Approche de recherche intelligente fondée sur le modèle des Topic Maps : application au domaine de la construction durable

Nebrasse Ellouze

CONSERVATOIRE NATIONAL DES ARTS ET METIERS

ECOLE NATIONALE DES SCIENCES DE L'INFORMATIQUE

THESE EN COTUTELLE
Préparée au sein des laboratoires RIADI-GDL (ENSI) et CEDRIC (Equipe ISID)

présentée par :

Nebrasse ELLOUZE

pour l'obtention du Diplôme de Doctorat en Informatique du CNAM (Paris) et de l'ENSI (Université de La Manouba)

Discipline/ Spécialité : Informatique

Approche de recherche intelligente fondée sur le modèle des Topic Maps

Application au domaine de la construction durable

Soutenue le 3 décembre 2010 au CNAM devant le jury d'examen :

Pr. **Jacky Akoka**, Professeur au CNAM, France	Examinateur
Pr. **Mohamed Ben Ahmed**, Professeur Emérite à l'université de la Manouba, Tunisie	Co-directeur de thèse
Pr. **Mokrane Bouzeghoub**, Professeur à l'UVSQ, France	Rapporteur
Dr. **Zoubida Kedad**, Maître de Conférences à l'UVSQ, France	Examinatrice
Dr. **Nadira Lammari**, Maître de Conférences au CNAM, France	Co-encadrante
Pr. **Elisabeth Métais**, Professeur au CNAM, France	Co-directrice de thèse
Pr. **Yacine Rezgui**, Professeur à l'université de Salford, UK	Rapporteur
Pr. **Max Silberztein**, Professeur à l'université de Franche-Conté, France	Examinateur

Dédicace

A mes très chers parents

A tous ceux que j'aime

Remerciements

C'est avec une grande émotion et beaucoup de sincérité que je voudrais exprimer ma gratitude à toutes les personnes ayant participé, soutenu et apprécié mon travail.

Tout d'abord, je tiens à remercier et exprimer toute ma reconnaissance auprès de mon directeur de thèse Pr. émérite **Mohamed Ben Ahmed** qui m'a initiée à la recherche et m'a toujours motivé, soutenu et encouragé. Méticuleux et perfectionniste, il m'a prodigué des conseils inestimables, dans tous les domaines, tout au long de ma thèse. Ses idées, son expérience et ses précieux conseils m'ont énormément aidée dans ce travail. Je le remercie pour sa disponibilité, son soutien et ses conseils nombreux et importants. Il a beaucoup contribué à la mise en valeur de mon travail, n'a cessé de m'encourager à avancer et m'a aidée à progresser à travers les difficultés et les doutes inhérents à tout travail de recherche.

Je souhaite remercier très vivement ma co-directrice de thèse Pr. **Elisabeth Métais** qui m'a accueilli pendant de longs séjours au laboratoire Cedric du Conservatoire National des Arts et Métiers, au cours desquels elle s'est montrée très disponible et accueillante pour discuter de mes travaux de thèse pendant de longues heures. Je la remercie également pour m'avoir guidé dans mes travaux, conseillé avec professionnalisme et une très grande expertise, sans jamais compter son temps ni perdre sa bonne humeur. Ses apports majeurs me permettent aujourd'hui de vous présenter cette thèse.

Je souhaite également remercier très chaleureusement ma co-encadrante, Dr. **Nadira Lammari** pour l'intérêt et la disponibilité qu'elle a manifestée à l'égard de mes recherches ainsi que pour tous les conseils et encouragements dont j'ai bénéficié tout au long de ce travail. Qu'elle trouve ici le témoignage de tout mon respect et ma reconnaissance et du plaisir que j'ai eu à travailler avec elle tout au long de ma thèse.

Qu'il me soit permis d'exprimer mes sincères remerciements à Pr. **Mokrane Bouzeghoub** et Pr. **Yacine Rezgui** pour accepter d'être mes rapporteurs de thèse. J'exprime également toute ma gratitude à Pr. **Jacky Akoka** et Dr. **Zoubida Kedad** qui m'ont fait l'honneur d'avoir accepté d'examiner ce travail.

Mes remerciements s'adressent aussi aux membres de l'équipe ISID pour les discussions scientifiques enrichissantes et l'ambiance amicale que nous avons partagée durant mes séjours au CNAM.

Enfin, mes sentiments les plus chaleureux sont pour ma famille. Je remercie mes parents qui m'ont toujours soutenue dans mes choix et qui m'ont toujours encouragée à aller de l'avant.

Résumé

Cette thèse aborde les problématiques liées à la construction de Topic Maps et à leur utilisation pour la recherche d'information dans le cadre défini par le Web sémantique (WS). Le WS a pour objectif de structurer les informations disponibles sur le Web. Pour cela, les ressources doivent être sémantiquement étiquetées par des métadonnées afin de permettre d'optimiser l'accès à ces ressources. Ces métadonnées sont actuellement spécifiées à l'aide des deux standards qui utilisent le langage XML : RDF et les Topic Maps.

Un contenu à organiser étant très souvent volumineux et sujet à enrichissement perpétuel, il est pratiquement impossible d'envisager une création et gestion d'une Topic Map, le décrivant, de façon manuelle. Plusieurs travaux de recherche ont concerné la construction de Topic Maps à partir de documents textuels [Ellouze et al. 2008a]. Cependant, aucune d'elles ne permet de traiter un contenu multilingue. De plus, bien que les Topic Maps soient, par définition, orientées utilisation (recherche d'information), peu d'entre elles prennent en compte les requêtes des utilisateurs.

Dans le cadre de cette thèse, nous avons donc conçu une approche que nous avons nommée ACTOM pour « **A**pproche de **C**onstruction d'une **TO**pic Map **M**ultilingue ». Cette dernière sert à organiser un contenu multilingue composé de documents textuels. Elle a pour avantage de faciliter la recherche d'information dans ce contenu. Notre approche est incrémentale et évolutive, elle est basée sur un processus automatisé, qui prend en compte des documents multilingues et l'évolution de la Topic Map selon le changement du contenu en entrée et l'usage de la Topic Map. Elle prend comme entrée un référentiel de documents que nous construisons suite à la segmentation thématique et à l'indexation sémantique de ces documents et un thésaurus du domaine pour l'ajout de liens ontologiques. Pour enrichir la Topic Map, nous nous basons sur deux ontologies générales et nous explorons toutes les questions potentielles relatives aux documents sources. Dans ACTOM, en plus des liens d'occurrences reliant un Topic à ses ressources, nous catégorisons les liens en deux catégories: (a) les liens ontologiques et (b) les liens d'usage. Nous proposons également d'étendre le modèle des Topic Maps défini par l'ISO en rajoutant aux caractéristiques d'un Topic des méta-propriétés servant à mesurer la pertinence des Topics plus précisément pour l'évaluation de la qualité et l'élagage dynamique de la Topic Map.

Mots clés : Topic Map, recherche d'information, enrichissement, documents multilingues, thésaurus, requêtes des utilisateurs, fusion, élagage, évolution.

Abstract

The research work in this thesis is related to Topic Map construction and their use in semantic annotation of web resources in order to help users find relevant information in these resources. The amount of information sources available today is very huge and continuously increasing, for that, it is impossible to create and maintain manually a Topic Map to represent and organize all these information. Many Topic Maps building approaches can be found in the literature [Ellouze et al. 2008a]. However, none of these approaches takes as input multilingual document content. In addition, although Topic Maps are basically dedicated to users navigation and information search, no one approach takes into consideration users requests in the Topic Map building process.

In this context, we have proposed ACTOM, a Topic Map building approach based on an automated process taking into account multilingual documents and Topic Map evolution according to content and usage changes. To enrich the Topic Map, we are based on a domain thesaurus and we propose also to explore all potential questions related to source documents in order to represent usage in the Topic Map. In our approach, we extend the Topic Map model that already exists by defining the usage links and a list of meta-properties associated to each Topic, these meta-properties are used in the Topic Map pruning process. In our approach ACTOM, we propose also to precise and enrich semantics of Topic Map links so, except occurrences links between Topics and resources, we classify Topic Map links in two different classes, those that we have called “ontological links” and those that we have named “usage links”.

Keywords: Topic Map, information search, enrichment, multilingual documents, thesaurus, user requests, merging, pruning, evolution.

Table des matières

Liste des tableaux

Liste des figures

CHAPITRE 1

Introduction

1.1 Contexte de travail

Notre travail se situe dans le contexte du **Web sémantique (WS) et la recherche d'information (RI)**. Le WS a pour objectif de structurer les informations disponibles sur le Web. Pour cela, les ressources doivent être sémantiquement enrichies par des métadonnées afin de permettre d'optimiser l'accès à ces ressources. Ces métadonnées sont actuellement spécifiées à l'aide des deux standards qui utilisent le langage XML : RDF et les Topic Maps qui permettent d'annoter le texte de la ressource Web afin d'en faciliter la recherche.

Selon la norme ISO/IEC 13250[1], les Topic Maps sont définies comme : « *Une technologie de codage de connaissances permettant de relier ces connaissances encodées aux ressources d'information pertinentes. Les Topics Maps sont organisées autour de Topics, qui représentent des sujets du monde du discours, des associations représentant les relations entre les sujets et des occurrences qui permettent de relier les sujets aux ressources d'informations pertinentes*. »

Les Topic Maps permettent d'une part de structurer le domaine de connaissance et d'autre part de classifier les documents dans un même formalisme. Une Topic Map est définie comme une structure sémantique organisée autour de Topics et de relations (ou associations) entre ces Topics, permettant d'organiser des contenus et des connaissances de différentes sources dans le but de faciliter la navigation et aider l'utilisateur à retrouver l'information pertinente dans ces contenus. Les Topic Maps peuvent être utilisées pour des portails. Dans le cadre de ressources multiples, la fusion de Topic Maps permet d'intégrer l'accès à plusieurs ressources.

Un Topic représente n'importe quel sujet à rechercher dans les ressources, il peut correspondre à une idée, un thème, un concept, un objet du monde réel (par exemple une personne ou un endroit), etc. Un Topic doit avoir un nom de base (*base name*), avec éventuellement des variantes (*variant names*) et peut s'insérer dans une hiérarchie de types. Tout Topic a une ou plusieurs occurrences qui regroupent toutes les informations permettant d'accéder aux différentes ressources concernées par ce Topic. Des associations « n-aires » peuvent relier les Topics et chaque Topic a un rôle dans une association. La limite de validité des Topics et des associations est définie par un contexte *(scope)* qui peut être par exemple une date ou une langue. Des propriétés (de type attributs-valeurs) tels que le langage ou le profil de l'utilisateur peuvent être associées aux ressources au moyen de facettes. Des liens

[1] http://www.isotopicmaps.org/sam/sam-model/

d'occurrence relient les Topics aux ressources. Le modèle des Topic Maps dispose d'un langage de spécification XTM (XML Topic Map) définie par le consortium TopicMaps.Org.

La finalité de nos travaux est de spécifier, concevoir et développer un outil d'aide à la construction de Topic Maps. L'aide fournie par un outil peut être de plusieurs natures :

- Assistance à la création manuelle (ex : éditeurs) ;
- Création automatique par l'outil.

L'état de l'art montre que des éditeurs d'aide à la spécification de Topic Maps ont déjà été proposés et dans cette thèse l'accent sera mis sur le deuxième type d'aide : la création automatique. Il s'agit de construire automatiquement l'ontologie, les autres liens sémantiques et les liens d'occurrence reliant les Topics aux documents, en essayant d'utiliser toutes les ressources possibles. Ce peut être à partir de textes en langage naturel, de bases de données existantes, de recherche sur le Web, par réutilisation d'ontologies existantes (fusion de plusieurs ontologies, expansion d'une ontologie existante, …), par analyse des requêtes de recherche de documents, etc.

Un des problèmes majeurs d'accès à l'information sur le Web est actuellement le multilinguisme. Les Topic Maps doivent pouvoir aider à unifier l'accès à des ressources multilingues. Une partie de la thèse sera consacrée à l'étude de production de Topic Maps permettant d'accéder à des ressources de différentes langues, par un même portail. Il s'agira dans un premier temps de choisir les meilleurs mécanismes disponibles dans les Topic Maps pour gérer le multilinguisme, puis de trouver des algorithmes de production de Topic Maps multilingues.

La Topic Map ainsi construite servira comme une cible de recherche pour l'utilisateur, il pourra naviguer dans les documents sources à travers les liens de la Topic Map ou l'interroger au moyen de requêtes en utilisant un langage de requête approprié.

L'application choisie pour tester l'outil développé sera le domaine de l'architecture durable, c'est-à-dire la recherche de méthodes de constructions écologiques préservant l'environnement et les sources d'énergies.

1.2 Problématique

Les Topic Maps conçues à l'origine comme un équivalent électronique d'index traditionnels, répondent à l'heure actuelle à un besoin d'organiser autour d'une vision métier des contenus et connaissances de différentes sources et de différentes langues. Grâce au

réseau de liens sémantiques entre les sujets qu'elles représentent, elles permettent une navigation facile et sélective améliorant ainsi la recherche de l'information dans les contenus. Un contenu à organiser étant très souvent volumineux et sujet à enrichissement perpétuel, il est pratiquement impossible d'envisager une création et gestion d'une Topic Map, le décrivant, de façon manuelle. Plusieurs travaux de recherche se sont penchés sur cette problématique. Beaucoup de propositions ont concerné la construction de Topic Maps à partir de ressources hétérogènes [Ellouze et al. 2008a]. Elles se distinguent principalement par les sources en entrée qui peuvent être des documents XML, des métadonnées RDF, des documents textuels, ou des bases de connaissances existantes telles que les thésaurus et la nature des techniques utilisées pour la production de Topic Maps.

La plupart des approches de construction de Topic Map proposent de réutiliser et d'adapter des techniques existantes telles que celles reliées à la conception de schémas conceptuels, à la fusion de donnés, aux techniques d'apprentissage et aux techniques de traitement automatique des langages naturels. Des travaux proposent de réutiliser des méthodes de construction d'ontologies et d'autres approches décrivent des processus de construction collaborative de Topic Maps.

Cependant, ces approches présentent plusieurs limites, en effet, à notre connaissance, aucune des approches existantes, excepté celle de Kasler [Kasler et al. 2006] qui utilise à la fois des documents écrits en anglais et en hongrois, ne permet l'obtention d'une Topic Map à partir d'un contenu multilingue. De plus, aucune des approches existantes n'exploitent plusieurs sources d'informations pour la construction d'une Topic Map. Bien que les Topic Maps soient, par définition, orientées utilisation (recherche d'information), peu d'entre elles prennent en compte les requêtes des utilisateurs pour la création et l'enrichissement de la Topic Map.

Par ailleurs, ces approches ne proposent pas de techniques pour l'évaluation de la qualité de la Topic Map générée et enfin, nous remarquons que le guidage méthodologique de l'utilisateur n'est pas pris en compte dans la majorité des travaux existants dans le domaine de la construction de Topic Map.

Dans ce contexte, nous proposons une approche incrémentale et évolutive de construction de Topic Map basée sur un processus automatisé, qui prend en compte des documents multilingues et l'évolution de la Topic Map selon le contenu et l'usage au cours du temps.

Notre approche prend comme entrée un référentiel sémantique construit à partir des documents textuels multilingues segmentés thématiquement et indexés sémantiquement et un thésaurus du domaine pour l'ajout de liens ontologiques et structurels entre les Topics. Pour l'enrichissement de la Topic Map, nous nous basons sur deux ontologies générales représentant la terminologie du langage commun dans les deux langues anglais et français afin de rajouter de nouveaux synonymes aux Topics et de nouveaux liens entre eux. Nous proposons aussi de prendre en compte l'usage de la Topic Map à travers l'intégration des requêtes des utilisateurs dans la Topic Map et la mise en œuvre de liens d'usage entre les questions potentielles extraites des sources d'interrogations disponibles relatives aux documents sources (les FAQ par exemple) et les réponses associées.

1.3 Contributions

Dans cette thèse, nous proposons ACTOM, une **A**pproche de **C**onstruction d'une **TO**pic Map **M**ultilingue ». Cette Topic Map sert à organiser un contenu multilingue composé de documents textuels. Elle a pour objectif de faciliter la recherche d'information dans ce contenu. Les principales contributions de cette thèse peuvent être résumées comme suit :

- Dans ACTOM, **nous produisons, en plus des liens reliant un Topic à ses ressources (les occurrences), deux autres catégories de liens : (a) les liens ontologiques et structurels et (b) les liens d'usage.** Les liens ontologiques et structurels regroupent les liens de spécialisation, le lien de composition ainsi que les liens associatifs, tels que ceux définis dans le standard (ANSI/NISO Z39.19-2005), que nous pourrions identifier suite à l'analyse des documents à organiser. Le lien d'usage est un hyper lien de type « répond à » (hyper lien questions/réponses) entre la question représentée comme un Topic et les réponses associées, c'est-à-dire les Topics référençant les documents qui permettent de répondre à la question. Nous proposons dans ce contexte de relier la question à chacun des mots clés la constituant via un hyper lien de type « est composé de » ;
- Nous proposons **d'étendre le modèle des Topic Map** défini par l'ISO en rajoutant aux caractéristiques d'un Topic **des méta-propriétés** servant à mesurer la pertinence des Topics dans le temps. Ces méta-propriétés sont initialisées à la création de la Topic Map. Ils nous renseignent sur l'importance des Topics et sur l'usage qu'on en fait lors de l'exploitation de la Topic Map. Ils sont utilisés pour

la gestion des évolutions de la Topic Map, plus précisément pour l'élagage de Topics considérés non pertinents ;

- Notre approche a pour originalité de construire **un référentiel de documents indexés thématiquement et sémantiquement** en associant à chaque document, la liste de ses segments qui représentent les thématiques abordés dans le document et la liste des termes et des concepts représentatifs de son contenu, pondérés avec leurs degrés de pertinence. Ce référentiel sert, d'une part, à compléter notre méta-modèle de Topic Map avec en particulier la possibilité d'indexer un Topic par un segment de document et d'autre part, à la construction et l'annotation de la Topic Map ;
- Notre approche **conjugue l'utilisation de quatre sources d'information pour la construction de la Topic Map** : **(a)** le référentiel de documents ; **(b)** un thésaurus bilingue (Français/Anglais) du domaine, **(c)** deux ontologies générales (WordNet pour l'anglais et WOLF pour le français) , et **(d)** un ensemble de scénarios d'usage sous forme de questions/réponses que nous recensons à partir de sources d'interrogations possibles telles que les questions relatives aux documents sources qu'un expert du domaine ou un utilisateur quelconque peut poser, les foires aux questions (FAQ), les traces des discussions téléphoniques et des consultations directes avec les travailleurs du domaine. De part son processus incrémental, notre approche est réutilisable à chaque enrichissement du contenu qu'elle organise ;
- Notre approche est **incrémentale et évolutive**, elle consiste à créer une Topic Map à partir de chaque document du référentiel en se basant sur le thésaurus du domaine pour l'ajout de liens ontologiques. Pour enrichir cette Topic Map, nous utilisons, d'une part les deux ontologies générales, et d'autre part les scénarios d'usage pour l'ajout de nouveaux Topics et des liens d'usage. Ce processus est répété pour tous les documents disponibles. Les Topics Maps résultantes sont intégrées deux à deux afin d'obtenir la Topic Map globale. **Pour l'étape d'intégration**, nous mettons l'accent sur la reconstruction de hiérarchies en adaptant les algorithmes proposées par notre équipe [Lammari et Métais, 2004], [Lammari et al. 2008] et en prenant en compte le thésaurus du domaine et les deux ontologies générales qui fournissent les chainons manquants ;

- ACTOM vise à produire une Topic Map globale permettant de structurer sémantiquement des concepts dans **plusieurs langues** tout en prenant en charge une des spécificités du multilinguisme qui est l'absence éventuelle de termes sémantiquement équivalents d'une langue à une autre ; ce qui est assez fréquent lorsque les contenus sont issus de différentes cultures. La Topic Map résultante offrira à l'utilisateur la possibilité de s'enrichir de connaissances se trouvant dans des documents écrits dans une langue autre que la sienne ;
- Nous abordons **la qualité de la Topic Map** à travers la définition d'une méthode **d'élagage dynamique (*pruning*)** de la Topic Map au niveau de l'affichage et ce grâce aux méta-propriétés des Topics ;
- La Topic Map servira comme **une cible de recherche** pour l'utilisateur. Cette recherche peut être : (1) Une recherche **exacte** qui permet d'effectuer une recherche « précise » grâce à un langage de requête approprié, dans le cas ou l'utilisateur connait ce qu'il cherche ; (2) Une recherche **connotative**, lorsque l'utilisateur exprime sa requête en langage naturel et le système reconnait les Topics subsumés par les termes de la requête ; (3) Une recherche **exploratoire** définie quand l'utilisateur veut se faire une idée du contenu du corpus. Il peut découvrir tous les domaines connexes à son centre d'intérêt qui peut s'étendre pour découvrir de nouveaux documents abordant des sujets connexes ; et (4) Une recherche **thématique** utilisée lorsque l'utilisateur cherche à explorer le corpus sur un thème particulier. Il peut alors naviguer dans l'espace sémantique représenté par la Topic Map globale et visualiser les documents appartenant à cette thématique ;
- Nous avons proposé une **recherche par navigation** sémantique basé sur le modèle des Topic Maps qui a pour intérêt de procurer à l'utilisateur une représentation des connaissances proche du modèle cognitif qu'il a sur le domaine. Nous mettons à la disposition de l'utilisateur un espace sémantique de navigation permettant de réaliser les différents types de recherche citées précédemment, de représenter au mieux la sémantique du domaine et les différentes relations (hiérarchique, similarité et associatives) existantes dans les documents sources et ce pour toutes les facettes linguistiques. Il s'agit également d'aider l'utilisateur à retrouver le plus rapidement possible les connaissances qu'il cherche en lui offrant des moyens de représentation et de visualisation lui

permettant de filtrer les connaissances et sélectionner uniquement celles qui sont pertinentes pour lui ;

- Nous avons **réalisé et implémenté une plateforme** permettant de mettre en œuvre l'approche proposée. L'application choisie pour tester la plateforme développée sera le domaine de la construction durable, en particulier les solutions pour l'économie d'énergie préservant l'environnement.

