]
11. ~~L'~~excitation ~~des~~ champs de fluorescence induite ~~pour~~ ~~la~~ résonance [...]

2.4 Développements et gestion de priorités

Les mécanismes de développement (“Expansion” (2.1) et “Expressions” (2.2)) sont réalisés en parallèle. Ils utilisent la même base de texte et candidats (issus de “A22 : Essaimage” (1)) mais ne forment pas les mêmes fenêtres.

À cause de la construction en skip-gram, certaines fenêtres peuvent êtres en conflits : lorsqu'une expansion est imbriquée dans une expression ou dans une autre expansion. *Haruspex* ne valide définitivement les expressions et expansions qu'après résolution des conflits éventuels.

2 cas se présentent :

Expansion strictement incluse ou autonome En cas de d'expansion strictement incluse ou autonome dans une expression ou une autre expansion, c'est à dire lorsque l'expression n'apparaît que incluse ou autonome et jamais incluant (par skip-gram) une autre expansion. Dans ce cas, la priorité est donnée à l'expansion incluse. Ce comportement suppose qu'il faut construire les expansions élément par élément en évitant les skip-grams si possible.

Par exemple, avec le candidat « Atelier » déjà connu et le texte contient 3 occurrences de « Atelier et Chantier de Bretagne » et 5 « Atelier et Chantier », alors l'expression « Atelier et Chantier » (8 occurrences) est favorisée sur le skip-gram « Atelier de Bretagne ». la prochaine itération donnera « Atelier et Chantier de Bretagne » (il y en aura 3).

Expansions intriquées Si les expansions sont intriquées : incluses l'une dans l'autre. C'est à dire que l'expansion incluse dans un développement est parfois à son tour *greedy* : elle inclut des lemmes par ailleurs. Le problème vient alors du mécanisme *greedy*. Dans ce cas, le comportement par défaut donne la priorité à la moins occurrente avec un comportement sans skip-gram (*non-greedy*). Ce dernier comportement suppose que l'expansion la plus occurrente pourra toujours se former, même après que certaines occurrences aient été consommées par l'expansion plus rare.

Par exemple, avec le candidat « Atelier » déjà connu ; le texte contient 3 occurrences de « Atelier et Chantier de Bretagne » ; 5 « Atelier et Chantier » ; 2 « Atelier de Bretagne en chantier » et 1 atelier de Bretagne, on compte :

(a) Occurrences d'expansions intriquées

	occ.
Atelier et chantier de Bretagne	3
Atelier naval de Bretagne	8
Atelier et chantier	5
Atelier et chantier de la Loire	6
Atelier de Bretagne en chantier	2
Atelier de Bretagne	1

(b) Décompte des expansions possibles selon 2 comportements

	greedy	*non-greedy*
atelier Bretagne	3+2+1+8	2+1
ateliers chantiers	3+5+2+6	5+3+6

TABLE C.1 – Priorité pour développements intriquées

D'après le tableau C.1, la priorité sera donné à la construction *non-greedy* (sans skip-gram) de l'expansion la plus rare {atelier Bretagne}.

3 A24. Récession

Il s'agit de retro-transformer les candidats dont le nombre d'occurrence est inférieur à un seuil. Cela apparaît lorsqu'un nouveau candidat se forme via “A23 : Développement (expansion et expression)” (2) et que les candidats d'origine (qui ont formé l'expression ou expansion) ne sont plus suffisamment occurrents. La récession est complexe puisqu'elle construit un arbre de récession du `candidat` vers un état stable antérieur.

État à l'issue des développements. Le texte, à l'issue des opérations précédentes est composé de `candidats` parfois simples, parfois composés, parfois composés en skip-gram. Les composants des `candidats` issus de développement ne comptent plus individuellement, ils sont « consommés » par la composition. Par exemple microscope à fluorescence ne compte pas pour les `candidat` microscope ou fluorescence individuellement.

Vérification des candidats. On compte les candidats à l'issue de la phase de développement. Si le nombre d'occurrences du candidat est inférieur à un seuil, alors le candidat est désactivé. Dans notre exemple, le seuil est de 3.
NOTE : le décompte ci-dessous concerne tout le texte. Les `candidats` {microscope, molécule, champs} occurrent ailleurs que dans les fenêtres du `candidat` « fluorescence » (seules fenêtres représentées précédemment).

candidat	dans nos fenêtres	par ailleurs	valide
microscope à fluorescence	3	0	oui
fluorescence de chlorophylle	3	0	oui
fluorescence de molécule	2	0	oui
champs de fluorescence	2	0	oui
fluorescence	1	0	non
champs	0	2	non
molécules	0	7	oui
microscope	0	3	oui

TABLE C.3 – Nombre d'occurrence et validité des candidats à la récession

Désactivation des candidats sous le seuil. Les `candidats` désactivés régressent. Un `candidat` qui régresse revient à une forme antérieure stable : lemme simple ou `candidat` moins composé. Ci-dessous ils sont soulignés en gris (fenêtre n°7). Dans notre exemple précédent, l'ensemble de candidats deviendrait :
{molécule, champs, champs à fluorescence, fluorescence de molécule, fluorescence de la chlorophylle, microscope à fluorescence}. Toute les occurrences non composées de « champs » et de « fluorescence » sont désactivées, redeviennent de simples lemmes. On obtiendrait le résultat suivant :

1. Europe. ~~Le~~ microscope à fluorescence ~~est un~~ outil optique [...]
2. ~~L'~~usage ~~du~~ microscope optique à fluorescence permet ~~des~~ mesures [...]
3. Luciférase. ~~Les~~ métriques ~~de~~ fluorescence mesurables de chlorophylle activée ou contaminée [...]
4. capteur utilise ~~la~~ fluorescence des molécules radiochimiques [...]
5. ~~L'~~étude systématique ~~de la~~ fluorescence de la chlorophylle contaminée [...]
6. préparation ~~pour les~~ champs à fluorescence induite ~~en~~ laboratoire [...]
7. transfection. ~~L'~~utilisation ~~de la~~ fluorescence ~~de~~ résonance ~~se~~ généralise [...]
8. ~~Lorsqu'elle est~~ autonome ~~la~~ propriété ~~de~~ fluorescence d'une molécule ~~est dite~~ primaire [...]
9. identification ~~de~~ zone de fluorescence de la chlorophylle répartie [...]
10. microscope à contraste de fluorescence implique ~~les~~ fluo-chromes [...]
11. ~~L'~~excitation ~~des~~ champs de fluorescence induite ~~pour la~~ résonance [...]

Le mécanisme de désactivation diffère d'une destruction totale. En effet, il permet un re-déploiement rapide de `candidat` désactivés sous certaines conditions (voir section 3).

Après une prochaine itération : désactivation de candidats composés Après une prochaine itération :

— le motif répété « champs à fluorescence induite » (fenêtres n°6 et n°11) pourrait devenir un nouveau candidat sans conséquences : « champs à fluorescence » serait consommé au profit d'un `candidat` plus complexe.
— le motif répété « fluorescence de chlorophylle contaminée » (fenêtres n°3 et n°5) pourrait devenir un nouveau candidat. Alors « fluorescence de chlorophylle » occurrerait 1 seule fois (fenêtre n°9), et entrerait en récession. Nous étudions ce cas ci-dessous.

La phase de récession de « fluorescence de chlorophylle » donnerait :

1. fluorescence de chlorophylle → fluorescence et chlorophylle
2. Une occurrence du `candidat` fluorescence de molécule désactivée est enregistrée (fenêtre n°9)
3. Concernant chlorophylle. Hypothèse 1 : chlorophylle est devenu `candidat` (entre temps) et occurre par ailleurs dans le texte :
   (a) imaginons que le `candidat` chlorophylle occurre déjà (au moins) 3 fois par ailleurs dans le texte
   (b) une nouvelle (4e) occurrence du `candidat` chlorophylle est enregistrée. Cette nouvelle occurrence du `candidat` chlorophylle est appelé `branche` de l'occurrence du `candidat` fluorescence de chlorophylle désactivée.
   (c) chlorophylle est toujours `candidat`
4. Concernant chlorophylle. Hypothèse 2 : « chlorophylle » n'est jamais devenu `candidat` (peu importe s'il occurre par ailleurs)
   (a) cette occurrence de chlorophylle redevient un simple lemme. Ce lemme est également une `branche` de l'occurrence du `candidat` fluorescence de chlorophylle désactivée.
5. Concernant fluorescence
   (a) le `candidat` fluorescence occurre 1 fois désactivée.
   (b) en réactivant le `candidat` fluorescence et en comptant la nouvelle occurrence ; il y a 2 occurrences du `candidat` fluorescence

(c) cela ne suffit pas. Une seconde occurrence du `candidat` fluorescence désactivé est enregistrée. Cette nouvelle occurrence du `candidat` fluorescence désactivé est appelé `branche` de l'occurrence du `candidat` fluorescence de chlorophylle désactivée.
(d) l'occurrence de « fluorescence » est donc un simple lemme

Une future récession de `candidats` complexes vers le `candidat` fluorescence de chlorophylle pourrait réactiver ce `candidat`. Comme pour toute formation de `candidat`, les composants sont consommés.

Destruction de l'arbre des occurrences de candidat désactivé Comme spécifié en section 3.2.3, tout développement d'un `candidat` (activé ou désactivé) détruit l'arbre des occurrences désactivées qui dépendent de lui. Par exemple dans notre cas, en fenêtre n°9, les `branche` de l'arbre de l'occurrence du `candidat` fluorescence de chlorophylle désactivé sont « fluorescence » et « chlorophylle » (`candidat` ou lemme). Le développement (expression, expansion) d'une de ces branches détruira l'arbre en amont. Par exemple, une expansion qui utiliserait cette occurrence de chlorophylle (pour faire chlorophylle verte par exemple), détruirait l'arbre comme indiqué en figure C.3.

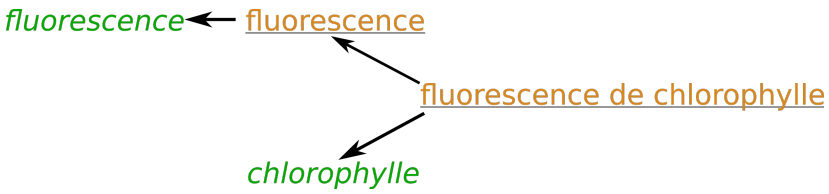


FIGURE C.2 – Arbre de récession de fluorescence de chlorophylle vers des formes stables. Les `candidats` sont soulignés, en orange les `candidats` désactivés, en vert les formes stables, en italique les lemmes simples.

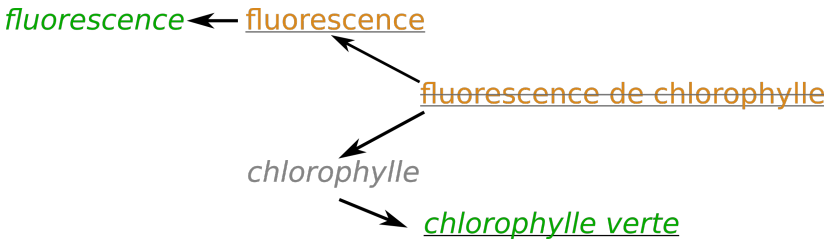


FIGURE C.3 – Destruction de l'arbre de récession d'une occurrence de « chlorophylle » par expansion.

Pour l'exemple nous n'avons pas pu produire de jeu de données suffisamment complexe, mais il est fréquent que 2 `candidats` composés régressent vers des composants qui se trouvent alors réactivés. Dans notre exemple, il faudrait qu'un autre `candidat` contenant fluorescence régresse. Or cela n'est pas cohérent dans notre exemple, en effet la phase d'essaimage (“A22 : Essaimage” (1)) nous certifie que nous avons toute les occurrences de « fluorescence ».

Publications et références personnelles

Revues d'audience internationale à comité de lecture
Quantin M., Hervy B., Laroche F. et Bernard A. *Haruspex : Multi-fuzzy-graph from Unstructured Texts Keyphrases Extraction* in Knowledge Based Systems. *Publication soumise en Septembre 2017*

Revues nationales à comité de lecture
Quantin M. *L'empreinte des ponts roulants dans les Halles Alstom* in Patrimoine Industriel, 66/67, 2015
Quantin M. *Récit historique et objet technique : outil de valorisation mutuelle* in Cahier d'Histoire du CNAM, 5, 2016
Teissier P.,Quantin M., Hervy B. *Extraction de connaissances en histoire des sciences. Cartographie d'une mémoire collective sur les matériaux* in Cahier du Centre François Viète, 2017

Congrès internationaux à comité de sélection et actes publiés
Quantin M., Hervy B., Laroche F. et Kerouanton J.-l. (2015). *Mass customization for cultural heritage 3D models* in Digital Heritage, 2015, Grenada
Quantin M., Hervy B., Laroche F. et Bernard A. *Supervised Process of Un-structured Data Analysis for Knowledge Chaining* in Wang L., éditeur : CIRP Design, 2016, Stockholm

Congrès nationaux et actes publiés
Hervy B., Quantin M., Laroche F., Bernard A. et Kerouanton J.-L. *Gestion de connaissances pour l'acquisition, le traitement et la valorisation des connaissances du patrimoine technique* in Ingénierie des Connaissances (IC), 2015, Rennes, France.
Quantin M. *Haruspex, outil de gestion de connaissances non structurées* in JIAP, 2016, Paris
Hervy B., Quantin M., Teissier P. *Extraction et chaînage supervisés de connaissances d'un corpus d'entretiens en histoire des sciences*, in EGC, 2017, Grenoble
Quantin M., Hervy B., Laroche F. *Extraction d'expressions et mise en réseau d'un corpus* in EDA, 2017, Lyon

Autres
Quantin M. *Personnalisation de masse pour le patrimoine industriel numérique* in GDR MACS, 2015, Nantes
Quantin M. *Représentation spatiales des connaissances* in Formation DG Patrimoine, 2015, Marseille
Quantin M. *Structuration de données numériques pour la conservation, l'analyse et la valorisation du patrimoine industriel* in Journée des doctorants CFV, 2015, Nantes
Quantin M. *Data production and retrieval* in _DayClick, 2016, Angers.
Quantin M. *Defining a semantic closeness indicator for a knowledge-based graph* in GDR MACS, 2016, Grenoble
Quantin M., Laroche F., André N., Villedieu F. *Rétroconception et maquettage d'un bâtiment mécanique de la Rome antique* in AIP PRIMECA, 2017, La Plagne
Quantin M., Laroche F., André N., Villedieu F. *Numérisation, modélisation et impression 3D pour la recherche historique et la muséographie. Cas d'étude sur la salle à manger de Néron* in SFHST, 2017, Strasbourg
Hervy B., Quantin M., Teissier P. *Haruspex : un outil numérique d'aide à l'analyse de corpus. Application à l'histoire de la chimie du solide* in SFHST, 2017, Strasbourg
Quantin M. *Modélisation du Mécanisme de la Salle à Manger Tournante de Néron* in Le Banquet, 2017, Marseille

Bibliographie

AAVV (2009). The london charterfor the computer-based visualisation of cultural heritage. 24

ACKOFF, R. L. (1989). From data to wisdom. *Journal of Applied Systems Analysis*, 16(1):3–9. 25

AKCA, D. (2017). 3d recording, documentation and management of cultural heritage. *Journal of Spatial Science*, 62(1):207–208. 35

ALATRISTA-SALAS, H., KERGOSIEN, E., ROCHE, M. et TEISSEIRE, M. (2014). Animitex project : Image analysis based on textual information. *In CEUR Workshop Proceedings*, volume 1318, pages 49–52. 40

ANETTE, H. (2004). *Combining machine learning and natural language processing for automatic keyword extraction*. Thèse de doctorat, Stockholm. 44

ARTHUR, D. et VASSILVITSKII, S. (2007). k-means++ : the advantages of careful seeding. *In Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1025. 50

ARUN, R., SURESH, V., MADHAVAN, C. E. V. et MURTY, M. N. (2010). On finding the natural number of topics with latent dirichlet allocation : Some observations. *In Lecture Notes in Computer Science*, volume 6118 LNAI, pages 391–402. Springer, Berlin, Heidelberg. 75

AUBRY, M., RUSSELL, B. et SIVIC, J. (2015). Visual geo-localization of non-photographic depictions via 2d-3d alignment. *In Visual Analysis and Geolocalization of Large-Scale Imagery*. Springer. 38