1.4 Organisation du mémoire

Ce mémoire est organisé en sept chapitres : Dans le **premier chapitre** (chapitre introductif), nous présentons le contexte de nos travaux de thèse, notre problématique ainsi que les contributions que nous avons réalisées dans ce travail.

Dans **le deuxième chapitre**, intitulé « Etat de l'art », nous présentons une vue d'ensemble des différents champs de recherche concernés par notre problématique à savoir le Web sémantique et les langages et les ressources disponibles pour l'annotation de ressources documentaires dans le cadre du WS (RDF, Topic Map, les ontologies). Nous décrivons ensuite les problèmes liés à la recherche d'information multilingues et les travaux existants pour les résoudre. Nous avons également réalisé un état de l'art détaillé sur les approches de construction de Topic Map, dans cet état nous avons passé en revue les principaux domaines de recherche intervenant dans la construction de Topic Map tels que ceux relatifs à la construction d'ontologies, aux techniques d'analyse linguistiques de documents textuels et aux méthodes d'intégration et de fusion d'ontologies. Nous avons par la suite réalisé une étude comparative entre les différentes approches et outils d'élaboration de Topic Map. Nous terminons ce premier chapitre par une synthèse de l'état de l'art qui comprend les critiques que nous avons pu dégagées sur les travaux existants et introduit l'approche que nous proposons.

Le **troisième chapitre** de ce mémoire, intitulé « Approche générale et méta-modèles », présente notre approche générale de recherche intelligente dans un référentiel de documents indexés sémantiquement et thématiquement fondée sur le modèle des Topic Maps. Dans ce chapitre, nous détaillons les deux méta-modèles que nous proposons, le méta-modèle de Topic Maps et celui du référentiel de documents multilingues annotés descriptivement et indexés sémantiquement et thématiquement. Nous terminons ce chapitre par expliciter les différents modes de recherche offerts par notre approche.

Le **quatrième chapitre**, intitulé, « Description détaillée de notre approche » s'intéresse à la description détaillée de notre approche ACTOM de construction de la Topic Map multilingue enrichie et annotée à partir du référentiel de documents segmentés thématiquement et indexés sémantiquement, du thésaurus du domaine, de deux ontologies générales, et d'un ensemble de scénarios d'usage sous forme de questions/ réponses extraites à partir de FAQ. Nous décrivons également la démarche de construction du référentiel à partir d'un ensemble de documents textuels multilingues. Nous présentons ensuite les problèmes que nous avons rencontrés lors de la construction liés au multilinguisme et les solutions que nous proposons pour les résoudre. Dans ce chapitre, nous mettons en exergue les différents choix réalisés tout en les justifiant et en présentant leurs avantages. Enfin nous terminons par un bilan pour mettre en évidence l'originalité de notre proposition par rapport aux travaux existants.

Dans **le cinquième chapitre**, nommé « Prise en compte de la qualité : méthode d'élagage de la Topic Map », nous nous intéressons à la qualité de la Topic Map générée, en particulier nous présentons le processus d'élagage évolutif à l'affichage de la Topic Map à travers la définition de notes (scores) attribuées à chaque Topic. Ces notes sont définies comme des méta-propriétés du Topic et nous renseignent sur son importance et sa pertinence par rapport à son usage dans le temps.

Le **sixième chapitre**, nommé « Plateforme de mise en œuvre de l'approche proposée et expérimentations », a pour but de décrire la plateforme mettant en œuvre l'approche et les modèles proposés et les résultats que nous avons obtenus grâce à nos expérimentations.

Le **dernier chapitre** permet de conclure le travail réalisé et propose des perspectives.

CHAPITRE 2

Etat de l'art

Ce chapitre présente le contexte de nos travaux de recherche, nous commençons tout d'abord par décrire brièvement les modèles et les langages proposés pour l'annotation sémantique des ressources dans le cadre du Web sémantique afin de faciliter la recherche d'information. Nous présentons en particulier le standard RDF et les ontologies proposés par le W3C et le modèle des Topic Maps proposé par l'ISO. Une étude comparative nous permet de dégager les particularités du modèle des Topic Map et ses avantages pour la recherche d'information. Dans la deuxième section de ce chapitre, nous étudions les problématiques liées à la recherche d'information multilingue et les approches proposées dans la littérature pour les résoudre. La troisième section est consacrée à la description des travaux existants pour la construction de Topic Map, nous réalisons une étude comparative de ces travaux selon des critères que nous définissions. A partir d'une synthèse de l'état de l'art, nous terminons par énoncer notre proposition pour la construction d'une Topic Map à partir de documents textuels multilingues, cette Topic Map servira comme une cible de recherche d'information pour l'utilisateur.

2.1 Le Web sémantique appliqué à la recherche d'information

2.1.1 Introduction au Web sémantique

Tim Berners-Lee [Berners-Lee et al. 2001] (fondateur et président du *Consortium World Wide Web* 'W3C') a attribué l'expression du Web sémantique en faisant référence à la vision du Web de demain comme un vaste espace d'échange de ressources entre humains et machines permettant une meilleure exploitation de masses de données disponibles sur le Web. L'objectif n'est pas de permettre aux machines de se comporter comme des êtres humains, mais simplement de développer des langages pour exprimer des informations d'une manière traitable par des machines. En effet, le traitement automatisé des données requiert une représentation de la sémantique compréhensible et échangeable par les machines.

Le Web sémantique peut être défini comme un substrat supportant des fonctions avancées pour la collaboration (homme-homme, homme-machine, machine-machine), qui permet de partager des ressources et de raisonner sur le contenu de ces dernières [Berners-Lee et al. 2001]. L'idée est de rendre explicite la sémantique des documents au travers de métadonnées ou d'annotations, afin de permettre aux agents logiciels d'effectuer des tâches de recherche et de sélection des ressources pour les utilisateurs.

Les recherches actuellement réalisées dans le domaine du Web sémantique s'appuient sur un existant riche venant de différents domaines. Par exemple, les systèmes de recherche

en Intelligence Artificielle, les systèmes de représentation et/ou l'ingénierie des connaissances ont permis d'étudier les problèmes liés à l'accès aux collections d'informations structurées, aux règles d'inférences et aux raisonnements automatiques bien avant le développement du Web. Cependant, l'application des résultats de ces recherches pose d'autres problèmes dus au changement du contexte de déploiement, le Web et ses dérives (Internet, Extranet, Intranet), la nécessité d'un niveau élevé d'interopérabilité, la diversité des usages, les standardisations, etc. Le défi du Web sémantique est de fournir un langage [Legrand, 2001] :

- Qui exprime à la fois les données et les règles de raisonnement sur ces données ;
- Qui permette aux règles de n'importe quel système de représentation des connaissances d'être transférées sur le Web.

L'architecture du Web sémantique proposée par W3C (*World Wide Web Consortium*) s'appuie sur une pyramide de langages dont seulement les couches basses sont aujourd'hui relativement stabilisées. La figure 2.1 montre une des visions de l'organisation en couches proposée par le W3C [Beners-Lee, 1998].

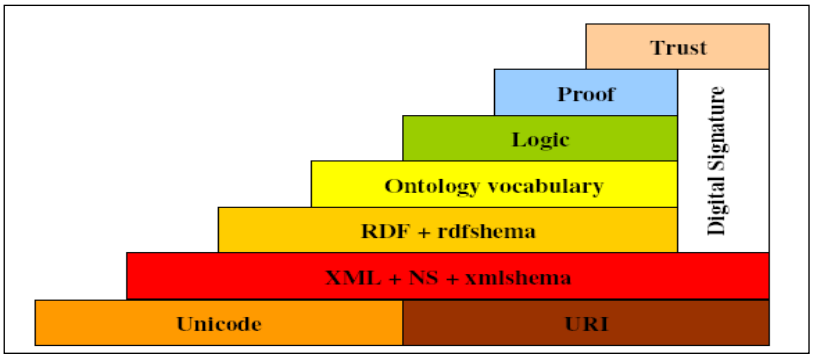


Figure 2.1 Architecture du Web sémantique [Beners-Lee, 1998]

Au niveau le plus bas se trouvent les données brutes codées par le standard Unicode, ces données possèdent une adresse URI (*Uniforme Ressource Identifier*) qui permet d'attribuer un identifiant unique à un ensemble de ressources. Ces données peuvent être structurées grâce à un langage de balises tels que XML (*eXtensible Markup language*), NS (*NameSpace*) ou *xmlshema*. La syntaxe XML peut être considérée comme un premier niveau de sémantique, elle permet aux utilisateurs de structurer les données en fonction de leur contenu sans rien dire de la signification des structures.

Pour attribuer une signification à cette structure et relier d'une façon pertinente les différents éléments, Tim Berners-Lee propose le standard RDF comme standard de

représentation, développé par le W3C. Les Topic Maps[2] ont été définies par l'ISO (*International Standards Organisation*) pour accomplir la même tâche.

Ces langages ont pour but de donner une organisation plus structurée des informations présentes sur le Web à travers une description sémantique des données fournies par XML.
La signification sémantique des données XML représentées par RDF ou Topic Maps, est largement insuffisante pour assurer une bonne distinction des différents concepts. Ce problème peut être résolu grâce à l'utilisation des ontologies.

Dans ce qui suit, nous présentons les techniques de structuration sémantique des données et les ressources utilisées pour l'annotation de ces données dans le cadre du Web sémantique, nous présentons tout d'abord les thésaurus, nous évoquons ensuite la notion de métadonnées et après, nous présentons trois standards de représentation de connaissances RDF/RDFS, Topic Maps et OWL. RDF, RDFS et OWL sont développés par le W3C, les Topic Maps par l'ISO.

2.1.2 Modèles de représentation de connaissances dans le cadre du Web sémantique

Le processus d'annotation sémantique consiste à ajouter (semi-)automatiquement des métadonnées structurées aux ressources documentaires du Web. Les annotations décrivent aussi bien le document dans son ensemble, comme son titre, son auteur, etc., que son contenu par des descripteurs provenant de ressources terminologiques et ontologiques [Bourigault et al. 2004] (comme les taxonomies, thésaurus, ontologies) pour normaliser la sémantique des annotations documentaires comme celle des concepts du domaine concerné.

Nous allons présenter les deux principales ressources permettant de représenter et de modéliser la connaissance d'un domaine : les thésaurus et les ontologies ensuite nous présentons les principaux langages et modèles utilisés pour l'annotation sémantique des ressources dans le cadre du Web sémantique.

2.1.2.1 Les Thésaurus

Un thésaurus est fondé sur une structuration hiérarchisée d'un ou plusieurs domaines de la connaissance et dans lequel les notions sont représentées par des termes d'une ou plusieurs langues naturelles et les relations entre notions par des signes conventionnels (AFNOR 1987). Les normes (ISO 2788 et ANSI Z39) ont permis d'uniformiser leur contenu en termes de

[2] International Organisation for standardization (ISO), International Electronichnical Commission (IEC) "Topic Map, Internationl Standard ISO/IEC 13250 :1999", 19 April 1999, http://www1.y12.doe.gov/capabilities/sgml/sc34/document/0322_files/iso13250-2nd-ed-v2.pdf

relations entre unités lexicales : équivalence, relations hiérarchiques et relations non taxonomiques (liens associatifs).

Les relations présentes dans un thésaurus répondant aux normes ANSI Z39 et ISO 2788 sont rappelées dans la figure 2.2. Ces relations sont « est plus spécifique que », «est plus générique que», « utiliser plutôt » et « utiliser pour désigner ».

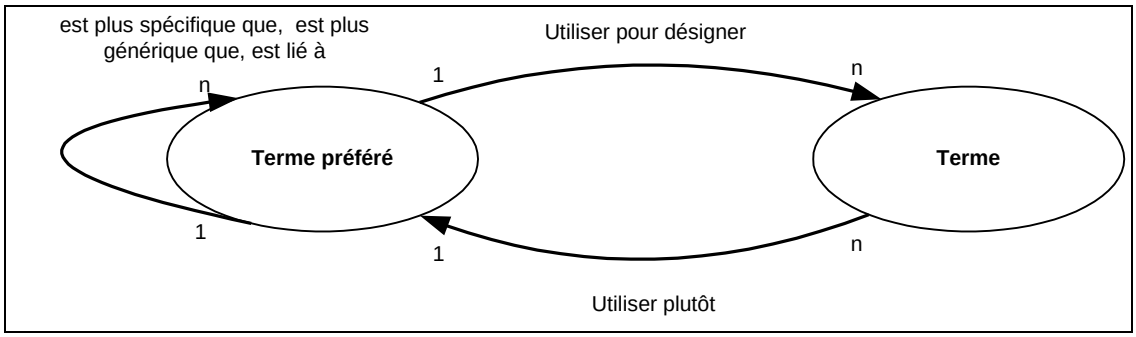


Figure 2.2 Relations entre les termes d'un thésaurus selon les normes ANSI Z39 et ISO 2788

Un thésaurus est donc considéré comme un vocabulaire contrôlé et structuré dans lequel les relations entre les termes du domaine considéré sont clairement spécifiées formant ainsi un réseau terminologique. La structuration hiérarchisée correspond à la relation d'hyperonymie permettant de structurer les termes du vocabulaire. On dit alors qu'un terme X est plus générique que (EPG) ou est plus spécifique que (EPS) qu'un terme Y, par exemple « Véhicule » a un sens plus général que « Automobile ».

D'autres relations constituent le réseau terminologique comme les relations de synonymie et les relations associatives (figure 2.2).

Synonymie : un terme X est utilisé pour désigner (UPD) ou utilisé plutôt (UP) qu'un terme Y, par exemple « Voiture » et « Automobile ».

Associative : un terme X est lié à un terme Y s'il y a une sorte de relation non sémantiquement spécifiée entre les deux, par exemple le terme « Conduite » est souvent associé au terme «Véhicule ».

Nous voulons ici souligner le fait que les thésaurus ne sont pas des ontologies : ils permettent de modéliser le vocabulaire d'un domaine ou d'une application mais ne fournissent pas de représentation de la connaissance de ce domaine ou de cette application. Par contre, ils peuvent être comme ressources pour l'aide à la création des ontologies [Charlet et al. 2004] ou de Topic Maps comme nous le verrons plus tard dans notre approche.

D'un point de vue de la représentation des connaissances, les thésaurus ont un faible degré de formalisation. Ce sont des collections de termes qui sont organisées suivant une ou plusieurs hiérarchies avec des relations entre termes. Les thésaurus n'ont pas de niveau

d'abstraction conceptuelle [Soergel et al. 2004]. La distinction entre un concept et sa lexicalisation n'est pas clairement établie. Les relations de synonymies sont établies entre les termes mais les concepts ne sont pas identifiés. Ceci s'explique par l'utilisation initiale des thésaurus, qui n'ont pas pour objectif de refléter comment le monde peut être compris en termes de sens mais en termes de terminologie et de catégories servant à l'indexation manuelle de documents d'un domaine.

De plus, la couverture sémantique des thésaurus est limitée. En effet, les relations entre termes sont vagues et ambiguës. Les liens sémantiques qu'ils contiennent reflètent parfois l'utilisation prévue du thésaurus plutôt que les liens sémantiques réels entre termes. Ces relations peuvent ainsi englober les relations « est une instance de » ou « est une partie de ». La relation associative « est lié à » est souvent difficile à exploiter car elle connecte des termes en considérant différents types de relations sémantiques.

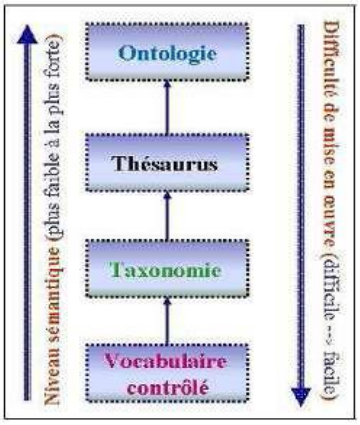


Figure 2.3 Différence entre les modèles terminologiques et ontologiques

2.1.2.2 Les ontologies

Le terme ontologie est issu du domaine de la philosophie de la connaissance. Il désigne l'ensemble des concepts d'un domaine ainsi que leurs relations. En Intelligence Artificielle [Smith, 2001], le terme ontologie désigne une organisation des concepts d'un domaine. Concrètement, une ontologie est une bibliothèque de termes ou des définitions de concepts, qui décrivent la structure de l'information pour un domaine donné ou une activité particulière. En 1993, Gruber a proposé sa définition qui reste jusqu'à présent la définition la plus citée dans les écrits en intelligence artificielle : « *An ontology is an explicit specification of a conceptualization* » [Gruber, 1993]. Depuis, plusieurs définitions de l'ontologie ont été proposées.

En 1997, Borst modifie légèrement la définition proposée par Gruber en énonçant : « *une ontologie est définit comme étant une spécification formelle d'une conceptualisation*

partagée » [Borst, 1997]. Cette définition de l'ontologie a ensuite été affinée par R. Studer et al. [Studer et al. 1998] comme « *spécification formelle et explicite d'une conceptualisation partagée* » :

- *Formelle :* l'ontologie doit être lisible par une machine, ce qui exclut le langage naturel ;
- *Explicite :* la définition explicite des concepts utilisés et des contraintes de leur utilisation ;
- *Conceptualisation :* le modèle abstrait d'un phénomène du monde réel par identification des concepts clefs de ce phénomène ;
- *Partagée :* l'ontologie n'est pas la propriété d'un individu, mais elle représente un consensus accepté par une communauté d'utilisateurs.

Une **ontologie** permet d'exprimer les concepts d'un domaine en les organisant hiérarchiquement et en définissant leurs propriétés sémantiques dans un langage de représentation des connaissances formel favorisant le partage d'une vue consensuelle sur ce domaine entre les applications informatiques qui en font usage [Bourigault et al. 2004].

Les **concepts** (aussi appelées « classes ») représentent les objets, abstraits ou concrets, réels ou fictifs, élémentaires ou composites, du monde réel. Ces concepts sont organisés par l'utilisation de la relation de subsumption, dans laquelle ils peuvent appartenir à plusieurs sur-concepts différents. Les **relations** représentent des interactions entre concepts permettant de construire des représentations complexes de la connaissance du domaine [Charlet et al. 2004].

Les **attributs** correspondent à des caractéristiques, des spécificités particulières, attachées à un concept et qui permettent de le définir de manière unique dans le domaine [Charlet et al. 2004]. Leurs valeurs sont littérales, i.e. de type primitif, comme une chaîne de caractère ou un nombre entier. Par exemple, un concept « Personne » peut avoir les attributs suivants : un « numéro de carte d'identité », une « date de naissance », etc.

2.1.2.3 Les métadonnées

Une métadonnée (du Grec, "méta", ce qui dépasse, englobe) est une donnée à propos d'une autre donnée. En sciences de l'information, les métadonnées sont des ensembles de données structurées décrivant des ressources physiques ou numériques, ou, sur un plan plus fonctionnel, « de l'information structurée qui décrit, explique, localise la ressource et en facilite la recherche, l'usage et la gestion » [NISO, 2004].

Nous utilisons généralement les métadonnées pour parler d'information descriptive à propos de ressources du Web. Toutefois les métadonnées peuvent répondre à de nombreux

objectifs, que ce soit l'identification d'une ressource satisfaisant un besoin particulier d'information, l'évaluation de sa pertinence ou enfin pour garder la trace des caractéristiques d'une ressource à des fins d'entretien ou d'utilisation à long terme. De nos jours, différentes communautés d'usagers comblent de tels besoins en utilisant une grande variété de normes de métadonnées. Afin de décrire une ressource, nous utilisons des métadonnées. Une notice contenant un ensemble d'attributs, ou éléments, nécessaires pour décrire la ressource en question est établie.

Par exemple, dans les bibliothèques, nous utilisons un ensemble de notices de métadonnées comprenant des éléments spécifiques tels que : auteur, titre, date de création ou de publication, sujet et cote, afin de retrouver un livre ou un document sur les tablettes. Le lien entre une notice de métadonnées et la ressource qu'elle décrit peut être fait de deux façons :

- Les métadonnées peuvent être contenues dans une notice séparée du document, comme c'est le cas pour une notice dans un catalogue de bibliothèque ;
- Les métadonnées peuvent être intégrées dans la ressource elle-même.

2.1.2.4 RDF/RDFS

RDF[3] (*Ressource Description Framework*) est né en 1997 à l'initiative du W3C Norme, comme son nom l'indique, RDF est un métalangage servant à encadrer la description de ressources, permettant de rendre plus structurée l'information nécessaire aux moteurs de recherche et, plus généralement, nécessaire à tout outil informatique analysant de façon automatisée des ressources Web. Pour ce faire, RDF propose de décrire une ressource par une forme de triplets (sujet, relation, objet) dont le sujet est un identificateur de la ressource elle-même.