AXARIDOU, A., CHRYSAKIS, I., GEORGIS, C., THEODORIDOU, M., DOERR, M., KONSTANTARAS, A. et MARAVELAKIS, E. (2014). 3d-systek : Recording and exploiting the production workflow of 3d-models in cultural heritage. *IISA 2014 - 5th International Conference on Information, Intelligence, Systems and Applications*, pages 51–56. 38

AZPEITIA, A., CUADROS, M., GAINES, S. et RIGAU, G. (2014). Nerc-fr : Supervised named entity recognition for french. *In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 8655 LNAI, pages 158–165, Brno, Czech Republic. 40

BENSAUDE-VINCENT, B. et TEISSIER, P. (2015). Building, preserving and using oral archives on materials research. an attempt towards the biography of research communities. *In 10th International Conference on the History of Chemistry (ICHC)*, Aveiro, Portugal. 99, 100

BERETTA, F. et VERNUS, P. (2012). Le projet symogih et la modélisation de l'information : une opération scientifique au service de l'histoire. *Les Carnets du LARHRA*, (1):81–107. 35

BERKHIN, P. (2006). A survey of clustering data mining techniques. *In* KOGAN, J., NICHOLAS, C. et TEBOULLE, M., éditeurs : *Grouping Multidimensional Data : Recent Advances in Clustering*, pages 25–71. Springer, Berlin/Heidelberg. 49, 51

BERNARD, A. et BARLIER, C. (2015). *Fabrication additive. Du prototypage rapide à l'impression 3D*. Technique et Ingénierie. Dunod. 62

BERNARD, M. et BOHET, B. (2017). *Littérométrie. Outils numériques pour l'analyse des textes littéraires*. 61

BERNERS-LEE, T. (2007). Giant global graph. *Decentralized Information Group*, page 29. 33

BERRY, D. M. (2011). The computatonal turn : Thinking about the digital humanities. *Culture Machine*, 12:1–22. 21, 22

BERTIN, J. (1983). *Semiology of Graphics*, volume 94. University of Wisconsin Press. 139

BIETTE, A. (2013). *L'Usine Bleue, Sucre des Îles, Sucre des Champs*. e+pi, regards d' édition. 123

BLEI, D. M., NG, A. Y. et JORDAN, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022. 51

BLUE, M., BUSH, B. et PUCKETT, J. (2002). Unified approach to fuzzy graph problems. *Fuzzy Sets and Systems*, 125(3):355–368. 32, 71

BOBILLO, F. et STRACCIA, U. (2011). Fuzzy ontology representation using owl 2. *International Journal of Approximate Reasoning*, 52(7):1073–1094. 31

BOBILLO, F. et STRACCIA, U. (2016). The fuzzy ontology reasoner fuzzydl. *Knowledge-Based Systems*, 95:12–34. 31

BOHNET, B. (2010). Very high accuracy and fast dependency parsing is not a contradiction. *Proceedings of the 23rd International Conference on Computational Linguistics (COLING '10)*, (August):89–97. 40

BONACICH, P. (2007). Some unique properties of eigenvector centrality. *Social Networks*, 29(4):555–564. 49

BOONYASOPON, P., RIEL, A., UYS, W., LOUW, L., TICHKIEWITCH, S. et DU PREEZ, N. (2011). Automatic knowledge extraction from manufacturing research publications. *CIRP Annals - Manufacturing Technology*, 60(1):477–480. 69

BORNET, C. et KAPLAN, F. (2017). A simple set of rules for characters and place recognition in french novels. *Frontiers in Digital Humanities*, 4:6. 40

BOURDON, F. et BOULET, V. (2015). Viaf : A hub for a multilingual access to varied collections. *World Library and Information Congress (IFLA)*. 29

BROOK WU, Y.-f., LI, Q., STEFAN BOT, R. et CHEN, X. (2005). Domain-specific keyphrase extraction. *In Conference on Information and Knowledge Management (CIKM)*, Bremen, Germany. 44

BRUNET, J. P., TAMAYO, P., GOLUB, T. R. et MESIROV, J. P. (2004). Metagenes and molecular pattern discovery using matrix factorization. *National Academy of Sciences of the USA*, 101(12):4164–4169. 74

BU, F., ZHU, X. Y. et LI, M. (2010). A new multiword expression metric and its applications. *Journal of Computer Science and Technology*, 26(1):3–13. 44, 45, 46, 82

BURNARD, L. (2012). Du literary and linguistic computing aux digital humanities : retour sur 40 ans de relations entre sciences humaines et informatique. *In* MOUNIER, P., éditeur : *EHESS seminaire Digital Humanities*. OpenEdition Press, Marseille, read/write édition. 22, 62

CARLSON, A., BETTERIDGE, J. et KISIEL, B. (2010). Toward an architecture for never-ending language learning. *In Artificial Intelligence (AAAI)*, pages 1306–1313. 41

CAROTHERS, G. (2014). Rdf 1.1 n-quads - a line-based syntax for an rdf datasets. 31

CESERANI, G. et ARMOND, T. D. (2015). British architects on the grand tour in eighteenth-century italy : Travels, people, places. Rapport technique. 58

CHAUDHURI, S. et DAYAL, U. (1997). An overview of data warehousing and olap technology. *ACM SIGMOD Record*, 26(1):65–74. 26, 35

CHAUNU, P. (1970). L'histoire sérielle. bilan et perspectives. *Revue Historique*, 243(2):297–320. 57

CHOAY, F. (2009). *Le patrimoine en questions, anthologie pour un combat*. Seuil. 23

CHURCH, K. W. et HANK, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1). 46

CILIBRASI, R. L. et VITÁNYI, P. M. B. (2007). The google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, 19(3):370–383. 45

CLAVERT, F. (2013). *L'histoire contemporaine à l'ère numérique*. 22

COLLAO, A. J., DIAZ-KOMMONEN, L., KAIPAINEN, M. et PIETARILA, J. (2003). Soft ontologies and similarity cluster tools to facilitate exploration and discovery of cultural heritage resources. *Proceedings - International Workshop on Database and Expert Systems Applications, DEXA*, pages 75–79. 69

COTTE, M. (2009). Les techniques numériques et l'histoire des techniques. le cas des maquettes virtuelles animées. *Documents pour l'histoire des techniques. Nouvelle série*, 18(18):7–21. 36, 38

CRAM, D. et DAILLE, B. (2016). Termsuite : Terminology extraction with term variant detection. *In Annual Meeting of the Association for Computational Linguistics*, pages 13–18, Berlin. Association for Computational Linguistics. 44, 45, 83, 92

CRESSIE, N. et WIKLE, C. K. (2011). *Statistics for spatio-temporal data*. Wiley. 57

CROFTS, N., DIONISSIADOU, I., DOERR, M. et STIFF, M. (1999). Définition du modèle conceptuel de référence du cidoc (crm). Rapport technique, Crozat. 32

DAGAN, I. et CHURCH, K. (1994). Termight : Identifying and translating technical terminology. *In Proceedings of the 4th Conference on Applied Natural Language Processing (ANLP'94)*, pages 34–40. 44

DAILLE, B., GAUSSIER, É. et LANGÉ, J.-M. (1994). Towards automatic extraction of monolingual and bilingual terminology. *In Proceedings of the 15th conference on Computational linguistics* -, volume 1, page 515, Morristown, NJ, USA. Association for Computational Linguistics. 44

DAMOVA, M. et DANNELLS, D. (2011). Reason-able view of linked data for cultural heritage. *In International Conference on Software, Services and Semantic Technologies S3T 2011*, pages 17–24. Springer Berlin Heidelberg. 33

DAVALLON, J. (2000). Le patrimoine : "une filiation inversée" ? *Espaces Temps*, 74(1):6–16. 126

DAVALLON, J. (2002). Comment se fabrique le patrimoine ? *Sciences Humaines. Hors série n°36*, mars/avril: http ://www.scienceshumaines.com/comment–fabriqu. 23

de CHADAREVIAN, S. (1997). Using interviews to write the history of science. *In* SÖDERQVIST, T., éditeur : *The Historiography of Contemporary Science and Technology*, pages 51–70. Amsterdam, harwood ac édition. 100

de MASSARY, X. et COSTE, G. (2007). Principes, méthode et conduite de l'inventaire général du patrimoine culturel. *Documents & Methodes, Ministère de la Culture et de la Communication*, (9). 23

DE ROO, B., VAN DE WEGHE, N., BOURGEOIS, J. et DE MAEYER, P. (2013). the temporal dimension in a 4d archaeological data model : Applicability of the geoinformation standard. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, II-2/W1(W1):111–121. 36

DEISY, C., GOWRI, M., BASKAR, S., KALAIARASI, S. M. A. et RAMRAJ, N. (2010). A novel term weighting scheme midf for text categorization. *Journal of Engineering Science and Technology*, 5(1):94–107. 45

DEMETRESCU, E. et FANINI, B. (2017). A white-box framework to oversee archaeological virtual reconstructions in space and time : Methods and tools. *Journal of Archaeological Science : Reports*, 14:500–514. 37

DESJARDIN, É., NOCENT, O. et de RUNZ, C. (2012). Prise en compte de l'imperfection des connaissances depuis la saisie des données jusqu'à la restitution 3d. *Archeologia e Calcolatori*, page 389. 38

DIJKSTRA, E. W. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik*, 1(1):269–271. 91

DOERR, M. (2003). The cidoc conceptual reference module : An ontological approach to semantic interoperability of metadata. *AI Magazine*, 24(3):75. 32

DRUCKER, J. (2011). Humanities approaches to graphical display. *Digital Humanities Quarterly*, 5(1):1–23. 23, 25

DUTERTRE, E. (2005). *Savons et savonneries, le modèle nantais*. e+pi, memo édition. 123

DYENS, O. (2008). *La condition inhumaine. Essai sur l'effroi technologique.* Flammarion, essais édition. 22

ENGUEHARD, C. (1993). Acquisition de terminologie à partir de gros corpus. *Informatique & Langue Naturelle*, pages p.373–384. 80

ENGUEHARD, C. (2001). Flexible-equality of terms : Definition and evaluation. *In* LARSEN, H. L., ANDREASEN, T., CHRISTIANSEN, H., KACPRZYK, J. et ZADROŻNY, S., éditeurs : *Flexible Query Answering Systems*, pages 289–300. Physica-Verlag HD, Heidelberg. 80

ENGUEHARD, C. (2005). Terminology. *In* ALTMAN, G., KÖHLER, R., PIOTROWSKI., R. G. et de GRUYTER, W., éditeurs : *Quantitative Linguistics*, pages pp.971–988. De Gruyter Mouton, Berlin/New York. 71

ENGUEHARD, C. et PANTERA, L. (1995). Automatic natural acquisition of a terminology. *Journal of quantitative linguistics*, 2(1):27–32. 44, 71, 77, 78

FANO, R. M. et HAWKINS, D. (1961). Transmission of information : A statistical theory of communications. *American Journal of Physics*, 29(11):793–794. 46

FATIMA, I., KHATTAK, A. M., LEE, Y.-K. et LEE, S. (2011). Automatic documents annotation by keyphrase extraction in digital libraries using taxonomy. *In* PARK J.J., YANG L.T., L. C., éditeur : *Future Information Technology (FutureTech)*, pages 47–56. Springer, Berlin, Heidelberg. 44

FÉNOGLIO, I. et GANASCIA, J.-G. (2008). Le logiciel medite : approche comparative de documents de genèse. *In L'édition du manuscrit - De l'archive de création au scriptorium électronique*, chapitre 10, pages 209–228. Academia-Bruylant. 60

FINLAYSON, M. A. et KULKARNI, N. (2011). Detecting multi-word expressions improves word sense disambiguation. *In Multiword Expressions : from Parsing and Generation to the Real World*, pages 20–24, Stroudsburg, PA, USA. Association for Computational Linguistics. 43

FRANK, E., PAYNTER, G., WITTEN, I., GUTWIN, C. et NEVILL-MANNING, C. (1999). Domain-specific keyphrase extraction. *In Proceedings of the 14th ACM international conference on Information and knowledge management - CIKM '05*, pages 668–673, Stockholm, Sweden. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. 44

FRANTZI, K., ANANIADOU, S. et MIMA, H. (2000). Automatic recognition of multi-word terms : the c-value/nc-value method. *International Journal on Digital Libraries*, 3(2):115–130. 44, 92

GARCIA-FERNANDEZ, J., ANSSI, J., AHN, Y. et FERNANDEZ, J. J. (2015). Quantitative + qualitative information for heritage conservation : An open science research for paving 'collaboratively' the way to historical-bim. *In Digital Heritage 2015*, pages 207–208, Grenada. IEEE. 37

GENET, J.-P. et ZORZI, A. (2011). *Les historiens et l'informatique un métier à réinventer*. École fran édition. 22

GOERZ, G. et SCHOLZ, M. (2010). Adaptation of nlp techniques to cultural heritage research and documentation. *Journal of Computing and Information Technology*, 18(4):317. 40

GOMES, L., REGINA PEREIRA BELLON, O. et SILVA, L. (2014). 3d reconstruction methods for digital preservation of cultural heritage : A survey. *Pattern Recognition Letters*, 50:3–14. 23, 35

GÓMEZ, L., KUIJPERS, B. et VAISMAN, A. (2017). Performing olap over graph data. *In International Workshop on Real-Time Business Intelligence and Analytics - BIRTE '17*, pages 1–8, New York, New York, USA. ACM Press. 27

GREENE, D., O'CALLAGHAN, D. et CUNNINGHAM, P. (2014). How many topics ? stability analysis for topic models. *In* T., C., F., E., E., H. et R., M., éditeurs : *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, pages 498–513, France. 74

GRIRA, N., CRUCIANU, M. et BOUJEMAA, N. (2004). Unsupervised and semi-supervised clustering : A brief survey. *A Review of Machine Learning Techniques for Processing Multimedia Content, Report of the MUSCLE European Network of Excellence (6th Framework Programme)*, pages 1–12. 40

GUPTA, S. et MANNING, C. D. (2014). Improved pattern learning for bootstrapped entity extraction. *Conference on Computational Natural Language Learning, CoNLL*, pages 98–108. 40

HAMERLY, G. et ELKAN, C. (2002). Alternatives to the k-means algorithm that find better clusterings. *Conference on Information and Knowledge Management (CIKM)*, 4(09):600–607. 50

HAN, H., GILES, C. L., MANAVOGLU, E., ZHA, H., ZHANG, Z. et FOX, E. a. (2003). Automatic document metadata extraction using support vector machines. *Proceedings of ACM/IEEE-CS joint conference on Digital libraries*, pages 37–48. 41

HARTOG, F. (1993). *Régimes d'historicité. Présentisme et expériences du temps*. 1993 édition. 22

HARTUV, E. et SHAMIR, R. (2000). A clustering algorithm based on graph connectivity. *Information Processing Letters*, 76(4-6):175–181. 55

HEIDEN, S. (2010). The txm platform : Building open-source textual analysis software compatible with the tei encoding scheme. *In 24th Pacific Asia Conference on Language, Information and Computation*, pages 389–398. Institute for Digital Enhancement of Cognitive Development, Waseda University. 59

HEIMBURGER, F. et RUIZ, É. (2012). Faire de l'histoire à l'ère numérique : retours d'expériences. *Revue d'histoire moderne et contemporaine*, 58-4(5):70–89. 33

HEINICH, N. (2009). *La fabrique du patrimoine*. Maison des Sciences de l'Homme (Éditions de la). 23, 41

HERVY, B. (2014). *Modélisation et intégration de données hétérogènes sur un cycle de vie produitcomplexe*. Thèse de doctorat, École Centrale de Nantes. 7, 9, 14, 24, 26, 37, 46, 122

HERVY, B., LAROCHE, F. et BERNARD, A. (2012). An information system for driving the future plm for museum : The dhrm, digital heritage reference model. *In Advanced Composite Materials and Processing ; Robotics ; Information Management and PLM ; Design Engineering*, pages 465–471. ASME. 35

HERVY, B., LAROCHE, F., KEROUANTON, J.-L., BERNARD, A., COURTIN, C., D'HAENE, L., GUILLET, B. et WAELS, A. (2014). Augmented historical scale model for museums : from curation to multi-modal promotion. *In Laval Virtual VRIC 14*, pages 10–13, Laval (France). 35, 126