La représentation d'un ensemble de ressources est un graphe étiqueté (diagramme noeuds et arêtes). Les nœuds, représentés par des ovales ou par des rectangles dans le cas des chaînes littérales, sont des URIs (*Uniforme Resource Identifier*) spécifiant des ressources ou bien des littéraux (dans le cas des noeuds feuille). Les arêtes spécifient des relations (ou des propriétés déclarées) entre ressources ou entre ressources et littéraux. Il est possible de définir un triplet RDF en un seul nœud ce qui permet de définir des relations d'arrité supérieures à deux. RDF permet donc de représenter des métadonnées attachées à des ressources. En tant

[3] http://www.w3.org/RDF/

que standard, il favorise l'interopérabilité entre les applications qui échangent des ressources sur le Web en facilitant le traitement automatique [Abel, 2004].

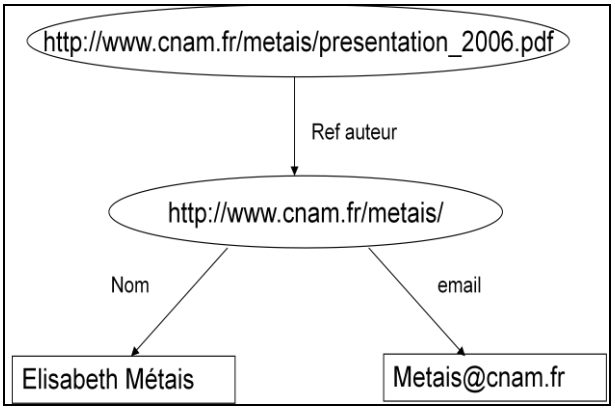


Figure 2.4 Exemple d'annotation sémantique en RDF

Le but général de RDF est de définir un modèle permettant de décrire les ressources sans préjuger du type d'utilisation que nous pourrons en faire. La définition du modèle est également neutre par rapport au domaine d'appartenance des ressources. Pour faciliter la définition des métadonnées, il est possible de définir des classes de ressources. C'est dans cette perspective que les Schémas RDF ont été créés.

Une collection de classes est appelée un schéma RDF. Les classes sont organisées en hiérarchie, et offrent une extensibilité grâce à la spécialisation en sous-classes. Au travers de l'acceptation et le partage de schémas, RDFS[4] facilite la réutilisation des définitions de métadonnées. En effet, RDFS ajoute à RDF la possibilité de définir des hiérarchies de classes et de propriétés dont l'applicabilité et le domaine de valeurs peuvent être contraints à l'aide des attributs *rdfs:domain* et *rdfs:range* [Laublet, 2007].

Cependant les schémas RDF ont des limitations qui ne permettent généralement pas leur emploi pour la description d'ontologies. Ces limitations concernent par exemple la définition de classes à partir d'opérateurs (intersection, complément) ou à partir de contraintes (restriction sur la cardinalité de propriétés).

2.1.2.5 Le langage OWL

Le langage OWL[5] (*Ontology Web Language*), recommandé par le W3C en février 2004, est le plus expressif des langages ontologiques pour le Web [Baget et al. 2004]. La conception d'OWL a bénéficié de plusieurs générations de langages de représentation des connaissances,

[4] RDF Schema: http://www.w3.org/TR/rdf-schema/
[5] http://www.w3.org/TR/owl-features/

d'une base théorique solide en logique et d'une volonté de la part de ses concepteurs pour créer un langage approprié à une utilisation dans le cadre du Web sémantique.

Le langage OWL sert à décrire des concepts et des relations ou attributs. Il a été conçu pour être utilisé par les applications qui traitent le contenu de l'information. Il facilite l'interopérabilité en fournissant plus de vocabulaire pour décrire les classes et les propriétés comme les approches orientées objets. Par exemple : les relations entre les classes, les cardinalités, l'égalité, le typage plus riche des propriétés, les caractéristiques des propriétés, etc.

OWL a été conçu pour satisfaire le besoin d'un langage d'ontologie du Web. Il ajoute plus de vocabulaire pour décrire les propriétés et les classes. Nous pouvons citer entre autre : les relations entre classes (par exemple la disjonction), les cardinalités (par exemple exactement un), l'égalité, typage plus riche des propriétés, caractéristiques des propriétés (par exemple la symétrie) et les classes énumérées. Bien que OWL soit dérivé de DAML+OIL qui est équivalent à un langage très expressif de la logique des descriptions, il n'est en tant que tel, équivalent à aucune logique de description. Certaines caractéristiques font qu'il n'existe aucun algorithme d'inférence décidable avec cette puissance d'expression. OWL fournit trois sous langages de plus en plus expressifs conçus pour l'usage des communautés spécifiques des utilisateurs et des développeurs : OWL Lite, OWL DL et OWL Full.

Le language **OWL Lite** convient aux utilisateurs qui ont principalement besoin d'une hiérarchie de classification et de contraintes simples. Par exemple, alors qu'il supporte des contraintes de cardinalité, il autorise seulement des valeurs de cardinalité 0 ou 1. Il devrait être plus simple de fournir un outil d'aide pour OWL Lite que ses parents plus expressifs. OWL Lite fournit un chemin rapide de migration pour les thésaurus et autres taxonomies. Les préfixes *rdf:* ou *rdfs:* sont utilisés lorsque les termes sont déjà présents dans RDF ou RDFS. Sinon, les termes sont introduits par OWL. Ainsi le terme *rdfs:subPropertyOf* indique que *subPropertyOf* est dans le vocabulaire RDFS ; par contre le terme *owl:class* indique que class est introduit par OWL.

Le language **OWL DL** correspond à une logique de description expressive. Il convient aux utilisateurs qui demandent un maximum d'expressivité tout en maintenant la complétude (garantie de calculer toutes les conclusions) et la décidabilité (tous les calculs doivent fournir en un temps fini). OWL DL inclut tous les constructeurs du langage OWL, mais ils peuvent être utilisés seulement sous certaines restrictions. OWL DL est appelé ainsi en raison de sa correspondance avec les logiques de description.

Le language **OWL Full** convient aux utilisateurs qui demandent un maximum d'expressivité avec la liberté syntaxique de RDF sans aucune garantie de calcul. Par exemple, une classe peut être traitée comme une collection d'individus et en même temps peut être vue comme un seul individu. OWL Full permet aussi à une ontologie d'augmenter le sens du vocabulaire prédéfini (RDF et OWL) [Azouaou, 2005].

2.1.2.6 Les Topic Maps

Outre les développements menés par le W3C, d'autres initiatives comme les Topic Maps sont nées pour l'annotation sémantique des ressources d'information et pour tenter de mettre au point des langages pour la modélisation d'ontologies ou la création de bases de connaissances. Parmi ces initiatives, nous retiendrons celle des Topic Maps, concurrent du langage RDF.

La norme Topic Maps a pour origine le travail réalisé il y a une dizaine d'années au sein du groupe Davenport autour de la problématique des index de documentation. Un des problèmes importants abordé par ce groupe était l'échange et la fusion d'index de collections hétérogènes de documents, problème qui est devenu crucial avec l'expansion du Web. Ces réflexions menées par le groupe Davenport ont été postérieurement complétées (1996) par les réflexions d'un groupe de travail appelé *Conventions pour l'application de HyTime*[6], réalisé dans le contexte de l'organisme GCA[7] et en particulier par les membres de ce groupe Michel Biezunski et Steve Newcomb. Ce travail a abouti en 2000 à un standard ISO[8] (ISO/IEC 13250:2000). Bien que l'objectif d'origine des Topic Maps fût la gestion d'index, cet objectif s'est élargi à l'annotation sémantique de ressources distribuées en général. Un groupe de travail, *TopicMaps.org* constitué en 2000 travaille sur l'adaptation de ce standard aux données du Web et à son portage en XML[9].

Le but des Topic Maps est de fournir une façon standard de modéliser la connaissance contenue dans les ressources d'information (documents, bases de données, vidéos, etc.). Une Topic Map est une structure abstraite organisée autour de Topics représentant des sujets que le créateur de la Topic Map souhaite décrire et pour lesquels des ressources sont disponibles pour fournir de la connaissance sur ces sujets et de relations (ou associations) entre ces Topics. Un Topic peut représenter n'importe quel sujet (adressable ou non) ; une idée, un thème, un concept, un objet du monde réel (une personne, un endroit, etc.). Un Topic possède

[6]HyTime est une norme ISO (10744:1992) basée sur SGML pour la description de documents hypermédia.
[7]*Graphical Communication Association* devenu OASIS (*Organization for the Advancement of Structured Information Systems*)
[8]http://www.isotopicmaps.org/rm4tm/
[9]Ce groupe a publié en 2001 une version XML de la DTD ou modèle de document de la norme Topic Maps appelé XTM. La première version de la DTD était en SGML et est connue sous le nom de HyTM

toujours au moins un nom, son *baseName*, et des variantes (*variant Names*). Il peut avoir d'autres noms comme c'est le cas dans le monde réel où les objets peuvent être dénommés de différentes manières (le *displayName* pour l'affichage et le *sortName* pour la recherche).

Un Topic peut s'insérer dans une hiérarchie de types. Tout Topic a une ou plusieurs occurrences qui regroupent toutes les informations permettant d'accéder aux différentes ressources concernées par ce Topic. Des associations n-aires peuvent relier les Topics et chaque Topic a un rôle dans une association.

La limite de validité des Topics et des associations est définie par un « scope » (contexte) qui peut être par exemple une date. Des propriétés (de type attributs-valeurs) tels que le langage ou le profil de l'utilisateur peuvent être associées aux ressources au moyen de facettes. La figure 2.3 présente un exemple de Topic Map.

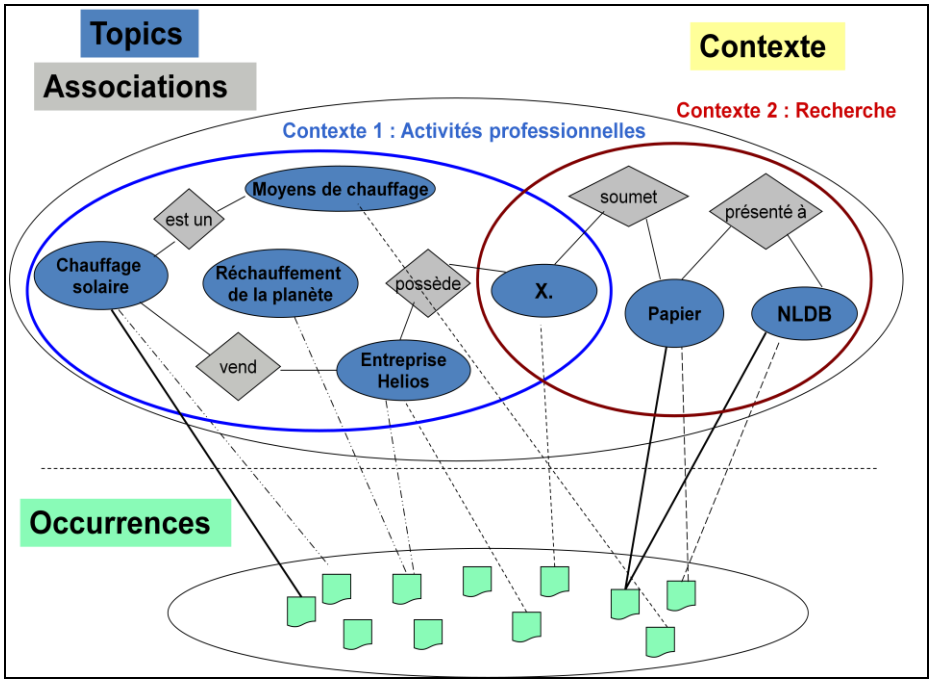


Figure 2.3 Exemple de Topic Map

Dans ce qui suit, nous décrivons en détail les composantes du modèle des Topic Maps.

Les Topics

Un **Topic** est la représentation informatique d'un *Sujet* appliqué à un ensemble de localisations (*Contexte*). Il peut désigner un auteur, une entité, un concept, etc. Le terme *Topic* réfère à un objet ou noeud d'une Topic Map qui représente un sujet. En d'autres termes, à un Topic correspond un sujet unique et inversement. Cependant, il peut arriver qu'un même sujet soit représenté par plusieurs Topics, dans le cas de la création de plusieurs Topic Map. Dans

une telle situation, il est nécessaire d'établir une seule et même identité pour les différents Topics. Ceci est réalisé par le concept de *subject indicator*. Tout Topic partageant un ou plusieurs *subjects indicators* sont considérés comme sémantiquement équivalents.

Prévus initialement pour représenter des ressources, les Topics peuvent servir aussi à représenter les entités d'une ontologie. Il est ainsi possible d'indiquer qu'un Topic est instance d'un autre, ce qui permet de faire un premier classement des Topics en type et instance. La définition des types des Topics dépend bien sûr de leur utilisation, des besoins de l'application et de la nature de l'information présente dans les documents [Pepper, 2000]. Plus généralement, un Topic est la représentation informatique d'un sujet plongé dans un contexte particulier. Dans un sens générique, un Topic est un objet composé de l'information qui le caractérise.

Un Topic peut avoir un ou plusieurs noms. Ces noms doivent permettre d'identifier un Topic sans ambiguïté au sein de la Topic Map, c'est-à-dire que deux Topics distincts ne peuvent pas partager un même nom à l'intérieur d'une même Topic Map. Sinon, lors du traitement de la Topic Map en utilisant le standard XTM, ils seront fusionnés. Le choix de ces noms doit être conforme aux usages dans lequel la Topic Map est utilisée.

Les Occurrences

Un Topic peut être lié à une ou plusieurs ressources d'information (figure 2.3), ressources pertinentes pour décrire ce Topic (sujet). Une **occurrence** (ressource d'information) peut être un article, une image, une vidéo, un rapport, un commentaire, etc. Ces occurrences se trouvent généralement en dehors de la Topic Map ce qui représente une séparation en deux niveaux (couches) des Topic Maps : d'une part les Topics et d'autre part leurs occurrences. Ces occurrences peuvent être classées selon leurs types : rapport de thèse, livre, article, image, son, etc, ou bien selon leurs langues : Français, Anglais dans le cas d'une application traitant des ressources multilingues.

Les Associations

Jusqu'à présent, les notions que nous avons abordées (Topic, sujet, type de Topic, occurrence) permettent d'organiser les ressources d'information selon un Topic ou sujet et de créer une indexation « directe ». La notion d'**Association** permet de créer des liens entre les différents Topics (figure 2.3). Créer des liens entre Topics permet de structurer les Topics les

uns par rapport aux autres et donc d'offrir aux utilisateurs un bon moyen de navigation et d'accès à l'information.

Une association permet de lier deux ou plusieurs Topics, appelés membres de l'association. Chaque membre joue un rôle dans l'association. Les associations peuvent également être classées selon leurs types qui sont eux-même définis en tant que Topics. Cette définition de type d'association, permet de regrouper l'ensemble des Topics qui ont un lien commun avec un autre Topic (un auteur et un ensemble de titres de publications par exemple). Cela améliore considérablement l'efficacité de la navigation.

Le contexte ou Scope

Les Topic Maps définissent également la notion de contexte comme le montre la figure 2.3. Un **contexte** permet de relier les caractéristiques d'un Topic (noms, occurrences et associations) à un contexte particulier permettant de lever certaines ambiguïtés. Une utilisation particulière du contexte est de préciser le format d'une donnée comme une date, ou bien la langue.

Les facettes

Non implémentées dans la norme XTM 1.0, les **facettes** permettent de compléter les informations à propos d'une Occurrence en ajoutant des informations de type attributs-valeurs comme par exemple la date, le type ou la langue des ressources (figure 2.4). Notons que cette information n'est pas ajoutée dans le document lui-même mais dans le composant Occurrence qui référence ensuite la ressource Web ce qui permet de rester indépendant de la ressource elle-même [Caussanel et al. 2002].

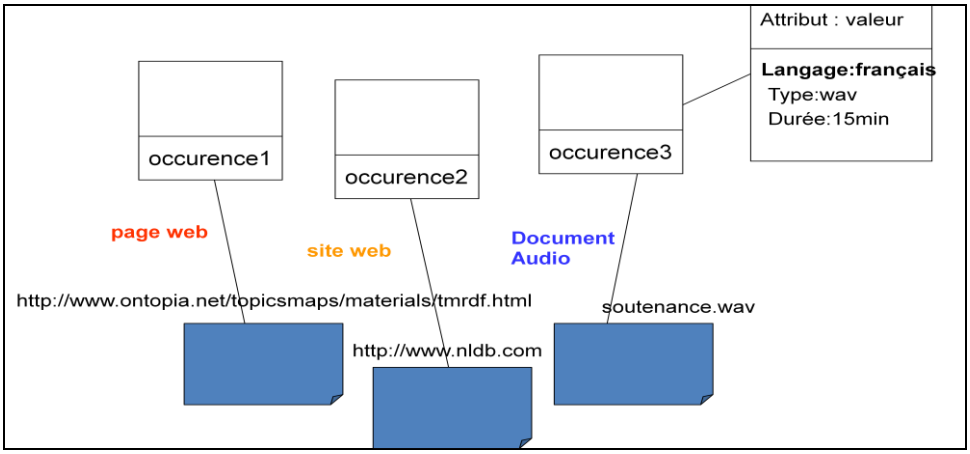


Figure 2.4 Notion de facette dans le modèle des Topic Maps

Le modèle Topic Map dispose d'un langage de spécification XTM (*XML Topic Map*)[10] qui repose sur la syntaxe XML défini par le consortium (*Topic Maps.Org)* et ce dans le but d'adapter le paradigme des Topic Maps au Web.

2.1.2.7 RDFS Vs Topic Maps, Ontologie Vs Topic Maps

Topic Map vs RDFS

Depuis plusieurs années des rapprochements ont été opérés au sein du Web sémantique entre les groupes de recherche investis dans l'utilisation des Topic Maps et ceux de RDFS. Pepper [Pepper, 2002a] montre que les deux modèles sont suffisamment similaires pour définir un mapping entre eux et dans les deux directions. Les principaux points de différence [Caussanel et al. 2002] entre ces deux standards sont : Les annotations, la prise en compte du contexte, les attributs et les occurrences.

Les annotations

Les Topic Maps ont une vision centrée sur les Topics alors que RDFS est plus centré sur les ressources. Les Topic Maps utilisent les Topics pour modéliser un réseau sémantique (niveau sémantique) au-dessus des ressources d'information (niveau ressources). Elles peuvent ne pas faire référence aux ressources. RDFS annote les ressources directement. Le principe de RDFS consiste à inscrire les annotations dans la ressource elle-même. Les Topic Map exploitent quant à elles un ou plusieurs fichiers spécifiques dans lesquels se trouvent les liens vers les ressources. Il s'agit d'une différence fondamentale puisque, dans le premier cas, une recherche suppose de parcourir chaque document pour en consulter le contenu et en évaluer la pertinence. Au contraire, dans le second cas, il s'agit de chercher le nœud du réseau dont la sémantique est proche de celle du sujet recherché donnant alors accès à l'ensemble des documents qui référencent ce sujet.

Concernant son usage proprement dit, la Topic Map est quant à elle davantage **orientée vers la navigation** au sein même de la représentation comme cela peut se faire au sein d'un site ou d'un document par le biais de liens hypertextes.

[10] XML Topic Maps (XTM) 1.0: www.TopicMaps.org Specification, 3 March (2001), http://www.topicmaps.org/xtm/1.0/

La prise en compte du contexte

Les Topic Map permettent d'affecter un Scope à chacune des caractéristiques d'un Topic. Cela permet de montrer les divers sens que peut recouvrir un Sujet en fonction du thème - lui aussi un Topic - dans lequel on se situe. En particulier, cette caractéristique permet de donner plusieurs noms à un Topic à la condition qu'ils appartiennent à des Scopes différents. Cette notion de contexte est absente du modèle RDFS. En ce qui concerne RDFS un mécanisme similaire à celui mis en place pour les Topic Maps est envisageable moyennant un développement spécifique.

Les attributs

Noms : RDFS comme les Topic Maps utilise la propriété *Noms* pour spécifier le nom d'un objet ou d'un concept. Cependant, uniquement en Topic Maps, il est possible de définir de manière standardisée le nom des Topics. En effet, la norme XTM permet de définir une infinité de types de nom mais la recommandation ISO indique trois principaux noms : le nom de base (obligatoire), le nom dédié à l'affichage, le nom utilisé pour des besoins de tris ou de classements.

Les occurrences

L'une des raisons de la puissance des Topic Maps réside dans la séparation en deux niveaux des Topics et de leurs occurrences. En effet, les Topic Maps sont superposées aux ressources et permettent de les décrire sans les modifier. Il est ainsi possible de créer plusieurs Topic Maps pour un même ensemble de ressources. Cette séparation n'existe pas en RDF.
D'après [Garshol, 2003], les deux standards RDF et Topic Maps sont compatibles et ont beaucoup de similarités, nous pouvons trouver des solutions pour convertir une Topic Map au format RDF et vice versa [Pepper, 2002a].

Topic Map Vs Ontologie

Une ontologie est une **conceptualisation formelle du réel,** partagée par une communauté à des fins d'échange. Techniquement, elle doit être exploitable par un programme. Elle est composée : d'une hiérarchie de concepts (lien « is a »), et d'autres liens sémantique (« part of », « localisé sur », « peut précéder », « utilisé par les mêmes utilisateurs », « synonyme », etc.). Les ontologies permettent de donner de la sémantique aux mots.

Basiquement, une Topic Map n'est pas la conceptualisation du réel, c'est une conceptualisation des sujets traités dans un ensemble de documents (autrefois index). Par conséquent, la notion de concept dans les ontologies est très différente de celle du modèle des Topic Maps : dans une Topic Map tout peut être représenté par un Topic (un concept, un thème, un sujet d'intérêt, une personne donnée, etc.).