HOCKEY, S. (2007). The history of humanities computing. *In* SCHREIBMAN, S., SIEMENS, R. et UNSWORTH, J., éditeurs : *A Companion to Digital Humanities*, pages 1–19. Blackwell Publishing Ltd, Malden, MA, USA, oxford : bl édition. 21

HODGE, G. (2000). *Systems of Knowledge Organization for Digital Libraries : Beyond Traditional Authority Files*. Numéro 91. 28

HOLZAPFEL, K., KOSUB, S., MAASS, M. G. et TAUBIG, H. (2006). The complexity of detecting fixed-density clusters. *Discrete Applied Mathematics*, 154(11):1547–1562. 56

HOYER, P. O. (2004). Non-negative matrix factorization with sparseness constraints. *The Journal of Machine Learning Research*, 5:1457–1469. 75

HUANG, A. (2008). Similarity measures for text document clustering. *In NZCSRSC*, numéro April, pages 49–56, Christchurch, New Zealand. 47

ICOMOS (1996). Principles for the recording of monuments, groups of buildings and sites. Rapport technique, Sofia (Bulgaria). 24

IHDE, D. (2009). *Postphenomenology and Technoscience : The Peking University Lectures*. SUNY Press. 138

ILLIEN, G., HOLOGNE, O., POUYLLAU, S., ALFONSI, G., TROEIRA, J.-P., DELAHOUSSE, J., DALBIN, S., von REKOWSKI, U., AUBRY, C. et HUOT, C. (2013). Enjeux professionnels. *Documentaliste-Sciences de l'Information*, 50(3):26–41. 30

ISO, I. O. f. S. (1988). Iso8601, data elements and interchange formats– representation of dates and times. Rapport technique. 58

ISO, I. O. f. S. (2015). Geographic information - temporal schema. Rapport technique. 36, 57

JACOMY, M., GIRARD, P., OOGHE, B. et VENTURINI, T. (2016). Hyphe, a curation-oriented approach to web crawling for the social sciences. 59

JACQUEMIN, C. (2001). *Spotting and discovering terms through natural language processing*. MIT Press, Cambridge Mass. 44

JAKAWAT, W. (2016). *Graphs enriched by Cubes (GreC) : a new approach for OLAP on information networks*. Thèse de doctorat, Université Lumière Lyon 2. 27

JIANG, X., HU, Y. et LI, H. (2009). A ranking approach to keyphrase extraction. *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '09*, (5):756. 44

JONES, S. et PAYNTER, G. W. (2002). Automatic extraction of document keyphrases for use in digital libraries : Evaluation and applications. *Journal of the American Society for Information Science and Technology*, 53(8):653–677. 44

KAPRALOV, M., POTLURU, V. K. et WOODRUFF, D. P. (2016). How to fake multiply by a gaussian matrix. *In International Conference on Machine Learning (ICML)*, New York City, NY, USA. 52

KARGER, D. R. et STEIN, C. (1996). A new approach to the minimum cut problem. *Journal of the ACM*, 43(4):601–640. 55

KASSNER, L., GRÖGER, C., MITSCHANG, B. et WESTKÄMPER, E. (2014). Product life cycle analytics – next generation data analytics on structured and unstructured data. *CIRP Conference on Intelligent Computation in Manufacturing Engineering*, 33:35–40. 69

KIM, S. N., BALDWIN, T. et KAN, M.-y. (2009). The use of topic representative words in text categorization. *In Proceedings of the fourteenth Australasian document computing symposium (ADCS 2009)*, pages 75–81, Sydney, Australia. 44

KITCHIN, R. (2014). Big data, new epistemologies and paradigm shifts. *Big Data & Society*, (June):1–12. 16, 21

KOLLER, D., FRISCHER, B. et HUMPHREYS, G. (2009). Research challenges for digital archives of 3d cultural heritage models. *Journal on Computing and Cultural Heritage*, 2(3):1–17. 35

KOLMOGOROV, A. N. (1963). On tables of random numbers. *Source : Sankhyā : The Indian Journal of Statistics, Series A*, 25(4):369–376. 46

KRAUTHAMMER, M. et NENADIC, G. (2004). Term identification in the biomedical literature. 44

KUHN, H. (1955). The hungarian method of solving the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97. 74

LADURIE, E. L. R. (1973). *Le Territoire de l'historien, t.1*. Collection édition. 22

LAKATOS, I., WORRALL, J. et CURRIE, G. (1980). *The methodology of scientific research programmes*. Cambridge University Press, Cambridge, New York. 22

LAMIREL, J.-c., DUGUÉ, N. et CUXAC, P. (2016). Performing and visualizing temporal analysis of large text data issued for open sources : Past and future methods. *In BDAS : Beyond Databases, Architectures and Structures. Ad- vanced Technologies for Data Mining and Knowledge Discovery*. 57

LANDAUER, T. K., FOLT, P. W. et LAHAM, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, 25(2):259–284. 51

LAROCHE, F. (2007). *Contribution à la sauvegarde des objets techniques anciens par l'archéologie industrielle avancée. Proposition d'un Modèle d'information de référence muséologique et d'une Méthode inter-disciplinaire pour la Capitalisation des connaissances du Patrimoine*. Thèse de doctorat, École Centrale Nantes. 13, 37

LAROCHE, F., BERNARD, A. et COTTE, M. (2008). Knowledge management for industrial heritage. *Methods and Tools for Effective Knowledge Life-Cycle-Management*, pages 1–586. 35

LAUNAY, M. (1986). Effet de sens, produit de quoi ? *Langages*, 21(82, numéro spécial Le signifiant):13–39. 138

LE, Q. et MIKOLOV, T. (2014). Distributed representations of sentences and documents. *International Conference on Machine Learning - ICML 2014*, 32:1188–1196. 55

le DEUFF, O. (2014). Le temps des changements. *In Le temps des humanit{é}s digitales*, chapitre 3, pages 115–172. Limoges, fyp édition. 22, 126

LEE, D. et SEUNG, H. (2001). Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, (1):556–562. 51

LEVENSHTEIN, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710. 80

LI, L., TANG, L., ZHU, H., ZHANG, H., YANG, F. et QIN, W. (2017). Semantic 3d modeling based on citygml for ancient chinese-style architectural roofs of digital heritage. *ISPRS International Journal of Geo-Information*, 6(5):132. 37

LIU, A. (2016). Where is cultural criticism in the digital humanities ? *In* GOLD, M. K., éditeur : *Debates in the Digital Humanities*, chapitre 29, page 632. University of Minnesota Press, Minneapolis. 23

LIU, H. et MOTODA, H. (2007). *Computational Methods of Feature Selection*. Chapman & Hall/CRC. 45

LOPES, N., ZIMMERMANN, A. et HOGAN, A. (2010). Rdf needs annotations. *W3C Workshop on RDF . . .* , pages 1–5. 31

MAIETTI, F., GIULIO, R. D., BALZANI, M., PIAIA, E., MEDICI, M. et FERRARI, F. (2017). Digital memory and integrated data capturing : innovations for an inclusive cultural heritage in europe through 3d semantic modelling. *In Mixed Reality and Gamification for Cultural Heritage*, pages 1–19. Springer International Publishing, Cham. 38

MALRAUX, A. (1965). *Le musée imaginaire*. 126

MANUEL, A., VÉRON, P. et LUCA, L. D. (2016). 2d/3d semantic annotation of spatialized images for the documentation and analysis of cultural heritage. *Proceedings of the 14th Eurographics Workshop on Graphics and Cultural Heritage*, pages 101–104. 35

MARRERO, M., URBANO, J., S ? ?NCHEZ-CUADRADO, S., MORATO, J. et G ? ?MEZ-BERB ? ?S, J. M. (2013). Named entity recognition : Fallacies, challenges and opportunities. *Computer Standards and Interfaces*, 35(5):482–489. 40

MARTIN, P. (2015). *L'or brun de l'estuaire, l'industriel, le port et le paysan*. Coiffard. 123

MARTIN, P. et PHILIPPE (2015). La production de guano artificiel, une étape dans la professionnalisation des fabricants d'engrais : l'exemple d'édouard derrien à nantes (1840-1860). *Annales de Bretagne et des pays de l'Ouest*, (122-1):161–190. 123

MASTERMAN, M. (1962). The intellect's new eye. *In Freeing the Mind*. The Times Publishing Company, London, times lite édition. 97, 138

MCCARTY, W. (2016). A telescope for the mind ? *In* GOLD, M. K., éditeur : *Debates in the Digital Humanities*, chapitre 8, page 490. University of Minnesota Press, Minneapolis. 23

MEI, Q. et ZHAI, C. (2005). Discovering evolutionary theme patterns from text. *In ACM Knowledge discovery in data mining (KDD)*, page 198, New York, New York, USA. ACM Press. 57

MIKOLOV, T., CORRADO, G., CHEN, K. et DEAN, J. (2013a). Efficient estimation of word representations in vector space. *Proceedings of the International Conference on Learning Representations (ICLR 2013)*, pages 1–12. 44, 54, 85

MIKOLOV, T., SUTSKEVER, I., CHEN, K., CORRADO, G. et DEAN, J. (2013b). Distributed representations of words and phrases and their compositionality. *In Advances in Neural Information Processing Systems 26 (NIPS 2013)*, pages 3111–3119, California (USA). 44

MILNE, D. et WITTEN, I. H. (2008). Learning to link with wikipedia. *In Proceeding of the 17th ACM Conference on Information and Knowledge Management (CIKM '08)*, pages 509–518, Napa Valley, USA (CA). ACM. 81

MITCHELL, T., COHEN, W., HRUSCHKA, E., TALUKDAR, P., BETTERIDGE, J., CARLSON, A., DALVI, B., GARDNER, M., KISIEL, B., KRISHNAMURTHY, J., LAO, N., MAZAITIS, K., MOHAMED, T., NAKASHOLE, N., PLATANIOS, E., RITTER, A., SAMADI, M., SETTLES, B., WANG, R., WIJAYA, D., GUPTA, A., CHEN, X., SAPAROV, A., GREAVES, M. et WELLING, J. (2015). Never-ending learning. *In Artificial Intelligence (AAAI)*, pages 2302–2310. 41

MORETTI, F. (2005). *Graphs, maps, trees : abstract models for a literary history*. Verso, London ; New York. 21, 23

MOUNIER, P. (2011). Du discours aux données....et retour ? 69

MOUNIER, P. (2017). Enquête sur une guerre souterraine au sein de la recherche. 22, 23

NADEAU, D. (2007). A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26. 40

NICOLAS, T., GAUGNE, R., TAVERNIER, C., GOURANTON, V. et ARNALDI, B. (2016). La tomographie, l'impression 3d et la réalité virtuelle au service de l'archéologie. *Les Nouvelles de l'archéologie*, (146):16–22. 36

NIETZSCHE, F. (1885). Mille et un buts. *In Ainsi Parlait Zarathoustra*. 133

NYFFENEGGER, F., RIVEST, L. et BRAESCH, C. (2016). Identifying plm themes, trends and clusters through ten years of scientific publications. *In* HARIK, R., RIVEST, L., BERNARD, A., EYNARD, B. et BOURAS, A., éditeurs : *International Conference on Product Lifecycle Management (PLM)*, pages p. 579–591, Columbia, SC, USA. Springer. 60

of VIRTUAL ARCHAEOLOGY, I. F. (2013). Principles of seville. international principles of virtual archeology. Rapport technique, Seville - Spain. 24

OGURA, H., AMANO, H. et KONDO, M. (2009). Feature selection with a measure of deviations from poisson in text categorization. *Expert Systems with Applications*, 36(3 PART 2):6826–6832. 45

OLIVER, A. et VAZQUEZ, M. E. (2015). Tbxtools : A free, fast and flexible tool for automatic terminology extraction. *Proceedings of Recent Advances in Natural Language Processing*, pages 473–479. 44

OLSEN, M. (1993). Signs, symbols and discourses : A new direction for computer-aided literature studies. *Computers and the Humanities*, 27(5-6):309–314. 22

PAATERO, P. et TAPPER, U. (1994). Positive matrix factorization : A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126. 51

PARANYUSHKIN, D. (2011). Identifying the pathways for meaning circulation using text network analysis. *Nodus Labs*, (December):26. 60

PARK, Y., BYRD, R. J. et BEGURAEV, B. K. (2002). Automatic glossary extraction : beyond terminology identification. *Association for Computational Linguistics*, 1:1–7. 46

PAUKKERI, M.-s., NIEMINEN, I. T., MATTI, P., PÖLLÄ, M. et HONKELA, T. (2008). A language-independent approach to keyphrase extraction and evaluation. *In COLING*, pages 83–86. 44

PAVLIDIS, G., KOUTSOUDIS, A., ARNAOUTOGLOU, F., TSIOUKAS, V. et CHAMZAS, C. (2007). Methods for 3d digitization of cultural heritage. *Journal of Cultural Heritage*, 8(1):93–98. 35

PESTRE, D. (1997). La production des savoirs entre académies et marché. *Revue d'économie industrielle*, 79(1):163–174. 110

PRENDERGAST, C. (2005). Evolution and literary history : A response to franco moretti. *New Left Review*, II(34). 23

PROTASIEWICZ, J., PEDRYCZ, W., KOZŁOWSKI, M., DADAS, S., STANISŁAWEK, T., KOPACZ, A. et GAŁEZEWSKA, M. (2016). A recommender system of reviewers and experts in reviewing problems. *Knowledge-Based Systems*, 106:164–178. 69

QUANTIN, M. (2016). Récit historique et objet technique : outil de valorisation mutuelle. *Cahier d'Histoire du CNAM*, 5. 125

QUANTIN, M., HERVY, B., LAROCHE, F. et BERNARD, A. (2016). Supervised process of un-structured data analysis for knowledge chaining. *In* WANG, L., éditeur : *CIRP Design 2016*, Stockholm. Elsevier. 123

QUANTIN, M., LAROCHE, F., ANDRÉ, N. et VILLEDIEU, F. (2017). Rétroconception et maquettage d'un batiment mécanique de la rome antique. *In AIP Primeca*, La Plagne. 62

QUATTRINI, R., PIERDICCA, R., MORBIDONI, C., STOCKERT, J. C., BLÁZQUEZ-CASTRO, A., CA, M. et HOROBIN, R. W. (2017). Knowledge-based data enrichment for hbim : Exploring high-quality models using the semantic-web. *Journal of Cultural Heritage*, pages 1–12. 36

RAPTI, A., TSOLIS, D., SIOUTAS, S. et TSAKALIDIS, A. (2015). A survey : Mining linked cultural heritage data. *In Proceedings of the 16th International Conference on Engineering Applications of Neural Networks (INNS) - EANN '15*, pages 1–6, New York, New York, USA. ACM Press. 33

REINERT, M. (1999). Quelques interrogations à propos de l'"objet" d'une analyse de discours de type statistique et de la réponse "alceste". *Langage & Société*, 90:57–79. 60

REMONDINO, F. (2011). Heritage recording and 3d modeling with photogrammetry and 3d scanning. *Remote Sensing*, 3(6): 1104–1138. 35

REMONDINO, F. et CAMPANA, S. (2014). *3D Modeling in Archaeology and Cultural Heritage Theory and Best Practices*. 35

REN, H., VIAUD, M.-L. et MELANÇON, G. (2017). Evolution temporelle de communautés représentatives : mesures et visualisation. *Extraction et Gestion des Connaissances (EGC)*, pages 417–422. 57

RICHMAN, A. E. et SCHONE, P. (2008). Mining wiki resources for multilingual named entity recognition. *In Proceedings of ACL-08 : HLT*, numéro June, pages 1–9. 81

RIEDL, M. et BIEMANN, C. (2015). A single word is not enough : Ranking multiword expressions using distributional semantics. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*, (September): 2430–2440. 44

RIEGL, A. (1984). *Le culte modeme des monuments. Son essence, sa genèse*. Seuil, Paris. 23

RIEL, A., TICHKIEWITCH, S., MOLCHO, G., SHPITALNI, M., UYS, W., UYS, E. et du PREEZ, N. (2008a). Improving Product Development Organisations using Knowledge Mining : Requirements, Methods and Tools. *Knowledge Management in Product Development*. 38

RIEL, A., UYS, W. et TICHKIEWITCH, S. (2008b). Mining knowledge in the digital enterprise. *In International Conference on Digital Enterprise Technology*, numéro October, Nantes. 60, 65, 69