La Topic Map peut donc inclure une ontologie structurant les Topics de la Topic Map correspondants à des concepts.

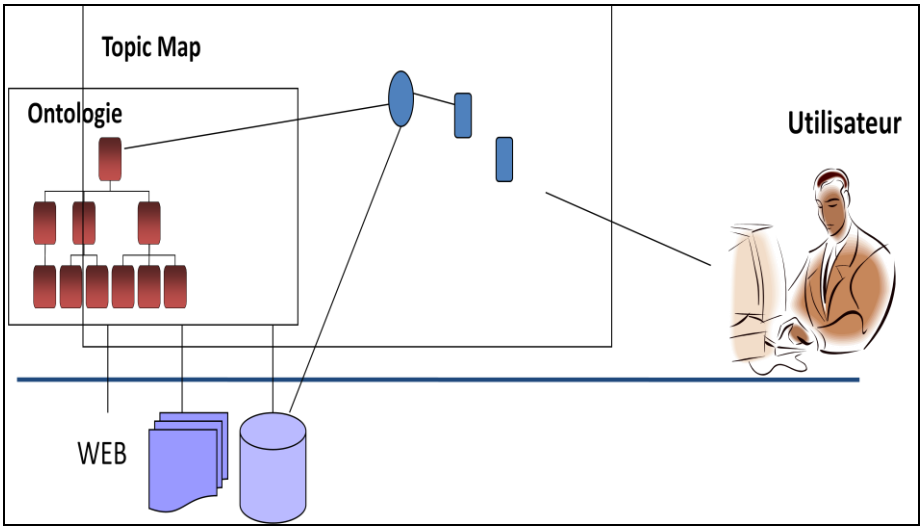


Figure 2.5 Topic Map Vs Ontologie

Par ailleurs, grâce au réseau de liens sémantiques entre les sujets qu'elles représentent, les Topic Maps sont orientées navigation et destinées à l'utilisateur final, les ontologies étant généralement destinées à des applications informatiques. Elles permettent une navigation facile et sélective améliorant ainsi la recherche de l'information dans des contenus de différentes sources.

2.2 Recherche d'information multilingue

Un des problèmes majeurs d'accès à l'information sur le Web est actuellement le multilinguisme, en effet, avec le développement d'internet au niveau mondial, les échanges de documents s'intensifient entre les pays, les cultures et par conséquent les corpus contiennent de plus en plus des documents écrits dans différentes langues. La recherche devient alors multilingue et doit retrouver tous les documents concernés par un besoin d'information, ces documents pouvant être écrits dans des langues différentes.

La notion de multilinguisme en RI peut se présenter sous différentes facettes [Oard et Dorr, 1996]. La facette à laquelle nous nous intéressons dans ce travail est la recherche

d'information par croisement de langues CLIR (*Cross-Language Information Retrieval)* [Grefenstette, 1998], [Greffenstette et al. 2005]. La principale problématique liée à la recherche d'information par croisement de langues est : **Comment à partir d'une requête exprimée dans une langue donnée, récupérer des documents écrits dans des langues différentes de celle de la requête ?**

La plupart des solutions proposées aujourd'hui ont adopté la traduction des documents et/ou des requêtes comme moyen pour mettre ces documents et ces requêtes dans un même référentiel (ou espace d'indexation) [Baziz, 2005]. Ceci revient soit à :

- **traduire la requête vers la langue des documents :** Il s'agit de présenter au moteur de recherche les traductions de cette requête dans les différentes langues souhaitées. Le système récupérera alors les différents documents correspondants à chaque traduction ;
- **traduire les documents vers la langue de la requête :** Les documents sont traduits dans la langue de la requête à l'aide d'outils de traduction. Le système de recherche d'information procède ensuite à une simple interrogation monolingue. Son principal inconvénient est lié à la taille de la base. Il n'est pas concevable de traduire une collection de documents dans toutes les langues souhaitées pour l'interrogation ;
- **traduire la requête et les documents dans un référentiel commun :** Dans ce cas, il s'agit de représenter la requête et les documents dans un même référentiel [Roussey et al. 2001]. Ce référentiel est souvent un vocabulaire multilingue prédéfini qui peut être par exemple un thésaurus (exemple EuroWordNet[11]). Cependant l'inconvénient de ce type de vocabulaire est qu'il n'est pas toujours disponible. Cette solution permet de résoudre partiellement le problème posé par la traduction. En effet, dans un corpus où les documents sont écrits dans n langues, il nous faut au pire des cas n-1 traducteurs. Cette méthode est la moins consolidée de toutes celles que nous avons citées et peu de résultats satisfaisants ont été constatés [Hahn et al. 2004].

Actuellement la plupart des travaux dans ce domaine se focalisent sur la traduction de la requête [Baziz et al. 2003]. Cette traduction est moins coûteuse que celle de tous les documents du corpus [Ballestros et Croft, 1998], [Gollins et Sanderson, 2000] et [Schauble et Braschler, 2000].

[11] http://www.illc.uva.nl/EuroWordNet/

2.2.1 Problèmes liés à la recherche d'information multilingue

Dans le contexte de la traduction de requêtes, de documents ou des deux, les systèmes tentent de résoudre plusieurs problèmes parmi lesquels, nous citons : l'absence de termes sémantiquement équivalents d'une langue à une autre, le fait qu'un terme possède plusieurs traductions et les problèmes d'ambigüité sémantique.

Absence de termes sémantiquement équivalents d'une langue à une autre

Il s'agit de retrouver pour chaque terme de la requête exprimé dans la langue source (L1), un ou plusieurs terme(s) sensé(s) le représenter dans la langue cible (L2). Or, une des spécificités du multilinguisme est l'absence éventuelle de termes sémantiquement équivalents d'une langue à une autre ; ce qui est assez fréquent lorsque les contenus sont issus de différentes cultures.

Problème de désambiguïsation

Le second problème est posé dans le cas où un terme possède plusieurs traductions et est lié au choix de la ou des meilleure(s) traduction(s), c'est le problème de la désambiguïsation. Des travaux proposent des techniques de désambiguïsation et d'expansion des termes de traductions de requêtes. Cette désambiguïsation s'appuie sur des concepts issus d'une base de données lexicographique externe (par exemple WordNet [Fellbaum, 1998]).

Problèmes d'ambigüité sémantique

Du fait de l'utilisation d'un processus de traduction automatique (pour la requête ou pour les documents ou pour les deux), la recherche d'information multilingue hérite les problèmes posés par la traduction automatique; les problèmes de la traduction automatique sont dus aux ambiguïtés sémantiques des termes. Nous distinguons trois types d'ambiguïtés sémantiques [Roussey et al. 2001] et nous illustrons ces ambiguïtés par des exemples tirés d'un corpus français anglais :

- La polysémie : Un même terme peut avoir différents sens. Le terme anglais « *plant* » illustre bien ce problème car il possède au moins trois sens différents (la plante, l'installation, le coup monté). Par contre, « power plant » n'a qu'un seul sens. Donc, si nous tenons compte des termes voisins aux termes ambigus, c'est-à-dire son contexte, nous arrivons à déterminer son sens exact ;

- L'homographie : Deux mots différents s'écrivent de la même façon, par exemple « livre » est soit la conjugaison du verbe « livrer », soit le nom synonyme d'ouvrage. Le mot « bois » peut être la conjugaison du verbe boire, soit il veut dire le matériau tiré de l'arbre utilisé comme combustible ou pour fabriquer du papier, des meubles ou des objets. La catégorie lexicale (verbe, nom, adjectif) permet de lever cette ambiguïté de sens ;
- Le sens large : Un terme qui a un sens très large, exemple « air » peut prendre un sens particulier dans certains domaines « air bag ». Pour résoudre ce problème, il faut identifier la totalité de l'expression pour déterminer le concept spécifique qui se cache derrière la combinaison des termes.

Nous distinguons trois classes de méthodes de traduction proposées dans la littérature [Harrathi, 2009] : les méthodes basées sur l'utilisation de traducteur automatique, les approches qui utilisent des corpus alignés (comparable ou parallèles) et les méthodes basées sur l'utilisation de dictionnaires bilingues.

2.2.2 Utilisation de traducteur automatique

Cette approche, appelée en anglais *Machine Translation method* ou *MT-based method*, s'appuie sur un logiciel de traduction automatique pour la traduction de la requête ou des documents [Radwan, 1994], [Pirkola, 1998]. Les systèmes de recherche basés sur les traducteurs automatiques sont utilisés pour obtenir un même texte dans plusieurs langues, avec ou sans l'aide d'un expert. Il s'avère que la traduction de la requête présente moins de précision que celle de la collection de documents [Oard et Hackett, 1997], qui contiennent un contexte d'information nettement plus important, ce qui diminue les risques de mauvaise traduction. Cependant, l'application de cette méthode consistant à traduire tous les documents dans toutes les langues désirées est trop compliquée à mettre en œuvre pour des corpus de taille importante. C'est donc la traduction de la requête qui est retenue lorsqu'on parle de recherche d'information par traduction automatique.

Cette méthode est en cours de consolidation et les résultats qu'elle donne ne sont pas très satisfaisants, car elle se base sur la traduction automatique dont les résultats ne sont pas toujours probants, surtout lorsque les requêtes sont courtes. En effet, plus les requêtes sont courtes, moins les résultats sont bons.

2.2.3 Utilisation de dictionnaire bilingue

Cette méthode, appelée en anglais *machine-readable dictionaries-based method,* se base sur l'expansion de la requête. Cela consiste à formuler la requête autrement en remplaçant les mots qui la composent par des variantes afin de récupérer des documents pertinents dans lesquels les termes saisis ne sont pas toujours présents. Cette reformulation se fait à l'aide de dictionnaires monolingues qui permettent la reformulation dans une même langue (synonymes, antonymes, etc.) et les dictionnaires bilingues qui permettent la reformulation dans des langues différentes [Aljlayl et Frieder, 2001].

2.2.4 Utilisation de corpus alignés (parallèles ou comparables)

Cette méthode basée sur le corpus [Schauble et Braschler, 2000], [Nassr et Boughanem, 2002], [Névéol et Ozdowska, 2005] utilise directement le contenu d'un ensemble de documents, regroupés dans un corpus soit pour la traduction ou pour la désambiguïsation des requêtes. Un corpus aligné est constitué d'un ensemble de documents exprimés dans une langue, alignés avec des documents dans une autre langue. L'alignement entre ces documents consiste à mettre en correspondance les documents de langues différentes selon un critère donné. Il peut être parallèle ou comparable.

L'alignement parallèle consiste à mettre en correspondance chaque document d'une langue source L1 avec le document représentant sa traduction dans la langue cible L2. Dans ce cas l'alignement peut être fait sur : le document, les paragraphes, les phrases ou les termes. Les corpus basés sur ce type d'alignement sont appelés les corpus parallèles [Névéol et Ozdowska, 2005].

L'alignement comparable plus délicat à réaliser [Sheridan et Ballerini, 1996], revient à mettre en correspondance des documents en se basant sur des critères comme par exemple la présence de mêmes dates, de mêmes noms de personnes dans des documents de langues différentes [Greffenstette, 1998], [Oard et Dorr, 1996]. Les corpus basés sur ce type d'alignement sont appelés les corpus comparables.

2.2.5 Quelques travaux sur la recherche d'information multilingue

Il existe plusieurs travaux dans le contexte de la recherche d'information multilingues. Nous citons par exemple les travaux suivants : Le projet UNL, le système Sydom et le projet MKBEEM qui nous ont paru particulièrement intéressants.

Le projet UNL[12] (*Universal Networking Language*)

Ce projet [Uchida, 2004] est lancé en 1996 par L'Université des Nations Unies à Tokyo, il a pour objectif de permettre un accès au Web dans toutes les langues du monde. La figure 2.6 illustre le principe de base de la traduction par UNL.

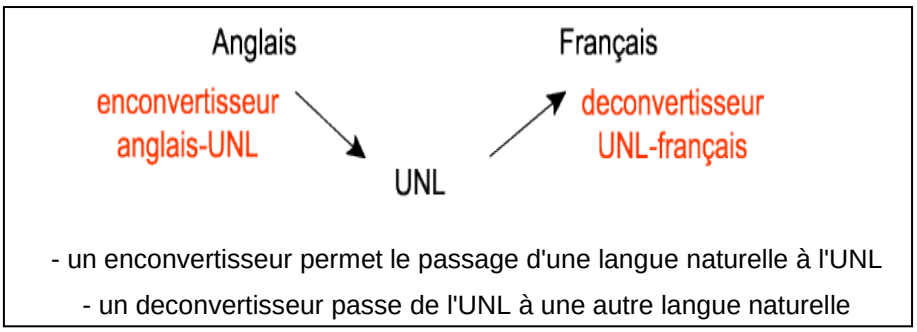


Figure 2.6 Principe de base de la traduction par UNL

Le modèle développé dans le projet UNL est fondé sur une représentation interlangue qui exprime sous forme de **graphes sémantiques** la structure abstraite des énoncés linguistiques. Le formalisme UNL permet de produire des textes dans la langue de son choix. Il permet de représenter n'importe quel texte sous forme d'un graphe conceptuel qui utilise des concepts universaux (*Universal Words*) et des relations entre ces concepts. Il comporte des nœuds étiquetés par les éléments d'une ontologie de concepts, des relations sémantiques binaires entre nœuds, ainsi qu'un ensemble d' « attributs » représentant des aspects de la sémantique des langues naturelles, et qui peuvent s'appliquer aux nœuds. La fondation UNDL (*United Nations Depository Libraries*) se doit d'être plus active auprès des instances de standardisation IETF, ITU, W3C,... L'intérêt particulier de ce formalisme de graphes parmi d'autres est que, développé en parallèle par de multiples laboratoires dans divers pays, il dispose d'outils (analyseurs de textes, générateurs automatiques de langues) dans une dizaine de langues différentes (12 actuellement). Cette situation en fait un candidat privilégié pour jouer le rôle de langage pivot dans la représentation d'un contenu destiné à être visible, après conversion, dans des applications multilingues.

Le langage UNL peut aussi être utilisé comme langage d'indexation dans des applications de recherche d'information sur de grandes bases (ou grands corpus) de documents. Cet aspect indexation apparaît comme fondamental. En effet, les moteurs de recherche existants pratiquent une indexation par mots-clés, éventuellement enrichie d'une information morphosyntaxique (lemmatiseur), et, pour les plus avancés d'entre eux, d'une information sémantique (réseau sémantique). La recherche par mots-clés, même enrichie, se

[12] http://clips.imag.fr/projets/unl/

heurte malgré tout aux deux obstacles qui sont d'une part l'ambiguïté des termes (ou la polysémie) et, d'autre part, l'absence de lien sémantique entre les termes de recherche qui restent des mots-clés isolés. L'expression de requêtes dans un langage de graphes offre une solution à ces deux problèmes : pour le premier, la distinction des homonymes par l'identification du mot à une entrée lexicale unique ; pour le second, l'existence de relations sémantiques entre les mots, qui permet de spécifier les rôles. L'utilisation de tels formalismes de représentation sémantique offre en outre des possibilités d'extension de l'indexation à des contextes multilingues.

Le système SyDoM

Le système de recherche d'information multilingue SyDoM s'appuie sur **un thésaurus sémantique** qui représente un nouveau genre d'ontologie défini dans [Roussey et al. 2002] [Pivano et al. 2004]. SyDoM se compose de différents modules, chacun de ces modules est dédié à une étape des processus d'indexation et de recherche des documents XML [Roussey et al. 2001]. SyDoM comprend :

- un module de gestion des thésaurus sémantiques, permettant de construire un langage documentaire utilisé pour annoter et interroger les documents XML. Ce langage se compose d'une modélisation du domaine à laquelle sont associés plusieurs vocabulaires ;
- un module d'indexation manuelle de documents en XML, permettant d'annoter les documents par des graphes conceptuels [Sowa, 1984] ;
- un module de recherche, permettant de construire une requête sous forme de graphes conceptuels et de récupérer la liste des documents répondant à cette requête.

Le thésaurus sémantique allie à une modélisation du domaine plusieurs terminologies. Ainsi les termes sont dissociés des notions qu'ils dénotent, ce qui permet de clarifier les relations entre les termes et les notions et d'identifier les relations terminologiques des relations sémantiques. Un thésaurus sémantique définit deux niveaux de connaissances (figure 2.7 et 2.8) :

- Le niveau conceptuel qui modélise le domaine d'étude formé de types de concepts ou de relations. Dans ce cas, il s'agit d'une conceptualisation du domaine résultant d'un consensus entre les différents acteurs d'un domaine particulier. Cette conceptualisation est utilisée comme langage pivot dans

SyDoM, elle est équivalente au support du modèle des graphes conceptuels de Sowa [Sowa, 1984] ;

- Le niveau terminologique est composé de l'ensemble des termes, le terme étant défini comme la manifestation linguistique d'un concept repéré dans un texte.

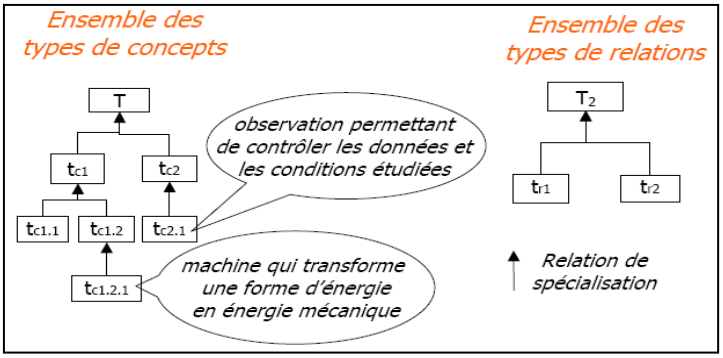


Figure 2.7 Thésaurus sémantique : Niveau conceptuel

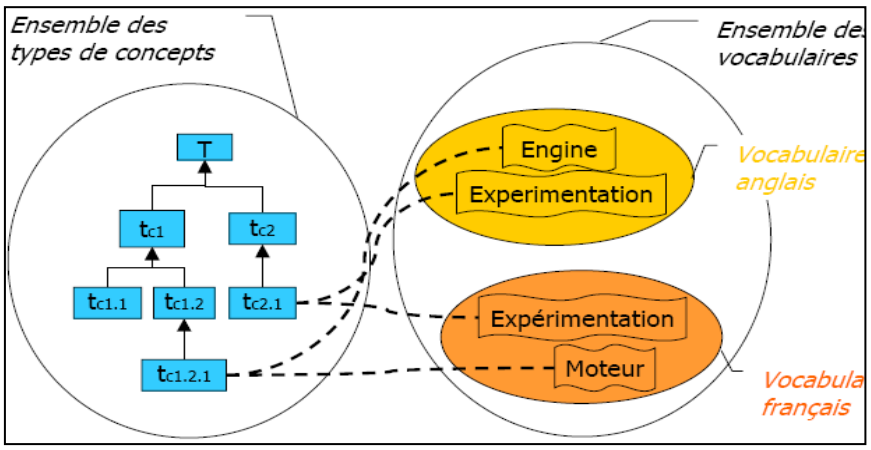


Figure 2.8 Thésaurus sémantique : Niveau terminologique

Dans la figure 2.9, V se compose de deux vocabulaires, un vocabulaire anglais V_{eng} et un vocabulaire français V_{fr}. Pour chacun des vocabulaires V_i, il existe une fonction λ_{Vi} qui associe à chaque type au moins un terme appartenant à V_i. Dans l'exemple de la figure 2.9, la fonction λ_{CVeng} fait correspondre au type $t_{c1.1.1} \in TC$, le terme anglais « Fuel » $\lambda_{CVeng}(t_{c1.1.1})$= « Fuel » et la fonction λ_{CVfr} fait correspondre au même type tc1.1.1, le terme français « Carburant » $\lambda_{CVfr}(t_{c1.1.1})$= « Carburant ».

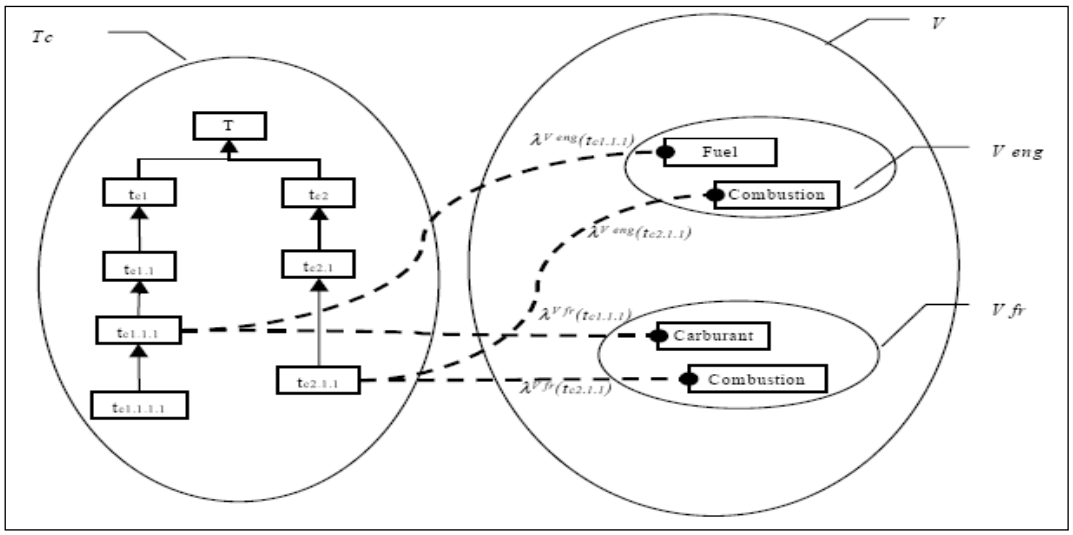


Figure 2.9 Un exemple de thésaurus sémantique [Harrathi, 2005]

A partir de ce thésaurus sémantique définissant le vocabulaire et les connaissances du domaine, les graphes sémantiques sont utilisés pour indexer les documents. Ces graphes peuvent être comparés aux graphes conceptuels de Sowa. Un exemple de graphe sémantique est présenté dans la figure 2.9 (dans le formalisme des graphes conceptuels). Ce graphe peut représenter l'index d'un document qui traite de la combustion du diesel.

Le projet MKBEEM

C'est une plate-forme de commerce électronique multilingue et pluriculturelle qui offre des interfaces homme-machine en langue naturelle respectueuse des langues et des cultures de ses clients internationaux. Le projet MKBEEM [Heinecke, 2003] (*Multilingual Knowledge-Based European Electronic Marketplace*) offre des services de médiation de commerce électronique multilingue (Finnois, Français, Espagnol et Anglais). L'approche technique est basée sur le couplage de raisonnements sémantiques (ontologies) avec le traitement automatique des langues naturelles. Cette technologie permet d'établir la correspondance entre une requête d'utilisateur en langue naturelle et la représentation ontologique du contenu de cette requête. Les ontologies sont exploitées pour la classification et l'indexation des produits ou des services dans les catalogues aussi bien que pour faciliter l'interprétation et l'inférence d'information pertinente en rapport avec la requête de l'utilisateur. En conclusion, ce projet se base sur trois scénarios : catalogage multilingue, traitement des requêtes d'utilisateur et enfin la transaction multilingue.

La section suivante dresse un état de l'art sur les approches de construction de Topic Maps, elle propose plus précisément une classification de ces approches selon les sources et

les techniques utilisées, elle présente aussi les différents domaines de recherche reliés à la construction de Topic Maps en particulier, les techniques d'analyse linguistiques de documents textuels, les méthodes de création et d'enrichissement d'ontologies et les approches de fusion de schémas conceptuels et d'ontologies. Nous donnons ensuite un bref aperçu sur les outils existants pour l'édition, la navigation et la visualisation de Topic Maps, enfin nous terminons par une étude comparative de ces approches selon des critères de comparaison que nous définissons, cette étude nous a permis de dégager les limites des travaux existants et énoncer notre proposition pour la construction de Topic Map à partir d'un contenu textuel multilingue.

2.3 Etat de l'art sur les approches de construction de Topic Maps

2.3.1 Introduction

Il existe dans la littérature plusieurs travaux sur la construction de Topic Maps [Ellouze et al. 2008a]. Ils se distinguent principalement par **les sources en entrée** et la nature des **techniques utilisées** pour la production de Topic Maps. Les sources peuvent être des documents XML, des métadonnées RDF, des bases de connaissances (thésaurus, ontologies, etc), des documents textuels ou des connaissances des experts du domaine.

En se basant sur l'état de l'art, nous remarquons que la construction de Topic Maps fait intervenir plusieurs domaines de recherche et les techniques utilisées par les approches proposées sont liées à ces domaines [Ellouze et al. 2008b]. Ces techniques sont variées et dépendent aussi des sources en entrée, en effet, elles peuvent englober des techniques de conception de schémas de données, des techniques de fusion de données, des techniques d'apprentissage ou de classification, dans le cas où les sources en entrée sont des documents textuels, certaines approches utilisent des techniques d'analyse linguistique de documents textuels, d'autres travaux s'inspirent de méthodologies de construction d'ontologies, enfin, notons que parmi les approches existantes, il y a celles qui se basent sur une construction collaborative de Topic Maps faisant intervenir plusieurs acteurs. La figure 2.10 récapitule l'état de l'art sur les approches de construction de Topic Maps [Ellouze et al. 2008b].

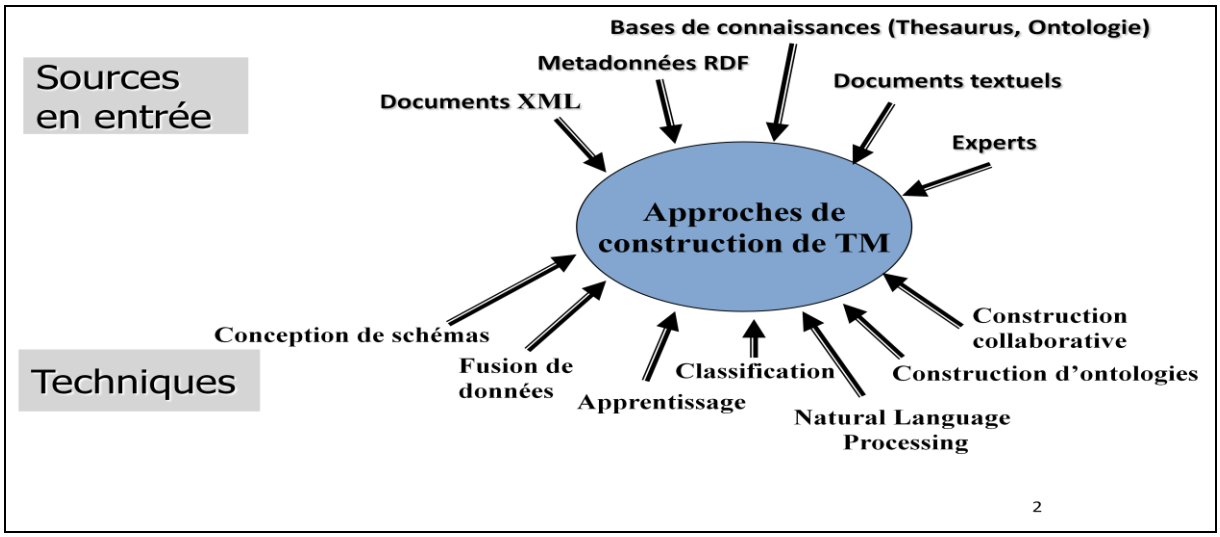


Figure 2.10 Etat de l'art sur les approches de construction de Topic Maps

2.3.2 Extraction de concepts et de relations à partir de documents textuels

Dans la littérature, plusieurs méthodes ont été proposées pour l'extraction de concepts et de relations entre concepts à partir de documents textes. Parmi ces méthodes, il y a celles qui ont été développées dans le cadre de la création et de l'enrichissement d'ontologies à partir de données textuelles. Ces méthodes se distinguent par le type de technique utilisée : technique statistique, techniques syntaxique ou encore technique de fouille de données. La plupart d'entre elles sont outillées.

2.3.2.1 Méthodes statistiques

Dans les méthodes basées sur les techniques statistiques, des mesures sont utilisées afin de sélectionner les concepts candidats. Parmi elles, on peut citer le nombre d'apparitions d'un terme au sein d'un corpus [Agirre et al. 2000] [Faatz et Steinmetz, 2002] [Parekh et al. 2004], l'information mutuelle, le tf×idf, le T-test ou encore les lois de distributions statistiques des termes [Velardi et al. 2001] [Xu et al. 2002] [Neshatian et Hejazi, 2004]. Notons que toutes ces techniques ne permettent pas d'extraire des relations entre concepts.

2.3.2.2 Méthodes syntaxiques

Les méthodes basées sur l'analyse syntaxique, quant à elles, utilisent les fonctions grammaticales d'un mot ou d'un groupe de mots au sein d'une phrase. Certaines d'entre elles posent l'hypothèse que les dépendances grammaticales reflètent des dépendances sémantiques [Bendaoud et al. 2007] [Roux et al. 2000]. D'autres utilisent des patrons syntaxiques [Hearst, 1992] [Maedche et Staab, 2000] [Stumme et al. 2006]. Ces méthodes ont l'avantage d'extraire

aussi bien des concepts mais aussi les relations entre concepts. Toutefois, ces relations ne sont pas toujours étiquetées sémantiquement. Une intervention humaine est donc nécessaire pour nommer ces relations.

2.3.2.3 Méthodes fondées sur des techniques de fouille de données

La dernière catégorie de méthodes exploite des techniques de fouille de données. A titre d'exemple, Han [Han et Karypis, 2000] et Neshatian [Neshatian et Hejazi, 2004] exploitent une ontologie et utilisent une technique de classification permettant de rapprocher des concepts candidats (contenus dans des documents) de concepts présents dans l'ontologie. Le principe est similaire dans les approches d'Agirre [Agirre et al. 2000] et Parekh [Parekh et al. 2004] qui regroupent, par une technique de clustering, des termes en fonction de leur nombre d'occurrences au sein du corpus. Chaque cluster représente alors la possibilité qu'une relation existe entre les concepts qu'il regroupe. D'autres méthodes proposent d'utiliser les corrélations fréquentes pouvant exister entre les termes d'un corpus. Ces approches consistent à extraire des règles d'association [Agrawal et Srikant, 1997] entre des termes candidats [Bendaoud et al. 2007] [Maedche et Staab, 2000] [Stumme et al. 2006]. A l'issu du processus, un ensemble de règles d'association est dérivé. Chacune des règles décrit l'existence d'une relation entre deux concepts. Un processus d'étiquetage manuel est par la suite exécuté afin de nommer les relations produites.

Il existe plusieurs outils d'extraction de concepts et/ou de relations entre concepts à partir des documents textuels. Parmi ces outils, nous citons Nomino [Dumas et al. 1997], Lexter [Bourigault, 1996], Fastr [Jacquemin et Bourigault, 2003], Mantex [Frath et al. 2000], Likes [Rousselot et al. 1996], Acabit [Daille, 1999], Syntex [Bourigault et al. 2005], OntoGen [Fortuna et al. 2006] et Text2Onto [Cimiano et Volker, 2005]. A titre d'exemple, l'outil Syntex est un analyseur de texte. Il sert à identifier les dépendances syntaxiques entre concepts. Text2Onto est un outil conçu pour construire des ontologies à partir de textes de manière complètement automatique. Il est composé de modules qui extraient à partir des textes des concepts, des relations entre ces concepts (relation d'équivalence, hiérarchiques, etc.) et des instances de concepts. Il repose sur l'architecture GATE[13] [Cunningham et al. 2002] pour prétraiter les textes. Les résultats sont dotés d'une mesure de confiance entre 0 et 1 obtenue à l'aide de différentes mesures combinables (*tf*×*idf*, RTF, entropie).

[13] http://gate.ac.uk/

Enfin, plusieurs **architectures d'ingénierie du texte** ont été développées pour les traitements linguistiques. Nous citons par exemple, l'architecture GATE (*General Architecture for Text Engineering*), UIMA [Ferrucci et Lally, 2004] ou Textpresso [Muller et al. 2004]. Ces architectures visent généralement l'annotation linguistique et l'exploration de corpus de taille moyenne pour l'extraction d'information. GATE a l'avantage de proposer une solution générique pour le traitement linguistique des documents textuels à travers un ensemble de modules paramétrables.

Si les plates-formes GATE et UIMA sont plutôt conçues comme des solutions génériques, le système Textpresso [Muller et al. 2004] a pour objectif de proposer une architecture capable de traiter des corpus de documents issus d'un domaine spécialisé. Cette plate-forme a été conçue pour la fouille des documents traitant de biologie, aussi bien des résumés que des articles complets. Son évaluation a porté sur un corpus relativement petit : 16000 résumés et 3000 articles en texte brut.

En règle générale, on dispose de très peu d'informations pour d'évaluer les performances de ces systèmes sur un corpus de documents. Un premier test nous a montré que GATE ne convient pas au traitement de gros corpus de documents : seuls de petits volumes de documents pouvaient être traités sans rencontrer des problèmes. Ceci s'explique par le fait que GATE ait été conçue comme un environnement puissant de développement et de conception d'applications de TAL dans le cadre de l'extraction d'information. Le passage à l'échelle n'était pas un objectif central. La plate-forme KIM [Popov et al. 2004], qui s'appuie sur GATE, tente cependant de satisfaire cette contrainte dans le cadre de projets d'annotation sémantique massive. Cette architecture est dédiée à l'enrichissement d'ontologies, l'indexation sémantique et la recherche d'information.

2.3.3 Méthodes de construction d'ontologies

De nombreux travaux ont été réalisés pour construire et maintenir des ontologies. Une ontologie, est une conceptualisation formelle du réel partagée par une communauté à des fins d'échange. Techniquement, elle est représentée sous la forme d'une hiérarchie « is a » de concepts. De nombreux autres liens sémantiques peuvent être stockés dans l'ontologie en fonction du domaine à conceptualiser, par exemple : « partie de », « est localisé sur », « a un effet sur », « évoque », etc. Il existe dans la littérature plusieurs méthodes de construction d'ontologies [Gomez-Perez et Macho, 2003] :

a) **les méthodes de construction par apprentissage** à savoir construction à partir de dictionnaires, à partir de documents XML, à partir de documents texte, à partir de bases de connaissances, à partir de schémas semi-structurés, à partir de schémas relationnels, etc., [Aussenac-Gilles et al. 2000], [Agirre et al. 2000], [Nobécourt, 2000], [Kietz et al. 2000], [Maedche et Staab, 2001], [Suryanto et Compton, 2001], [Dietel et al. 2001], [Maedche et Staab, 2002], [Khan et Luo, 2002], [Veraldi et al. 2002], [Gomez-Perez et Macho, 2003], [Meadche et al. 2003], [Maedche et Staab, 2004], [Hernandez, 2005], etc ;
b) **les méthodes de construction par réutilisation** c'est à dire par fusion d'ontologies existantes de même thématique ou encore par intégration d'ontologies de thématiques différentes [Gangemi et al. 1999], [Noy et Musen, 2000], [McGuiness et al. 2000] ;
c) **les méthodes de construction par re-engineering d'une ontologie existante** [Gomez-Perez et Rojas-Amaya, 1999] ;
d) **les méthodes de construction from scratch** [Gruninger et Fox. 1995], [Fernandez et al. 1999], [Schnurr et al. 2000], [Davies et al. 2002], [Sure et al. 2003], etc.

De nombreux outils sont développés pour la construction d'ontologies, nous citons par exemple Protege-2000[14], OntoStudio[15] (successeur de OntoEdit), c'est un environnement pour la construction et la gestion d'ontologies, CORPORUM-OntoExtract[16], Text-To-Onto, OntoBuilder[17] ,c'est un outil pour la création d'ontologies à partir de sources semi-structurées, iPROMPT [Noy et Musen, 2003] c'est un outil interactif pour la fusion d'ontologies, OntoLingua[18], CHIMAERA[19], c'est un outil pour la création et la gestion d'ontologies distribuées sur le Web, une de ses principales fonctions est la fusion d'ontologies. OntoSaurus[20] est un éditeur pour l'ontologie SENSUS. Tous ces outils disposent de la fonction d'édition, certains offrent aussi des fonctionnalités de fusion et/ou de validation et/ou de raisonnement.

[14] http://protege.stanford.edu/
[15] http://www.ontoprise.de
[16] http://www.Ontoknowledge.org/del/shtml
[17] http://ie.technion.ac.il/OntoBuilder
[18] http://www-ksl.stanford.edu/software/chimaera/
[19] http://ontolingua.stanford.edu/
[20] http://www.isi.edu/isd/ontosaurus.html

Parmi les méthodologies proposées pour la construction d'ontologies, nous citons par exemple : METHONTOLOGY [Fernandez et al. 1997] et TERMINAE [Aussenac et al. 2000].

METHONTOLOGY [Fernandez et al. 1997] qui s'applique à clarifier les différentes étapes de la construction en respectant des activités de gestion de projets (planification, assurance qualité), de développement (spécification, conceptualisation, formalisation, implémentation, maintenance) et des activités de support (intégration, évaluation, documentation).

TERMINAE, la méthodologie de Aussenac et al. [Aussenac et al. 2000] qui propose une approche pour sélectionner les concepts, leurs propriétés, les relations et leur regroupement. Cette méthodologie a été est une composante de la plate forme RFIEC (RFIEC est une plateforme regroupant un ensemble de résultats associé aux compétences locales d'équipes travaillant autour de l'analyse et la représentation de textes (outils, méthodologies, corpus, ressources linguistiques). Elle repose sur l'utilisation d'outils de traitement automatique des langues analysant les termes de textes et les relations lexicales. Les termes sont regroupés suivant leur contexte et facilitent la création de concepts et de relations sémantiques. Les concepts et relations sont ensuite formalisés dans un modèle.

La méthode de Hernandez **[Hernandez et Mothe, 2006]**, qui consiste à prolonger la méthode TERMINAE en intégrant les ressources terminologiques qui sont les thésaurus. Ils proposent de transformer un thésaurus en une ontologie de domaine. Les auteurs définissent une méthode qui permet d'extraire les éléments du schéma conceptuel de l'ontologie à partir d'un thésaurus et de documents textuels. Leur approche est fondée sur un ensemble de règles de transformation. Ces règles exploitent les liens «est plus spécifique que», «est plus générique que», « Utiliser plutôt » (UP) et « Utiliser pour désigner » (UPD) d'un thésaurus pour générer les concepts de l'ontologie, les labels associés à chacun de ces concepts et la hiérarchie de concepts.

La méthode proposée repose sur trois étapes : La **première étape** vise à extraire du thésaurus un ensemble de concepts ainsi que leurs variations lexicales. Par l'utilisation d'un thésaurus, Hernandez vise à proposer un mécanisme automatique de regroupement des labels d'un même concept. La **deuxième étape** permet de structurer les concepts de l'ontologie à partir de la détection de relations taxonomiques et associatives dans le thésaurus et dans le corpus. Cette étape soulève différentes problématiques de la construction d'ontologies. L'une d'elles relève de la difficulté à organiser les concepts par des relations taxonomiques. Pour

résoudre ces problèmes, les relations hiérarchiques entre termes du thésaurus sont utilisées pour aider à la détection de ces relations. Cependant, un des inconvénients des thésaurus est que le niveau hiérarchique le plus général est souvent composé de nombreux termes. Afin d'organiser les concepts à partir d'un niveau d'abstraction comportant un nombre limité de concepts, l'auteur propose l'utilisation d'une ontologie générique. Cette ontologie est utilisée pour définir semi-automatiquement les types abstraits du domaine et structurer l'ontologie.
Une autre problématique que fait intervenir cette étape est la détection de relations associatives entre concepts et la désignation de ces relations sémantiques. Elle propose un mécanisme visant à proposer de façon semi-automatique ces relations ainsi que leur label. Le mécanisme repose sur l'analyse syntaxique (**la troisième étape**) du corpus de référence qui permet d'extraire les syntagmes constituant le lexique du corpus ainsi que le contexte dans lequel ils apparaissent (noms et verbes qu'ils régissent et par qui ils sont régis). La formalisation de l'ontologie est réalisée à la fin de ces différentes étapes après la validation des éléments proposés par un expert du domaine.
Parmi les outils le plus connus d'élaboration d'ontologies, nous citons : Text-To-Onto et OntoBuilder.

Text-To-Onto, développée à l'Institut AIFB de l'Université de Karlsruhe, est une application d'extraction d'ontologies à partir de corpus ou de documents Web qui permet également la réutilisation d'ontologies existantes [Maedche et Staab, 2001]. Text-To-Onto est intégrée à la plate-forme logicielle KAON qui permet l'édition et la maintenance d'ontologies [Bozsak et al. 2002]. KAON utilise le langage de représentation RDFS et est orientée vers l'utilisation des ontologies sur le Web, l'application KAON Portal permettant la recherche et le parcours d'ontologies via un navigateur Web.

OntoBuilder, développée au Technion d'Haifa, permet de bâtir une ontologie à partir de ressources Web [Roitman et Gal, 2006]. L'extraction de l'ontologie à partir de fichiers XML est suivie d'une phase de raffinage guidée par l'utilisateur. Onto-Builder autorise aussi la fusion d'ontologies extraites de différents sites Web.

2.3.4 Intégration de schémas conceptuels et d'ontologies

Les approches d'intégration d'ontologies sont généralement basées sur deux étapes :

- Etape d'alignement/matching qui consiste à trouver les similarités entre les ontologies à fusionner ;

- Etape de fusion/merging qui consiste à intégrer les deux ontologies sources selon les similarités trouvées et construire une nouvelle ontologie.

Les approches d'intégration existantes se distinguent les unes des autres par les techniques d'alignement et de fusion utilisées. Nous commençons tout d'abord par présenter un état de l'art sur les travaux de matching entre deux schémas/ontologies.

2.3.4.1 Alignement de schémas conceptuels et d'ontologies

Le processus de matching ou d'alignement consiste à trouver des correspondances sémantiques entre les éléments de deux schémas. Il prend en entrée deux schémas/ontologies, représenté chacun comme un ensemble d'éléments (des tables, des éléments XML, des classes, des propriétés, des concepts) et détermine en sortie les relations (par exemple : équivalence, subsomption) entre ces éléments.

De nombreux travaux portent aujourd'hui sur l'alignement de schémas et d'ontologies, des études récentes telles que [Kalfoglou et Schorlemmer, 2003], [Shvaiko et Euzenat, 2005] présentent une synthèse des différentes approches proposées. Ces approches utilisent des techniques variées qui exploitent différents types d'information, les noms des éléments, les types des données, la structure de la représentation des éléments, les caractéristiques des données, etc.

Des travaux de classification et de comparaison des approches d'alignement ont été aussi réalisés ces dernières années [Rahm et Bernstein, 2001], [Shvaiko et Euzenat, 2004], en se basant sur ces travaux, on peut distinguer trois classes d'approches : les approches terminologiques, les approches structurelles et les approches sémantiques.