ROBINEAU, É. (2011). *Raffinage et raffineries de sucre à Nantes*. e+pi, Nantes, memo édition. 123

ROCHCONGAR, Y. (2003). *Capitaines d'industrie à Nantes au 19e siècle*. 123

ROSE, S., ENGEL, D., CRAMER, N. et COWLEY, W. (2010). Automatic keyword extraction from individual documents. *In Text Mining : Applications and Theory*, pages 1–20. John Wiley & Sons, Ltd, Chichester, UK. 43, 44

ROSENZWEIG, R. (2003). Scarcity or abundance ? preserving the past in a digital era. *The American Historical Review*, 108(3): 735–762. 23

ROSNAY, J. D. (1995). *L'homme symbiotique : regards sur le troisième millénaire*. Seuil. 22

ROWLEY, J. (2007). The wisdom hierarchy : representations of the dikw hierarchy. *Journal of Information Science*, 33(2):163–180. 25

RUECKER, S., RADZIKOWSKA, M., MICHURA, P., FIORENTINO, C. et CLEMENT, T. (2009). Visualizing repetition in text. *Digital Studies / Le champ numérique*, 1(3):1–9. 59

RUSKIN, J. (2008). *Les sept lampes de l'architecture*. Klincksieck. 23

SAG, I. A., BALDWIN, T., BOND, F., COPESTAKE, A. et FLICKINGER, D. (2002). Multiword expressions : A pain in the neck for nlp. *In Computational Linguistics and Intelligent Text Processing*, pages 1–15. Springer Berlin Heidelberg. 71

SALTON, G. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA. 45

SALTON, G., YANG, C. et YU, C. (1973). Contribution to the theory of indexing. Rapport technique 5, Ithaca, NY, USA. 86

SATOSHI SEKINE, K. S. et NOBATA, C. (2002). Extended named entity hierarchy. *In Third International Conference on Language Resources and Evaluation (LREC 2002)*, pages 1818–1824. 40

SCHAEFFER, S. E. (2007). Graph clustering. *Computer Science Review*, 1(1):27–64. 55

SCHMID, H. (1994). Probabilistic part-of-speech tagging using decision trees. *Proceedings of the International Conference on New Methods in Language Processing*, 4:44–49. 40, 72

SCHREIBMAN, S., SIEMENS, R. et UNSWORTH, J. (2004). *A Companion to Digital Humanities*. Blackwell Publishing Ltd, Malden, MA, USA. 21

SCLANO, F. et VELARDI, P. (2007). Termextractor : a web application to learn the common terminology of interest groups and research communities. *In Proceedings of the 9th Conference on Terminology and Artificial Intelligence (TIA 2007)*, pages 8–9, Sophia Antopolis (France). 44

SCOPIGNO, R., CALLIERI, M., CIGNONI, P., CORSINI, M., DELLEPIANE, M., PONCHIO, F. et RANZUGLIA, G. (2011). 3d models for cultural heritage : Beyond plain visualization. *IEEE Computer*, 44(7):48–55. 36

SHAHNAZ, F. et BERRY, M. W. (2006). Document clustering using nonnegative matrix factorization. *Information Processing & Management*, 42(2):373–386. 76

SHANNON, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423. 86

SHANNON, P., MARKIEL, A., OZIER, O., BALIGA, N. S., WANG, J. T., RAMAGE, D., AMIN, N., SCHWIKOWSKI, B. et IDEKER, T. (2003). Cytoscape : A software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–2504. 94

SHNEIDERMAN, B. (1996). The eyes have it : a task by data type taxonomy for informatio nvisualizations. *Proceedings 1996 IEEE Symposium on Visual Languages*, pages 336–343. 128

SIBSON, R. (1973). Slink : an optimally efficient algorithm for the single-link cluster method. 49

SINCLAIR, P., LEWIS, P., MARTINEZ, K., ADDIS, M. et PRIDEAUX, D. (2006). Semantic web integration of cultural heritage sources. *In Proceedings of the 15th international conference on World Wide Web - WWW '06*, page 1047, New York, New York, USA. ACM Press. 30

SINCLAIR, S. et ROCKWELL, G. (2014). Les potentialités du texte numérique. *In* EBERLE-SINATRA, M. et VITALI ROSATI, M., éditeurs : *Pratiques de l'édition numérique*, chapitre 12, page 219. Les Presses de l'Université de Montréal, Montréal (CA). 59

SLADE, J., JONES, C. B. et ROSIN, P. L. (2017). Automatic semantic and geometric enrichment of citygml building models using hog-based template matching. *In Advances in 3D Geoinformation*, pages 357–372. Springer, Cham. 37

SOROKIN, P. (1956). *Fads and Foibles in Modern Sociology and Related Sciences*. Regnery Publishing, Gateway Editions, Chicago. 139

SOUILI, A., CAVALLUCCI, D. et ROUSSELOT, F. (2015). Natural language processing (nlp) - a solution for knowledge extraction from patent unstructured data. *In Procedia Engineering*, volume 131, pages 635–643. 69

SRINIVASAN, R. et HUANG, J. (2005). Fluid ontologies for digital museums. *International Journal on Digital Libraries*, 5(3):193–204. 33

STANCO, F., BATTIATO, S. et GALLO, G. (2011). *Digital imaging for cultural heritage preservation : Analysis, restoration, and reconstruction of ancient artworks*. 35

STIEGLER, B. (2007). General pharmacology issues . there is no simple pharmakon questions de pharmacologie générale . il n ' y a pas de simple pharmakon. *Psychotropes*, 13(3):27–54. 22

STIEGLER, B. (2009). Technologies culturelles et économie de la contribution. *Culture et recherche*, (121):31. 21

STILO, G. et VELARDI, P. (2016). Efficient temporal mining of micro-blog texts and its application to event discovery. *Data Mining and Knowledge Discovery*, 30(2):372–402. 10, 57

SULA, C. A. (2013). Digital humanities and libraries : A conceptual model. *Journal of Library Administration*, 53(1):10–26. 9, 15

TEH, Y. W., JORDAN, M. I., BEAL, M. et BLEI, D. (2006). Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581. 52

TEISSIER, P. (2007). *L'émergence de la chimie du solide en France (1950-2000) de la formation d'une communauté à sa disparition*. Thèse de doctorat, Paris 10. 99

TEISSIER, P. (2014). *Une histoire de la chimie du solide. Synthèses, formes, identités*. Hermann, Paris. 106, 108

TEISSIER, P. et LOEVE, S. (2013). Archives orales : construire, conserver et contextualiser la mémoire savante contemporaine. *In Séminaire interdisciplinaire du Centre d'Alembert*. Université d'Orsay. 100

TEPPER, M. et SAPIRO, G. (2016). Compressed nonnegative matrix factorization is fast and accurate. *IEEE Transaction on Signal Processing (TSP)*, 64(9):2269–2283. 52

TJOA, S. K. et LIU, K. J. R. (2010). Multiplicative update rules for nonnegative matrix factorization with co-occurrence constraints. *In IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings (ICASSP)*, pages 449–452. IEEE. 51

TOKUNAGA, T. et MAKOTO, I. (1994). Text categorization based on weighted inverse document frequency. Rapport technique, Okayama. 45

TURNEY, P. D. (2000). Learning algorithms for keyphrase extraction. *Information Retrieval*, 2(4):303–336. 43

TURNEY, P. D. et PANTEL, P. (2010). From frequency to meaning : Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188. 71

UNESCO (2003). Charte sur la conservation du patrimoine numérique. 24

UNESCO (2016). Orientations devant guider la mise en œuvre de la convention du patrimoine mondial (fr). Rapport technique, Paris. 24

VALLET, J.-M., LUCA, L. D. et FEILLOU, M. (2012). Une nouvelle approche spatio-temporelle et analytique pour la conservation des peintures murales sur le long terme. *In Situ*, (19). 38

van ECK, N. J., WALTMAN, L., NOYONS, E. C. M. et BUTER, R. K. (2010). Automatic term identification for bibliometric mapping. *Scientometrics*, 82(3):581–596. 44

VEYNE, P. (1971). *Comment on écrit l'histoire*. Paris, seuil édition. 138

VICO, G. (1725). *La Science nouvelle*. Gallimard édition. 22

WAGNER, R. a. et FISCHER, M. J. (1974). The string-to-string correction problem. *Journal of the ACM*, 21(1):168–173. 80