Notons aussi que la plupart des travaux sur le matching de schémas et d'ontologies combinent l'utilisation de ces trois approches de façon à rendre le processus de génération des mappings le plus efficace possible [Kefi et Reynaud, 2006]. Les techniques terminologiques sont appliquées en priorité. Elles permettent de générer les alignements les plus probables en exploitant les noms des éléments. Ensuite, les techniques structurelles et sémantiques sont appliquées afin de trouver des alignements supplémentaires, lorsque l'exploitation des chaînes de caractères ne suffit pas.

Approches terminologiques

Les approches terminologiques sont généralement appliquées en priorité dans le processus de matching. Elles exploitent les labels et les synonymes des termes désignant les concepts à aligner [Maedche et Staab, 2002], [Euzenat et Valtchev, 2004]. Dans cette classe

d'approche, on peut distinguer deux sous classes : l'approche syntaxique et l'approche lexicale, appelée aussi linguistique. L'approche syntaxique considère les labels des concepts comme des chaînes de caractères et effectue la correspondance entre ces labels à travers les mesures de (dis)-similarité des chaînes de caractères; parmi les algorithmes de base portant sur des chaînes de caractères (*String-based algorithms*), nous citons :

- *EqualDistance Algorithm* : Cet algorithme est le plus simple. Il retourne la valeur 1 si les deux chaînes de caractères sont les mêmes et 0 sinon ;
- *SubString Equivalence* : Cet algorithme est une variation du précédent. Il considère que deux chaînes de caractères sont similaires, quand l'une est une sous-chaîne de l'autre ;
- *Edit Distance Algorithm (Levenshtein Distance)*: En général, une distance d'édition entre deux objets, est le coût minimal des opérations que l'on doit appliquer à l'un des objets afin d'obtenir le deuxième objet. Pour les chaînes de caractères, cette distance s'appelle *Levenshtein Distance*, et c'est le nombre minimum d'insertions, délétions, et substitutions de caractères requis, pour transformer une chaîne de caractères en une autre. Par exemple : EditDistance (NKN, Nikon) =0.4.

Les approches lexicales ou linguistiques exploitent la sémantique des labels des concepts et effectue la correspondance à travers les relations lexicales (par exemple, synonymie, hyponymie, etc.). Pour cela, ces approches font appel à des ressources auxiliaires tels que les dictionnaires de synonymes et d'hyponymes comme WordNet [Fellbaum, 1998], mais aussi à des thésaurus ou des ressources sémantiques spécifiques aux domaines étudiés.
Les techniques terminologiques s'avèrent efficaces, cependant, elles ne permettent pas de trouver l'ensemble des rapprochements possibles. Certains travaux proposent donc de compléter les approches basées sur des techniques terminologiques par d'autres techniques basées sur l'exploitation de la structure.

Approches structurelles

Les approches structurelles exploitent la structure des schémas ou des ontologies à aligner souvent représentés sous forme de graphes. Dans ce cas, les concepts ne sont pas étudiés séparément, ils sont considérés en tenant compte de leur position dans la hiérarchie des concepts. Les algorithmes implémentant ces techniques sont basés sur des heuristiques

qui considèrent, par exemple, que des éléments de deux schémas sont similaires si leurs sous-concepts directs et/ou leurs super-concepts directs et/ou leurs concepts frères sont similaires [Do et Rahm, 2001], [Noy et Musen, 2001], [Thanh Le et al. 2004]. Ces techniques structurelles peuvent être basées sur la notion de point fixe [Melnik et al. 2002]. Dans S-Match [Giunchiglia et al. 2004], le problème de matching est vu comme un problème de satisfiabilité d'un ensemble de formules du calcul propositionel. Les graphes et les correspondances à tester sont traduits en formules de la logique propositionnelle en considérant la position des concepts dans le graphe et non seulement leur nom.

Approches sémantiques

Les techniques terminologiques et structurelles sont très souvent combinées entre elles. Certaines approches peuvent aussi recourir à des ressources extérieures aux schémas/ontologies à aligner telles que WordNet et proposent de compléter leurs résultats d'alignement par l'utilisation de techniques de matching sémantique [Kefi et Reynaud, 2006] [Fellah et al. 2008].

Les approches sémantiques [Aleksovski et al. 2006], [Stuckensschmidt et al. 2004], [Van Hage et al. 2005] exploitent les interprétations des concepts à aligner et utilisent généralement des connaissances de domaine tels que, les lexiques, les thésaurus, des ontologies de domaine. Elles peuvent aussi utiliser le voisinage ou les instances associées au concept pour définir son contexte et comprendre sa sémantique. Ces approches définissent des fonctions de similarité sous la forme de relations sémantiques telles que l'hyperonymie, l'hyponymie, méronymie, partie de, …) entre les intentions des concepts.

Dans les trois classes d'approches, WordNet est souvent utilisé comme ressource complémentaire car il est une bonne source d'information sur les synonymies et fournit une hiérarchie de concepts basée sur les relations de généralisation/spécialisation. La figure 2.11 présente la classification des approches d'alignement proposée par [Shvaiko et Euzenat, 2004].

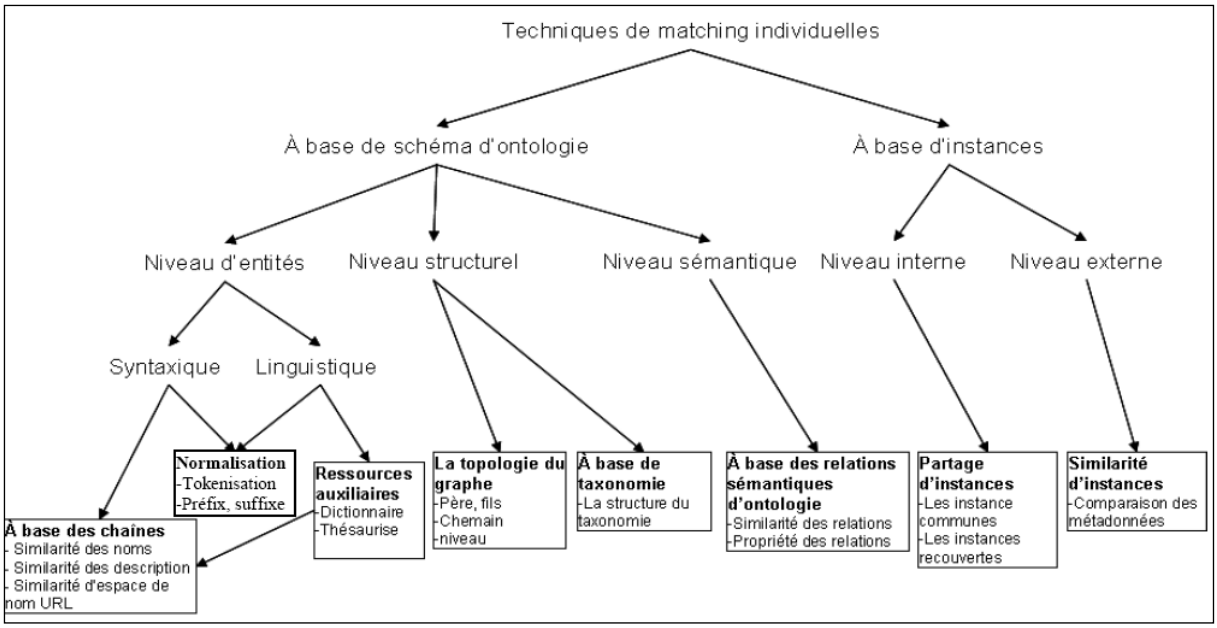


Figure 2.11 Classification des approches d'alignement proposée par [Shvaiko et Euzenat, 2004]

Approches basées sur la combinaison de plusieurs techniques

En se basant sur l'état de l'art sur le matching de schémas, les techniques terminologiques, structurelles et sémantique [Kefi et Reynaud, 2006] sont très souvent combinées entre elles pour améliorer les résultats du processus d'alignement, cette combinaison peut se faire selon différentes stratégies : la combinaison séquentielle en choisissant un ordre d'exécution (deux ou plusieurs techniques dans un même algorithme) comme le montre la figure 2.12 :

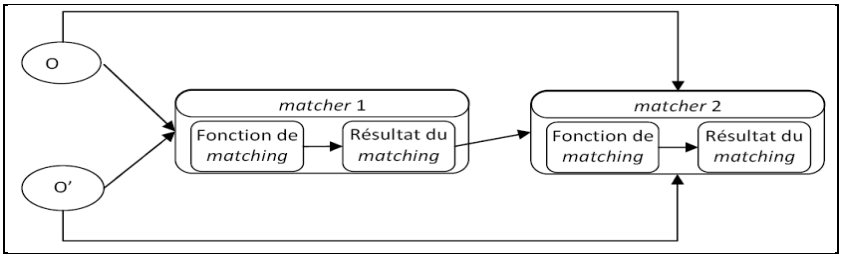


Figure 2.12 Combinaison séquentielle des systèmes de matching

Le deuxième type de combinaison comme le montre la figure 2.13 est la combinaison parallèle qui consiste à lancer parallèlement plusieurs matchers, puis par la suite à combiner leurs résultats de matching.

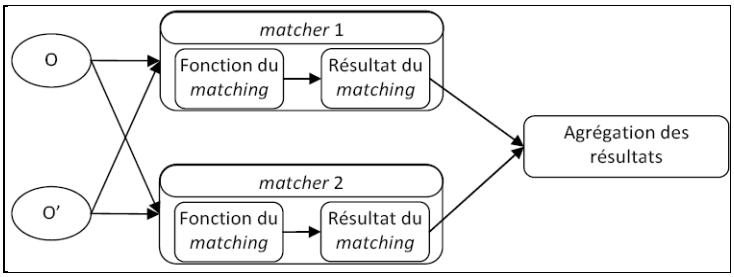


Figure 2.13 Combinaison parallèle des systèmes de matching

En conclusion, on remarque que la majorité des approches de matching se basent sur un processus semi-automatique et une application séquentielle de techniques terminologiques suivies des techniques structurelles et sémantiques, ces approches exploitent très souvent des connaissances du domaine (thésaurus, ontologies ou des alignements précédemment établis) et font intervenir l'utilisateur afin de compléter et valider les résultats d'alignement.

Parmi les travaux portant sur le matching de schémas, nous citons : COMA [Do et Rahm, 2001], Cupid [Madhavan et al. 2001], LSD [Doan et al. 2001], MOMIS [Bergamaschi et al. 2001], SemInt [Li et Clifton, 2000] et Similarity Flooding [Melnik et al. 2002].

Cupid est une approche de matching basée sur les schémas qui combine un matching linguistique avec un matching structurel. Elle se base sur la comparaison de schémas sans recourir aux instances. Le système SemInt se base sur la similarité entre les noms des éléments. Il repose sur une architecture neuronale afin de déterminer les éléments à mettre en correspondance. Le modèle Similarity Flooding repose sur une analyse de la structure des données. Cette approche implémente un algorithme de matching (appelé SF) basé sur des calculs de points fixes pour déterminer les correspondances sémantiques, cet algorithme se base sur l'idée de propagation de similarité. LSD utilise des techniques d'apprentissage automatique pour exécuter un matching bases sur les instances ainsi que sur les informations sur le schéma de données. LSD fait correspondre les nouvelles sources de donnes à un schéma global précédemment déterminé. COMA est une architecture qui elle aussi combine plusieurs systèmes de matching et se basent sur l'agrégation des résultats de ces matching.

Toutes les méthodes d'alignement déterminent des correspondances entre les entités en utilisant des mesures de similarité. Pour cela, de nombreuses mesures de similarité entre concepts ont été définies dans le cadre de l'alignement d'ontologies.

2.3.4.2 Mesures de similarités

Deux approches principales existent dans la littérature pour la mesure de similarité entre concepts dans une ontologie : (1) une approche basée sur les arcs qui utilise la structure

arborescente de l'ontologie et (2) une approche qui se base sur le contenu informatif des différents concepts en intégrant des mesures statistiques. D'autres approches proposent de combiner les deux (la structure de l'ontologie et le contenu informatif).

Approches basées sur les arcs

Les approches basées sur les arcs reposent uniquement sur la structure de l'ontologie. Les deux mesures les plus utilisées sont la mesure de Rada [Rada et al. 1989] et celle de Wu et Palmer [Wu et Palmer, 1994]. Elles se basent sur la distance en termes de nombre d'arcs séparant un concept d'un autre. [Rada et al. 1989] suggèrent que, pour mesurer la distance entre deux concepts ontologiques, on se base sur le nombre d'arcs minimum à parcourir pour aller du concept c_1 au concept c_2.

Dans le même ordre d'idée, [Wu et Palmer, 1994] définissent la similarité en fonction de la distance qui sépare deux concepts ontologiques dans la hiérarchie et également par leur position par rapport à la racine.

Ces mesures ont l'avantage d'être faciles à implémenter et peuvent donner une idée sur le lien sémantique entre les concepts. Cependant, elles ne prennent pas en compte le contenu du concept lui-même, ce qui peut conduire, dans certains cas, à une marginalisation de l'apport du concept en termes d'information.

Approches basées sur le contenu informatif

Les mesures de similarité suivant cette approche sont fondées sur la notion de Contenu Informatif qui utilise conjointement l'ontologie et le corpus. Le Contenu Informatif d'un concept traduit sa pertinence dans le corpus en tenant compte de sa spécificité ou de sa généralité. Pour ce faire, la fréquence des concepts dans le corpus est calculée et elle regroupe la fréquence d'apparition du concept lui-même ainsi que les concepts qu'il subsume (concepts fils). Les deux mesures les plus connues dans cette catégorie sont celles de Resnik [Resnik 1995] et Lin [Lin, 1998].

Resnik [Resnik, 1995] définit la similarité sémantique entre deux concepts par la quantité d'information qu'ils partagent : elle est égale au contenu informatif du concept le plus spécifique (plus petit généralisant *ppg*) qui subsume les deux concepts dans l'ontologie.
La mesure de Lin [Lin, 1998] considère que la description d'un concept est son contenu en information, et que les caractéristiques communes aux deux concepts sont quantifiées par le contenu en information du concept le plus spécifique généralisant les deux concepts. Cette

mesure permet de prendre en compte le concept le plus spécifique subsumant *c1* et *c2* ainsi que le contenu en information des concepts comparés. Contrairement à la mesure proposée par Resnik, elle permet de différencier la similarité entre plusieurs couples de concepts ayant le même subsumeur le plus spécifique.

Approches hybrides basées sur le contenu informatif et sur les arcs

La mesure de [Jiang et Conrath, 1997] prend en compte à la fois le contenu informatif du *ppg* et celui des concepts concernés. Par conséquent, elle peut pallier les limites de la mesure de Resnik. Dans le tableau 2.1, nous présentons les formules des mesures de similarité les plus connues.

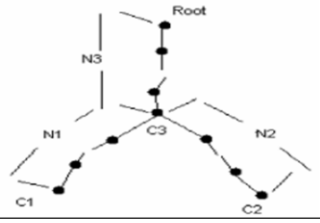
Mesure de Rada

$S(c_1,c_2)=\frac{1}{Dist_{edge}(c_1,c_2)}$ avec $Dist_{edge}(c_1,c_2)=nbLiens_{courtChemin}(c_1,c_2)$

Mesure de Wu-Palmer

$S(c_1,c_2)=\frac{2\times N3}{N1+N2+2\times N3}$ égale à $S(c_1,c_2)=\frac{2\times depth(c)}{depth_c(c_1)+depth_c(c_2)}$

Où c est le *PPG* (le plus petit généralisant) de c_1 et c_2 (en nombre d'arcs), *depth(c)* est le nombre d'arcs qui sépare c de la racine et *depthc*(c_i) avec i le nombre d'arcs qui séparent c_i de la racine en passant par c.



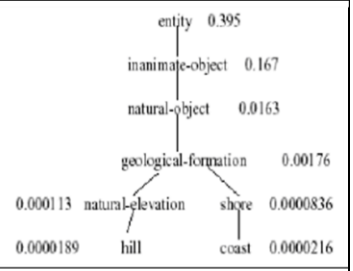
Mesure de Resnik

$S(c_1,c_2)=CI(ppg(c_1,c_2))$, où *ppg* est le plus petit généralisant et CI le contenu informationnel calculé comme suit:

$CI(c)=-\log(P(c))$

Où *P(c)* est la probabilité de retrouver une instance du concept c, calculées par : $P(c)=\frac{frequence(c)}{N}$ où N est le nombre total de concepts.

Voici un extrait de WordNet, le nombre attaché à chaque noeud est P(c)



Mesure de Lin

$S(c_1,c_2)=\frac{2\times\log(pms(c_1,c_2))}{\log(p(c_1))+\log(p(c_2))}$

Mesure de Jiang-Conrath

$S(c_1,c_2)=\frac{1}{Dist(c_1,c_2)}$ où $Dist(c_1,c_2)=\sum_{c\in pcc(c_1,c_2)} Poids(c, père(c))$ et $Poids(c, père(c))=CI(c)-CI(p)$

avec *pcc(c1,c2)* est l'ensemble des concepts du plus court chemin, *père(c)* est le concept parent de c.

Jiang simplifie le calcul de la distance par la formule : $Dist(c_1,c_2)=CI(c_1)+CI(c_2)-2\times CI(ppg(c_1,c_2))$

Tableau 2.1 Mesures de similarités conceptuelles

2.3.4.3 Fusion d'ontologies

Plusieurs travaux de recherche se sont intéressés à la fusion d'ontologies, nous citons par exemple : Ontomorph [McGregor et al. 1999], Chimaera [McGuinness et al. 2000] et PROMPT [Noy et Musen, 2000], Anchor-PROMPT [Noy et Musen, 2003], FCA-MERGE

[Stumme et Maedche, 2001] et QOM [Ehrig et Staab, 2004]. Une évaluation détaillée de ces outils est fournie dans [Noy et Musen, 2000].

Dans Anchor-PROMPT [Noy et Musen, 2003], les auteurs proposent une méthode de comparaison des ontologies de domaine par la définition des correspondances entre leurs concepts. Ils choisissent deux paires de concepts équivalents comme référence. Chaque paire appartient à une ontologie. Ensuite, ils sélectionnent tous les concepts intermédiaires deux à deux qui occupent les mêmes positions dans deux chemins de même longueur, reliant les deux concepts de la même paire. Cela permet aux auteurs de juger si ces paires de concepts sont équivalents ou pas. Selon les auteurs, deux concepts se trouvant dans la même position entre deux concepts équivalents, sont également équivalents. Anchor-PROMPT suppose que les deux ontologies sont construites de la même façon. Ce n'est pas le cas dans la réalité.

Dans FCA-Merge [Stumme et Maedche, 2001], les auteurs définissent une méthode formelle et ascendante de fusion des ontologies en se basant sur un ensemble de documents. Ils appliquent des techniques de traitement du langage naturel et d'analyse formel de concepts pour dériver le treillis des concepts. Ce dernier est exploré et transformé en une ontologie par l'intervention de l'être humain. Par contre, plusieurs travaux sur l'alignement des ontologies ont été développés. Ils traitent une étape du processus de fusion qui est la découverte des correspondances entre les entités des ontologies à aligner.

Quelques auteurs comme [Aleksovski et al. 2006] ont montré l'avantage d'utiliser la connaissance du domaine dans des cas définis. D'autres approches comme ASCO [Bach et al. 2004], GLUE [Doan et al. 2004], QOM [Ehrig et Staab, 2004] et OLA [Euzenat et al. 2005] ont été développées pour supporter le processus d'alignement des ontologies. OLA par exemple décrit les ontologies comme deux graphes-OWL et utilise la mesure de similarité de Valtchev [Valtchev, 1999] pour comparer les entités appartenant à la même catégorie (propriété, instance).

2.3.5 Approches de construction de Topic Maps

Nous proposons dans ce qui suit une classification des approches de construction de Topic Map, nous distinguons quatre classes : les approches basées sur la génération automatique de Topic Map à partir de RDF ou de documents XML, les approches utilisant une démarche de construction collaborative faisant intervenir plusieurs utilisateurs, les approches fondées sur des techniques d'analyse linguistiques de documents textuels pour la

création et l'enrichissement de Topic Map et enfin les approches basées sur des techniques de fusion de plusieurs Topic Maps pour l'élaboration d'une Topic Map globale et unifiée.

2.3.5.1 Génération automatique de Topic Maps

Des travaux de construction de Topic Maps proposent de générer automatiquement des Topic Maps soit à partir de métadonnées RDF comme par exemple les travaux de [Ogievetsky 01], [Pepper, 2002b], [Gronmo, 2002] et [Garshol, 2003], soit à partir de documents XML en explorant les balises XML comme le montre la figure 2.14 et en utilisant des transformations XSLT [Reynolds et Kimber, 2002] [Librelotto et al. 2008], ou bien à partir de schémas de bases de données [Pepper, 2002b] comme l'illustre la figure 2.15. Ces approches nécessitent souvent l'intervention d'experts du domaine pour corriger et améliorer les Topic Maps générées automatiquement.

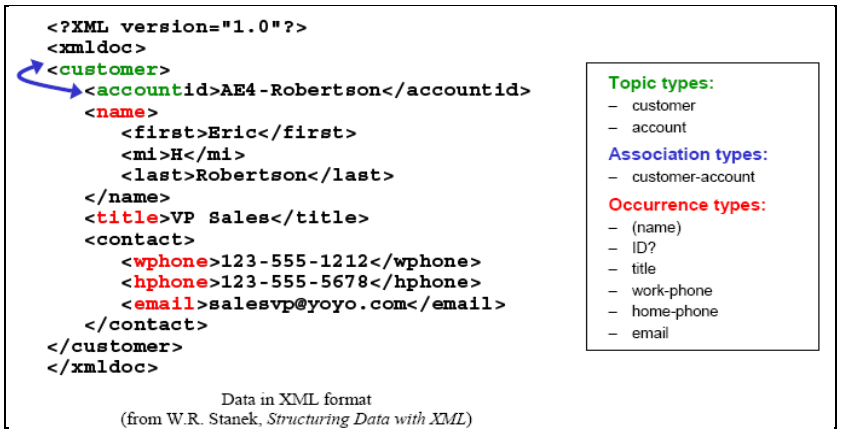


Figure 2.14 Génération automatique de Topic Map à partir données au format XML [Pepper, 2002b]

Gronmo **[Gronmo, 2002]** dans ses travaux, s'est intéressé à la traduction de métadonnées RDF en Topic Map, il propose un processus de construction en trois étapes qui accepte en entrée des sources de données hétérogènes (bases de données relationnelles, sites Web, systèmes d'information, etc), Avant de commencer, il propose de traduire ces données selon le modèle RDF, la première étape consiste à identifier les sujets dont parlent ces données, ensuite traduite ces sujets en des triplets RDF (sujet, propriété, valeur). La dernière étape concerne la transformation de ces triplets en des Topics ou des caractéristiques de Topics. Un triplet peut être considéré comme un Topic, une association, un nom de Topic ou une occurrence. Enfin, puisque chaque Topic Map provient d'une source de données, ces Topic Maps seront par la suite fusionnées pour obtenir une Topic Map globale décrivant

toutes les sources en entrée, les techniques de fusion utilisées sont celles proposées par le modèle des Topic Maps.

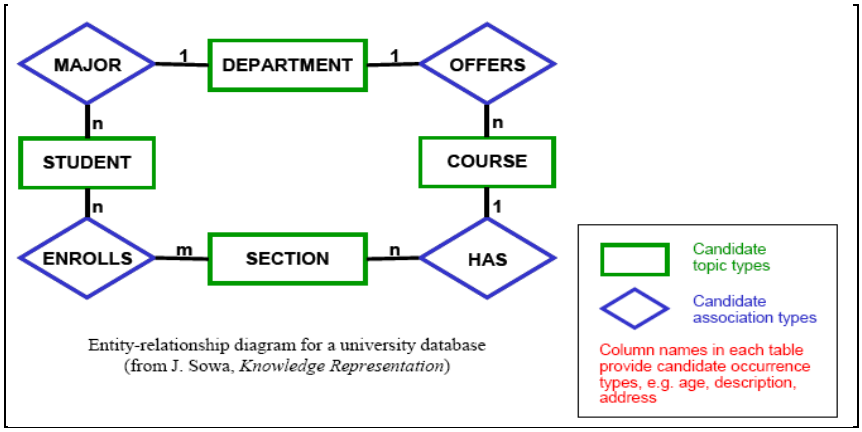


Figure 2.15 Génération automatique de Topic Map à partir de schémas de BD [Pepper, 2002b]

Plusieurs approches se sont intéressées aux mappings entre le standard RDFS et les Topic Maps **[Moore, 2001]**. En effet, le modèle des données correspondant aux deux formalismes peut être structuré sur trois niveaux : le niveau sémantique, le niveau de représentation objets et le niveau du formalisme. L'intégration des modèles est rendue possible grâce à la vision commune du niveau intermédiaire (de la représentation objets) sous forme de données semi-structurées, c'est-à-dire sous forme de graphe. Il suffit alors de réaliser une transformation d'un graphe Topic Map en un graphe RDFS, le graphe Topic Map apparaissant d'ailleurs beaucoup moins complexe. Nous pouvons trouver dans [Newcomb, 2002] la structure de graphe abstrait pour la représentation des relations entre Topics. Ainsi, il choisit de représenter les notions de Topic et d'association par des classes et la notion de membre d'association et de contexte d'association par des propriétés RDFS, il obtient alors des déclarations telles que :

```
1 <rdfs:Class rdf:ID="topic"/>,
2 <rdfs:Class rdf:ID="association"/>,
3 <rdfs:Property rdf:ID="associationMember"/>
4 <rdfs:Property rdf:ID="associationScope"/>
```

[Ogievetsky, 2001] montre une implémentation complète de la façon de transformer le modèle Topic Map en RDFS. Une autre approche présentée dans **[Roberson et Dicheva, 2007]** prend en entrée des sites Web et propose une génération automatique des éléments de la Topic Map en utilisant des techniques de crawling sur les documents HTML. Le crawling consiste à visiter un site Web à partir de son URL et ensuite traverser toutes les pages de ce

site en suivant les hyperliens contenus dans ces pages, ils proposent aussi un ensemble d'heuristiques pour extraire les informations à partir des documents HTML cette approche a été implémentée et l'outil développé est un plug-in pour l'éditeur TM4L qui permet d'extraire automatiquement des Topics et des relations entre ces Topics à partir d'un site Web spécifié par l'utilisateur.

2.3.5.2 Construction collaborative de Topic Maps

Certains travaux comme ceux présentés dans [Ahmed, 2003], [Lavik et al. 2004], [Zaher et al. 2006] et [Dicheva et Dichev, 2006] décrivent des processus de construction collaborative de Topic Maps. Cette co-construction fait intervenir plusieurs utilisateurs, qui peuvent être distants et connectés à travers un réseau comme par exemple le travail de Ahmed **[Ahmed, 2003]** qui propose une application peer to peer permettant d'échanger des Topic Maps ou des fragments de Topic Map dans un environnement distribué, chaque utilisateur construit un fragment de Topic Map appelé aussi Topic Map locale, ensuite ces fragments sont fusionnés pour obtenir une Topic Map globale et unifiée. BrainBank Learning **[Lavik et al. 2004]** est un environnement pour le e-learning basé sur une construction collaborative de Topic Map, c'est une application Web où chaque apprenant construit sa propre Topic Map et cette dernière peut, par la suite, être consultée par d'autres apprenants, chaque Topic est relié aux ressources qui parlent de ce Topic donc lors de la navigation, chaque apprenant peut accéder à tous ces documents qui sont sauvegardés sur le serveur de l'application BrainBank learning.

[Zaher et al. 2006] proposent une approche de co-construction collaborative d'un annuaire métier fondée sur le modèle des Topic Maps pour la société « Airbus-Ingénierie ». Cet annuaire offre la possibilité aux ingénieurs de rechercher des collègues par navigation dans une carte de thèmes multipoints de vue. Chacun d'eux peut y déclarer ses caractéristiques métier de façon très détaillée, selon une structure d'index comportant plusieurs arborescences qu'il peut aussi contribuer à enrichir thématiquement.

Pour la mise en œuvre de leur approche, les auteurs définissent le méta-modèle HyperTopic comme une extension du modèle des Topic Maps, ce méta-modèle vise à la fois la représentation des objets métiers et celle des acteurs sociaux et de leur activité. Il offre un support pour la manipulation de la sémantique des notions relevant de ces deux familles d'objets. Ce projet a été développé au sein de l'équipe Tech-CICO de l'Université de Technologie de Troyes, comme un exemple de l'approche du « Web socio-sémantique» proposée par [Zacklad et al. 2003b].

Une plateforme a été également implémentée dans le cadre de ce projet, appelé « Agoræ », cette plateforme a pour objectif de permettre à plusieurs acteurs de partager un contenu métier en se basant sur la structure d'une Topic Map multipoints de vue.

Le travail décrit dans **[Dicheva et Dichev, 2006]** propose l'environnement TM4L (Topic Maps for Learning) qui permet la création, la maintenance et l'utilisation de ressources pédagogiques disponibles en ligne dans le domaine du e-learning basé sur le modèle des Topic Maps.

Une autre utilisation du modèle des Topic Maps dans le domaine du e-learning est le projet MEMEORAe (MEMoire ORganisationnelle Appliquée à l'e-learning) [Abel et al. 2003], **[Benayache, 2005]**. Dans ce projet, les auteurs proposent de gérer les ressources pédagogiques d'une organisation au moyen d'une « mémoire organisationnelle de formation » basée sur des ontologies. Les ressources pédagogiques sont mises directement à la disposition des apprenants et chaque apprenant pourra sélectionner et organiser les ressources pertinentes, selon ses choix. Pour indexer ces ressources et fournir un bon moyen de visualisation et de navigation de la mémoire, les auteurs ont fait le choix d'utiliser la norme ISO Topic Maps.

2.3.5.3 Approches basées sur des techniques d'analyse linguistique de documents textes

Des techniques d'apprentissage et de traitement du langage naturel [Legrand et Soto, 2002a], [Legrand et Soto, 2002b], [Folch et Habet, 2002], [Böhm et al. 2002], [Korthaus et al. 2004], [Kasler et al. 2006] ont été appliquées pour l'élaboration de Topic Map à partir de documents textes. Nous citons par exemple la méthode de **[Legrand et Soto, 2002a]** qui proposent un système de recherche d'information sur le Web basé sur les Topic Maps, elle applique un algorithme de classification conceptuelle (AFC Analyse Formelle de Concepts) pour le filtrage de la Topic Map extraite à partir de site Web, cet algorithme consiste à regrouper les Topics qui ont des propriétés communes en des clusters des Topics, et ce pour avoir différents niveaux dans la Topic Map. Ces clusters serviront après dans l'étape de visualisation de la Topic Map.

L'approche décrite dans **[Folch et Habert, 2002]** utilise aussi des techniques d'analyse de documents textuels pour construire une Topic Map multi points de vues à partir de forums de discussion, ces techniques sont bases sur l'identification de classes de mots en utilisant des patrons linguistiques, ces classes constituent des Topics potentiels de la Topic Map résultante.

L'approche présentée dans **[Kasler et al. 2006]** consiste à développer un outil pour la génération semi-automatique de Topic Maps à partir d'un ensemble de documents textes (des

articles d'une conférence pendant 10 années disponibles dans deux langues, anglais et hongrois) en utilisant des techniques d'apprentissage (*machine learning techniques*). L'objectif de cet outil est de fournir un portail permettant de gérer et structurer le contenu de cette base documentaire. Tout d'abord, les auteurs commencent par transformer les documents sources en un format unifié (XML), ensuite, ils procèdent à l'extraction des Topics et des associations à partir des documents, cette étape est basée sur deux techniques : l'utilisation de la structure et des métadonnées descriptives des documents pour identifier des Topics tels que l'auteur, le titre de l'article, l'affiliation de l'auteur ; la deuxième technique consiste à utiliser une ressource externe (par exemple une ontologie) pour extraire des mots clés à partir des documents et les représenter sous forme de Topics, les liens entre les Topics sont définis grâce à l'ontologie externe avec l'aide d'experts du domaine. La Topic Map résultante est ensuite chargée dans un serveur de stockage pour être par la suite visualisée par les utilisateurs.

2.3.5.4 Construction par fusion de Topic Maps

Un des objectifs du modèle de Topic Map est de permettre **l'intégration de ressources hétérogènes et distribuées** provenant de différentes sources et disponibles dans différentes langues, plusieurs travaux [Ouziri, 2006], [Korthaus et al. 2006], [Godehardt et Bhatti, 2008], [Jiang et al. 2009] ont adopté le modèle les Topic Maps pour répondre à ce problème. La plupart de ces travaux utilisent des règles de fusion définies dans la norme XTM pour fusionner plusieurs Topic Maps, ces Topic Maps peuvent être écrites dans des syntaxes différentes ou générées selon des techniques variées ou provenant de sources différentes.

Comme nous l'avons vu précédemment, le modèle Topic Maps dispose de mécanismes flexibles pour l'annotation sémantique de ressources. Néanmoins le but de ce standard n'est pas seulement l'annotation de ressources en fonction des sujets fournis par le créateur de la Topic Map ou par une communauté d'utilisateurs précise, mais aussi l'interopérabilité sémantique des Topic Maps créés, c'est-à-dire la possibilité d'échanger et de fusionner des Topic Maps entre communautés différentes.

L'obstacle à cette interopérabilité sémantique est la difficulté d'établir des correspondances entre les sujets représentés par des Topics qui existent dans des Topic Maps différentes, car le même sujet peut être exprimé par des Topics ayant des noms différents et inversement des Topics ayant le même nom peuvent représenter des sujets différents (c'est le cas des Topics ayant des noms polysémiques). Afin de réduire cet obstacle, des groupes

d'activité[21] travaillent dans le cadre des organismes ISO et OASIS sur la normalisation d'un certain nombre de répertoires de sujets, appelés aussi des *sujets publics*. Chaque Topic peut alors faire référence à un sujet normalisé. Deux Topics faisant référence au même sujet public peuvent être fusionnés donnant lieu ainsi à un nouveau Topic qui contient l'union des caractéristiques (occurrences, associations, noms) des deux Topics.

Les **règles de fusion** selon le modèle des Topic Maps sont résumées comme suit :

- Lorsque deux Topic Maps sont fusionnées, les Topics **faisant référence au même sujet public** peuvent être fusionnés en unn seul Topic ;
- Lorsque deux Topics sont fusionnés, le Topic résultat contient **l'union des caractéristiques** (occurrences, associations, noms) des deux Topics sources.

Notons que la plupart des approches de construction de Topic Maps décrites précédemment [Folch et Habert, 2002] [Librelotto et al. 2004] [Zaher et al. 2006] utilisent des techniques de fusion dans le processus de construction puisqu'elles prennent en compte plusieurs types de ressources, une Topic Map est construite pour chaque type de ressource et après les Topic Maps résultantes sont fusionnées pour avoir une seule Topic Map unifiée et décrivant toutes les ressources.

L'approche définie par **[Ouziri, 2006]** est elle aussi basée sur des techniques de fusion de Topic Maps pour l'intégration sémantique de ressources pédagogiques disponibles sur le Web dans le contexte du e-learning. L'auteur définit une approche à trois étapes basée sur le modèle des Topic Maps, la première consiste à la représentation d'objets pédagogiques distribués sous forme de Topic Maps, ensuite enrichir sémantiquement ces objets en utilisant une ontologie externe de domaine et enfin intégrer sémantiquement ces objets. Ces objets appartiennent à différentes ressources pédagogiques distribuées, une Topic Map est construite pour chaque objet et la phase d'intégration revient à fusionner ces Topic Maps pour obtenir une seule Topic Map unifiée et cohérente permettant de représenter le contenu pédagogique des ressources en entrée.

Un autre travail présenté dans **[Jiang et al. 2009]** dans le domaine du e-learning propose un outil de formation et d'enseignement pédagogique, cet outil (appelé, ETM ToolKit) fournit des services de recherche, de navigation dans des ressources pédagogiques volumineuses et distribuées permettant à des enseignants de construire des bases de connaissances partageables et utilisables par des étudiants. ETM ToolKit utilise un nouveau méta-modèle de Topic Map appelé ETM pour « *Extended Topic Map* » [Lu et al. 2008]

[20] http://www.oasis-open.org/apps/org/workgroup/tm-pubsubj.

permettant d'intégrer sémantiquement des ressources pédagogiques distribuées. Ce méta-modèle est une extension du modèle des Topic Maps, défini dans le but de supporter des applications de fusion de Topic Maps distantes. Les auteurs [Lu et al. 2008] proposent des algorithmes et des mesures de similarité pour la fusion de Topic Maps, ces algorithmes sont basés sur le calcul des similarités terminologiques et des similarités sémantiques (en utilisant WordNet) entre les composants des Topic Maps sources.

Travaux sur la fusion de Topic Maps

Le modèle des Topic Maps définit une fonction générique de fusion appelée **« *MergeMap* »** qui se base sur les règles de fusion présentées dans le paragraphe précédent. Ces règles utilisent des principes **d'équivalence** pour déterminer si deux ou plusieurs Topics peuvent être fusionnés, ces principes permettent seulement d'évaluer l'égalité entre les entités composant une Topic Map. Ils ne prennent pas en compte la similarité entre ces entités.

Pour répondre à ce problème, plusieurs travaux de recherché se sont intéressés à la fusion de Topic Map et plus particulièrement à la mesure de similarité entre les éléments de deux Topic Maps, nous citons par exemple les travaux de **[Maicher et Witschel, 2004]** qui ont définit **SIM *(Subject Identity Measure)*** pour mesurer la similarité entre deux Topics en se basant sur la similarité de leurs noms et de leurs occurrences. Cependant, le processus de calcul de similarité consiste à comparer seulement des chaînes de caractères composant les noms des Topics et les occurrences. La structure hiérarchique et les associations ne sont pas prises en compte dans la fusion. L'approche proposée par **[Wu et al. 2006]** appelée ***Topic and Occurrence-oriented Merging* (TOM)** se base sur le principe que deux Topics peuvent être fusionnés s'il existe une similarité entre leur noms et les occurrences ou les ressources liées à ces Topics. **[Kim et al. 2007]** proposent **TM-MAP**, une technique de matching multi-stratégies qui permet de mesurer les similarités selon quatres facettes : similarité basée sur les noms, similarité basée sur les propriétés des Topics, similarité basée sur la hiérarchie de Topics et similarité basée sur les associations. Dans [Kim et al. 2007], les auteurs proposent un processus de matching des différentes entités d'une Topic Map à savoir : les Topics, les occurrences, les associations, et les relations hiérarchiques, ce processus prend en entrée deux Topic Maps sous format XTM et un dictionnaire du domaine et est composé de sept étapes :

1) Initialisation qui consiste à récupérer les PSI (*Public Subject Identifier*) et les index des Topics : ID, Nom, Occurrence, Type (qui peut être soit type de Topic

ou Topic instance) et Scope, à partir des documents XTM correspondant aux deux Topic Maps sources ;

2) Génération des paires d'entités ;
3) Sélection des pairs à aligner ;
4) Calcul de similarité entre les entités basé sur des techniques linguistiques, le résultat est un matching composé prenant en compte le résultat des matching (exécutés indépendamment) basés sur les noms, les occurrences, les hiérarchies et les associations ;
5) Regroupement des quatre opérations de matching pour générer une seule mesure de similarité pour chaque paire d'entités ;
6) Choix des matching candidats pour chaque entité (ayant le meilleur résultat de similarité (selon un seuil max et min déjà fixés) ;
7) Post-matching qui consiste à corriger les erreurs des résultats des alignements générés automatiquement par des experts du domaine.

Cependant ces approches utilisent des techniques de similarité terminologiques c'est-à-dire que la comparaison est basée sur les chaines de caractères (qui composent les noms des Topics par exemple et des techniques de similarité structurelles qui prend en compte la hiérarchie des Topics et leurs positions dans la Topic Map (les pères et les fils d'un Topic). La similarité sémantique entre les Topics n'est pas prise en compte dans la plupart des approches

Pour répondre à ce besoin, des travaux comme **[Korthaus et al. 2009] et [Lu et Feng, 2009]** proposent une approche de fusion pour trouver des correspondances entre ontologies en se basant sur les caractéristiques et les contraintes syntaxiques et sémantiques des Topic Maps. Un autre travail récent [Lu et al. 2009] [Lu et al. 2010] a été également proposé dans le cadre des applications d'intégration de ressources hétérogènes, cette étude concerne principalement l'extension de la fonction de fusion définie dans le modèle des Topic Maps et la proposition de nouvelles mesures pour calculer les similarités entre les éléments d'une Topic Map. Dans leur travaux, [Lu et al. 2010] se basent sur le méta-modèle *Extended Topic Map* défini dans [Lu et al. 2008].

2.3.6 Outils d'édition et de visualisation de Topic Maps

Plusieurs outils ont été développés pour l'édition et la visualisation des Topic Maps. Nous distinguons deux types d'outils, le premier (type Web) affiche des index à partir desquels il est possible de choisir un Topic et de voir les informations qui sont rattachées à ce

dernier, le deuxième type est caractérisé par une navigation graphique sous forme d'arbre parfois dynamique et en temps réel. Parmi les éditeurs manipulant les Topic Maps, nous citons : l'outil K42, Mondeca, TM4J, Topic Map Designer, Ontopia Navigator Framework, TM4L et UNIVIT.

K42[22]: L'outil K42 développé par la société Empolis représente les Topic Maps sous forme d'arbre hyperbolique. L'interface fournit par K42 permet à l'utilisateur de regarder toute l'information sur un écran simple, montrant les liens entre les Topics et où ils apparaissent dans la hiérarchie. En dehors de cette navigation de type Web, d'autres types de visualisation sont disponibles. La figure 2.16 illustre un exemple de Topic Map représentée avec l'outil K42.