WANG, X. et MCCALLUM, A. (2005). A note on topical n-grams. Rapport technique, Amherst. 44, 71

WELGER-BARBOZA, C. (2001). *Du musée virtuel au musée médiathèque : le patrimoine à l'ère du document numérique*. l'Harmattan. 23

WHITTLE, P. (1951). *Hypothesis Testing in Time Series Analysis*. Almqvist & Wiksells boktr., Uppsala. 56

WILKENS, M. (2015). Mapping and modeling centuries of literary geography across millions of books. *In Global Digital Humanities (DH)*, Sydney, Australia. 58

YAGHOOBZADEH, Y. et SCHÜTZE, H. (2015). Corpus-level fine-grained entity typing using contextual information. *In Empirical Methods in Natural Language Processing*, numéro September, pages 715–725, Lisbon, Portugal. 40

YAKEL, E., CONWAY, P., HEDSTROM, M. et WALLACE, D. (2011). Digital curation for digital natives. *Journal of Education for Library & Information Science*, 52(1):23–31. 34

ZACCHIROLI, S. (2017). Software heritage : Scholarly and educational synergies with preserving our software commons. *In Conference on Innovation and Technology in Computer Science Education - ITiCSE '17*, pages 3–3, New York, New York, USA. ACM Press. 33

ZALAMEA, O., ELINBAUM, P., EF, O., QMBOFT, D., FO, D., NJTNP, V. O., MB, U., EF, T., PNBSDBT, M. B. T., CARRIZO, S., YULN, M. et PANTAZIS, G. (2013). From a citygml to an ontology-based approach to support preventive conservation of built cultural heritage. *In Association of Geographic Information Laboratories in Europe (AGILE)*, volume 7616, pages 73–90. 37

ZHAO, P., LI, X., XIN, D. et HAN, J. (2011). Graph cube : On warehousing and olap multidimensional networks. *In ACM SIGMOD International Conference on Management of data*, pages 853–864, New York, New York, USA. ACM Press. 27

ZHOU, Z.-H. et ZHANG, M.-L. (2007). Multi-instance multilabel learning with application to scene classification. *Neural Information Processing Systems (NIPS)*, pages 2038–2048. 40

ZINS, C. (2007). Conceptual approaches for defining data, information, and knowledge. *Journal of the American Society for Information Science and Technology*, 58(4):479–493. 25

ZOGHLAMI, A., RUNZ, C. D., AKDAG, H., ZAGHDOUD, M. et GHEZALA, H. B. (2011). Handling imperfect spatiotemporal information from the conceptual modeling to database structures. *In International Symposium on Spatial Data Quality*, pages pp165–170, Coimbra, Portugal. 58

ZOU, Y., KIVINIEMI, A. et JONES, S. W. (2017). Retrieving similar cases for construction project risk management using natural language processing techniques. *Automation in Construction*, 80:66–76. 69

Liste des URL

a. EKLI : https://elki-project.github.io/

b. ClustViz : http://biit.cs.ut.ee/clustvis

c. Hyphe : http://hyphe.medialab.sciences-po.fr/

d. Le Trameur : http://www.tal.univ-paris3.fr/trameur/

e. NooJ : http://www.nooj4nlp.net/pages/nooj.html

f. Voyant-Tools : http://voyant-tools.org/

g. Google n-gram : https://books.google.com/ngrams

h. TextDNA : http://graphics.cs.wisc.edu/Vis/SequenceSurveyor/TextDNA.html

i. TXM : http://textometrie.ens-lyon.fr/

j. MONK : http://www.monkproject.org/

k. BioTex : http://tubo.lirmm.fr/biotex/

l. FastKwic : http://www.cnrtl.fr/outils/fastkwic/

m. AntX : http://www.laurenceanthony.net/software/

n. TAPoRware : http://taporware.ualberta.ca/

o. Unitex : http://www-igm.univ-mlv.fr/~unitex/

p. TermSuite : https://termsuite.github.io/

q. Sketch Engine : https://www.sketchengine.co.uk/

r. Lexico5 : http://www.lexi-co.com/

s. Medite : http://obvil.paris-sorbonne.fr/developpements/medite

t. ConcQuest : http://olivier.kraif.u-grenoble3.fr/ConcQuest/concquest.php

u. Tanagra : https://eric.univ-lyon2.fr/~ricco/tanagra/fr/tanagra.html

v. IRaMuTeQ : http://iramuteq.org/

w. Alceste : http://www.image-zafar.com/Logiciel.html

x. SplitsTree : http://ab.inf.uni-tuebingen.de/software/splitstree4/

y. Texture : textexture.com

z. WORDij : http://wordij.net/

aa. Linkurious : https://linkurio.us/

ab. Netlytic : https://netlytic.org

ac. SCITE : http://www.emiracle.eu

ad. TextObserver : http://textopol.u-pec.fr/textobserver/

ae. Prospero : prosperologie.org

af. Palladio : http://hdlab.stanford.edu/palladio/

ag. UAM CorpusTool : http://www.corpustool.com

ah. ANNIS : http://corpus-tools.org/annis

ai. CATMA : http://catma.de/

aj. GoogleAPI Natural Language : https://cloud.google.com/natural-language/

ak. Dandelion API : https://dandelion.eu/

al. Ambiverse : https://www.ambiverse.com

am. TextRazor : https://www.textrazor.com

an. Rosette : https://www.rosette.com/

ao. Aylien : https://aylien.com/

ap. Poemage : http://www.sci.utah.edu/~nmccurdy/Poemage/

aq. CORLI : http://explorationdecorpus.corpusecrits.huma-num.fr/

ar. TaPoR : http://tapor.ca/home

Titre : Proposition de chaînage des connaissances historiques et patrimoniales : approche multi-échelles et multi-critères de corpus textuels

Mots clés : multi-graphe flou, extraction de terminologie, patrimoine, connaissances

Résumé : Les humanités défient les capacités du numérique depuis 60 ans. Les années 90 marquent une rupture, énonçant l'espoir d'une interprétation qualitative automatique de données interopérables devenues « connaissances ». Depuis 2010, une vague de désillusion ternit ces perspectives, le foisonnement des humanités numériques s'intensifie.
Au cœur de ce paysage complexe, nous proposons une méthode implémentée produisant différentes vues d'un corpus textuel pour (1) l'analyse en interaction avec les connaissances qualitatives de l'historien et (2) la documentation numérique en contexte patrimonial (musée, site) auprès d'un public avisé. Les vues du corpus sont des graphes multiples pondérés, les documents sont des sommets liés par des arêtes renseignant les proximités sémantiques, temporelles et spatiales.
Cette méthode vise à co-créer des connaissances historiques. À l'utopie d'une modélisation des connaissances qualitatives de l'historien, nous préférons l'heuristique pragmatique: l'interprétation de quantifications du corpus suscite l'émergence de nouvelles certitudes et hypothèses.
Par ailleurs, notre approche (type OLAP) ouvre des parcours et accès personnalisés à chaque usager pour la documentation/analyse du patrimoine numérique voire 3D.
Plusieurs cas d'étude valident les méthodes proposées et ouvrent des perspectives d'applications industrielles.

Title : Mining technical textual data for historical heritage and knowledge development

Keywords : fuzzy multi graph, term mining, heritage, knowledge

Abstract : Humanities challenges computer sciences since 60 years. The 90's marks a break, announcing qualitative analysis and interpretation of interoperable data, which became «knowledge». Since 2010, a disillusionment tarnishes the prospects, Digital Hmanities diversity increases.
At the core of this complex background, we propose an implemented method producing various «views» of textual corpus in History. This views enable (1) interactive analysis with qualitative knowledge of the historian and (2) digital documentation of heritage on site (e.g. museum) for an advanced visitor.
Corpus views are weighted multi graphs. Documents are vertices linked by edges. Each edge contains semantic, temporal or spatial proximity information. This method aims at co-creating historical knowledge. Facing the utopian modeling of qualitative knowledge in history, we designed a pragmatic process : the historian analyses quantitative data of a known corpus, this generates new hypothesis and certainties.
Our approach (OLAP like) chart paths and customized access for each user to digital heritage documentation. These paths may meet 3D heritage data.
Several use cases validate the proposed method and open perspectives of industrial application.