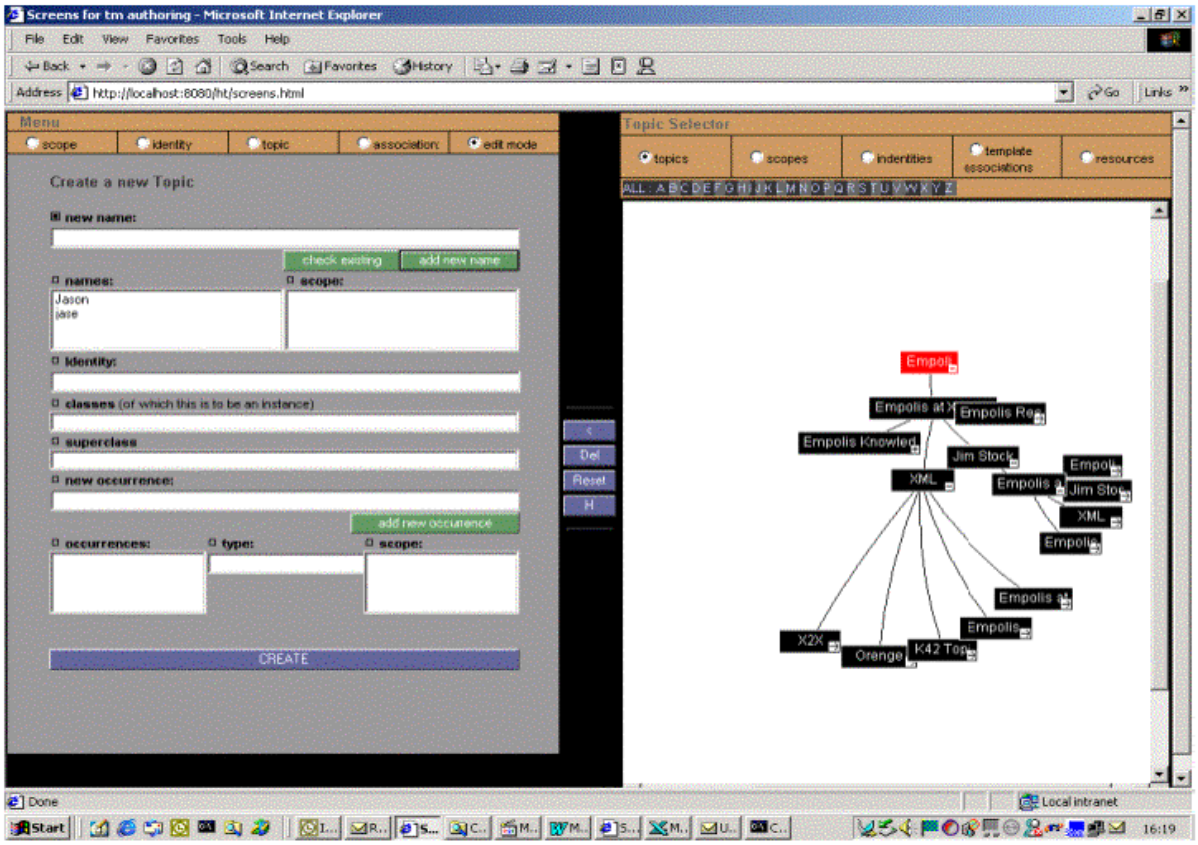


Figure 2.16 Application Empolis K42 TMV

Mondeca[23] : Le produit développé par Mondeca se présente sous la forme d'une interface Web permettant la saisie et la création de Topic Maps en temps réel, selon les besoins de l'utilisateur, ou selon ce qu'il a le droit de voir. Une fois la Topic Map saisie, ou bien après l'import d'un fichier au format Topic Map (format XTM), des fichiers XML

[22] Empolis Release K42 V2.0 : http://k42.empolis.co.uk/
[23] Mondeca : http://www.mondeca.com/

contenant l'information et répondant à une certaine DTD, propriétaire de Mondeca, sont générés.

Le navigateur développé par Mondeca construit des représentations graphiques en temps réel comme l'illustre la figure 2.17

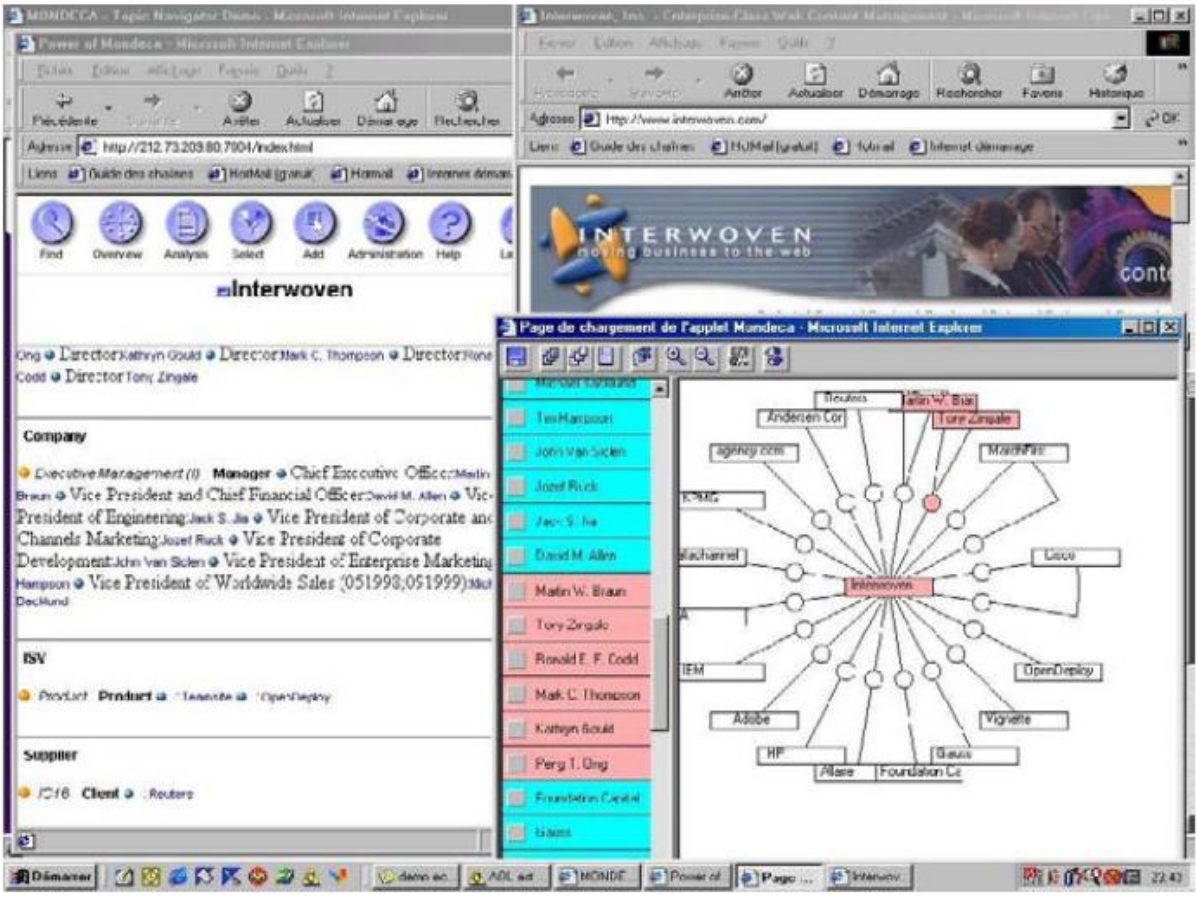


Figure 2.17 Navigateur Mondeca

TM4J[24] **:** TM4J représente une suite de packages Java qui fournissent des interfaces de navigation et transfèrent des réalisations existantes pour l'importation, la manipulation et l'exportation des Topic Maps codées pour se conformer au DTD de XTM. Ce dispositif inclut les fonctionnalités suivantes :

- Modèle objet conforme aux spécifications de la norme XTM 1.0 ;
- Un ensemble simple de lignes de commande pour la création et la mise à jour des Topic Maps ;
- Possibilité de fusionner deux Topic Maps ou plus ;
- Stockage des Topic Maps dans une base de données relationnelle ;

[24] http://tm4j.org/

- Possibilité de l'importation des données de XTM et de LTM d'Ontopia (*Linear Topic Map*) ;
- Possibilité d'exportation des Topic Maps en format XTM.

La Figure 2.18 est un exemple de visualisation dynamique de Topic Map fournie par l'outil TM4J. Cette représentation utilise le logiciel TheBrain[25] .

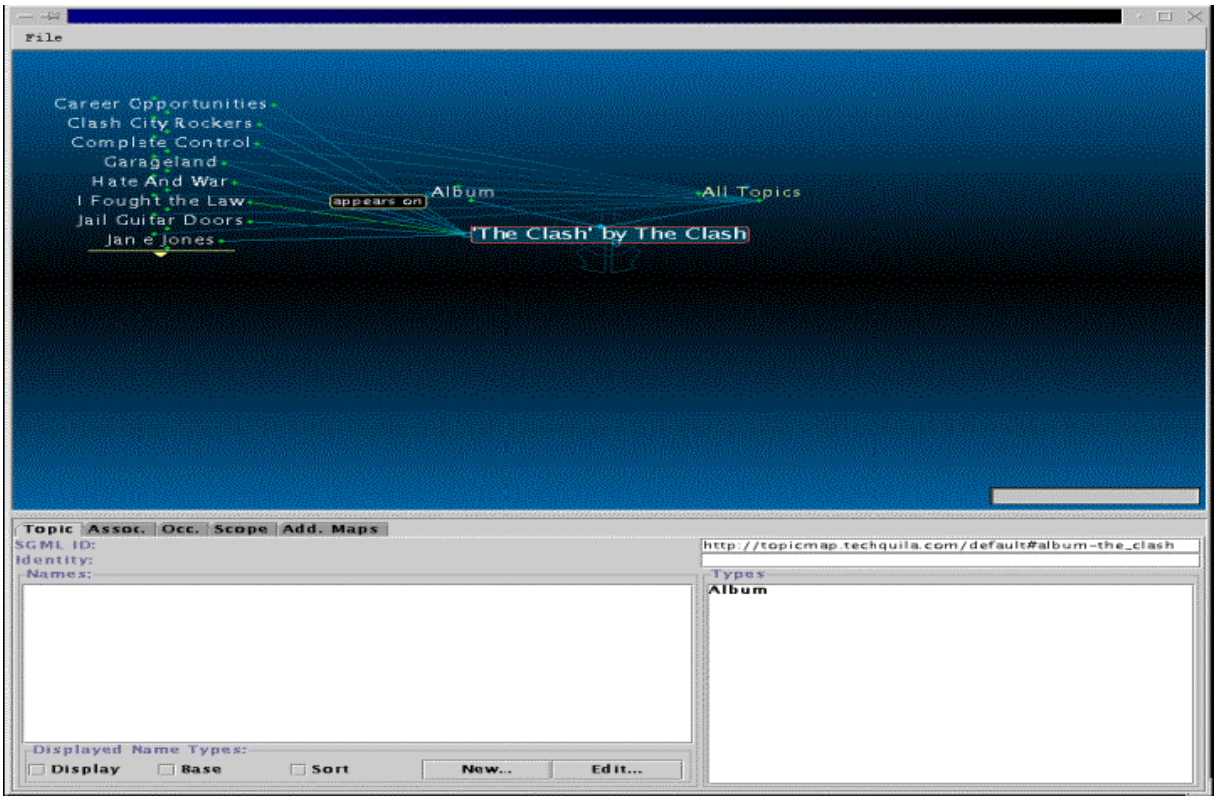


Figure 2.18 Visualisation avec le logiciel TheBrain

Topic Map Designer[26] : Topic Map Designer est un éditeur de Topic Maps qui supporte uniquement la norme ISO 13250, il permet d'exporter des Topic Maps au format XTM 1.0, mais il n'est pas destiné à être un environnement générique de création des Topic Maps. Topic Map Designer est facile à maîtriser. Dans le Treeview sur le côté gauche se trouve tous les Topics et associations. En choisissant un concept à partir du Treeview, toutes les informations sur le Topic (ou Association) sont affichées. Tous les changements (mises à jours) sont immédiatement reflétés dans Treeview.

[25] http://www.thebrain.com
[26] Topicmao-Design : http://www.topicmap-design.com/

Ontopia Navigator Framework[27] **:** *Ontopia Navigator Framework* est destiné au développement rapide des applications conformes aux enchaînements de Java 2 Enterprise Edition (J2EE) en utilisant des Topic Maps. Ontopia Navigator Framework est un langage de script XML optimisé pour le développement des applications de Topic Map. L'utilisation de ce langage permet aux utilisateurs de rassembler facilement les informations de la Topic Map, d'effectuer des opérations complexes comme la définition des scopes, la sélection par nom, etc, et de produire des résultats en format HTML, XML, ou n'importe quelle autre format. Il est conçu pour être facile à apprendre, puissant et extensible. La technique de navigation dans Ontopia est similaire à celle des navigateurs Web, l'utilisateur clique sur un lien pour ouvrir un nouveau Topic, une association ou une ressource. La figure 2.19 illustre un exemple de visualisation avec le navigateur Ontopia.

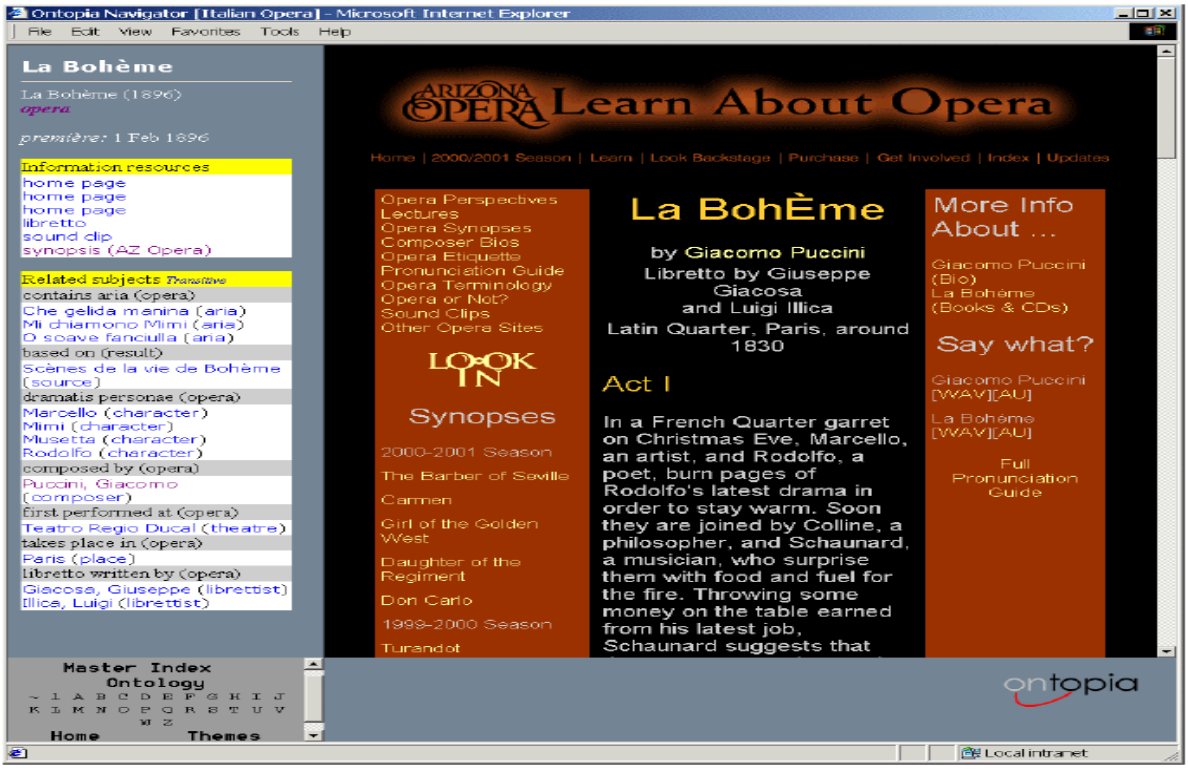


Figure 2.19 Navigateur Ontopia

TM4L[28] : l'environnement TM4L ou Topic Maps 4 E-Learning, est composé principalement d'une interface d'édition et d'une interface de visualisation de Topic Maps (*TM4L Viewer*). L'interface d'édition de TM4L (*TM4L Editor*) est un éditeur d'ontologie permettant à l'utilisateur la construction d'ontologie de ressources pédagogiques en utilisant

[27] Ontopia Solution – Navigator Framework : http://www.ontopia.net/
[28] TM4L: http://compsci.wssu.edu/iis/nsdl/index.html

des Topic Maps. Le contenu pédagogique ainsi crée est entièrement conforme à la norme XTM. L'éditeur de TM4L est basé sur la norme Topic Maps, en effet les objets manipulés sont : des Topics, des associations, des ressources et des contextes. Il inclut quatre sections différentes : Topic Map, Topics, Relations, et Thèmes.

La figure 2.20 montre quelques exemples d'interfaces de TM4L Editor et la figure 2.21 illustre un exmple de visualisation de Topic Map avec le même outil TM4L Editor. Il est également possible de visualiser la Topic Map avec TM4L Viewer comme le montre la figure 2.22 ou encore avec l'outil TMNav[29] (une application Java pour naviguer dans une Topic Map) comme l'illustre la figure 2.23.

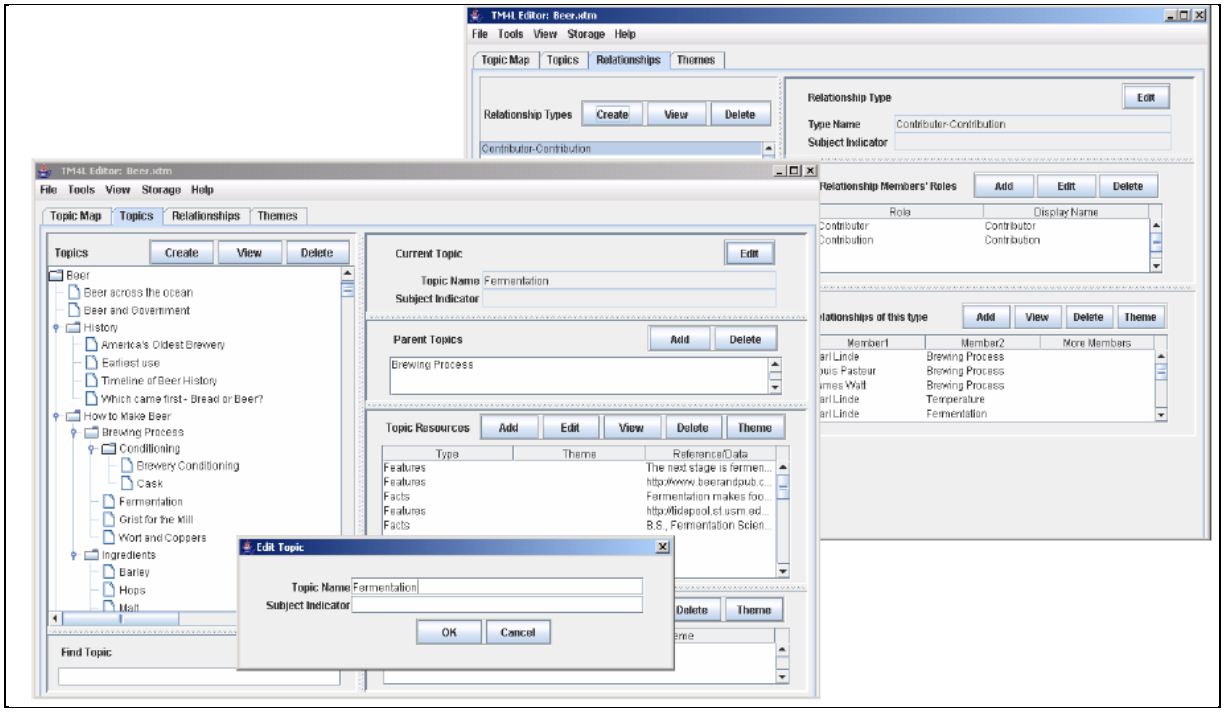


Figure 2.20 Exemple d'interfaces de TM4L Editor : création de Topics et d'associations

[29] http://tm4j.org/tmnav.html

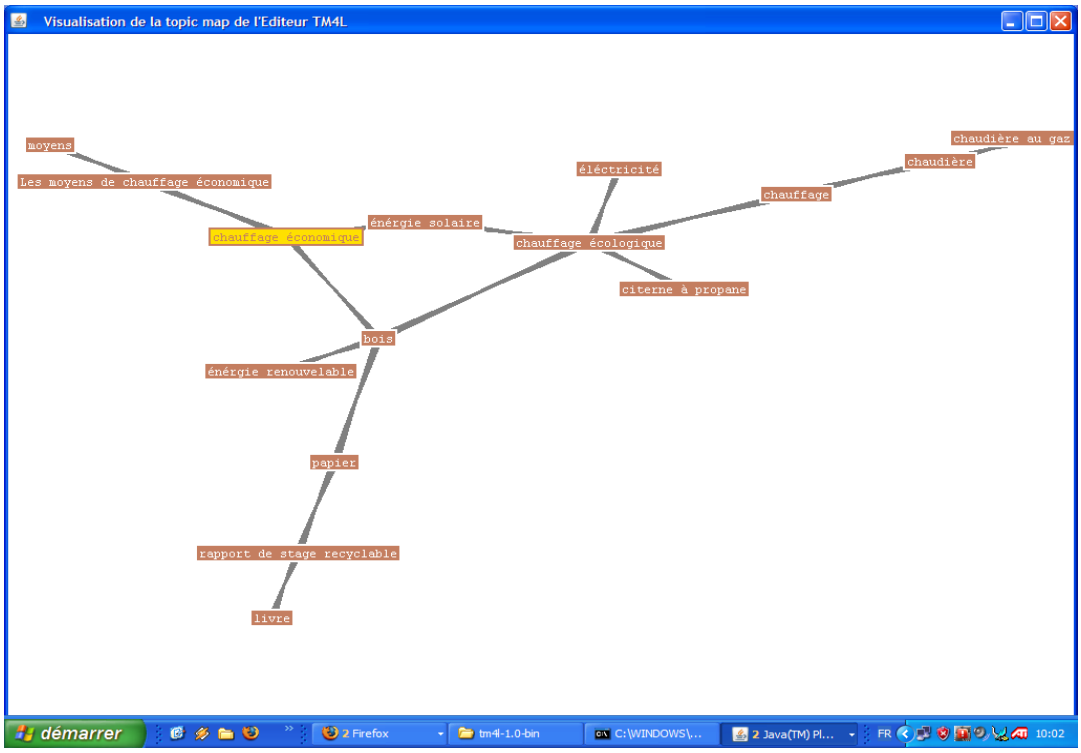


Figure 2.21 Exemple de visualisation avec TM4L Editor

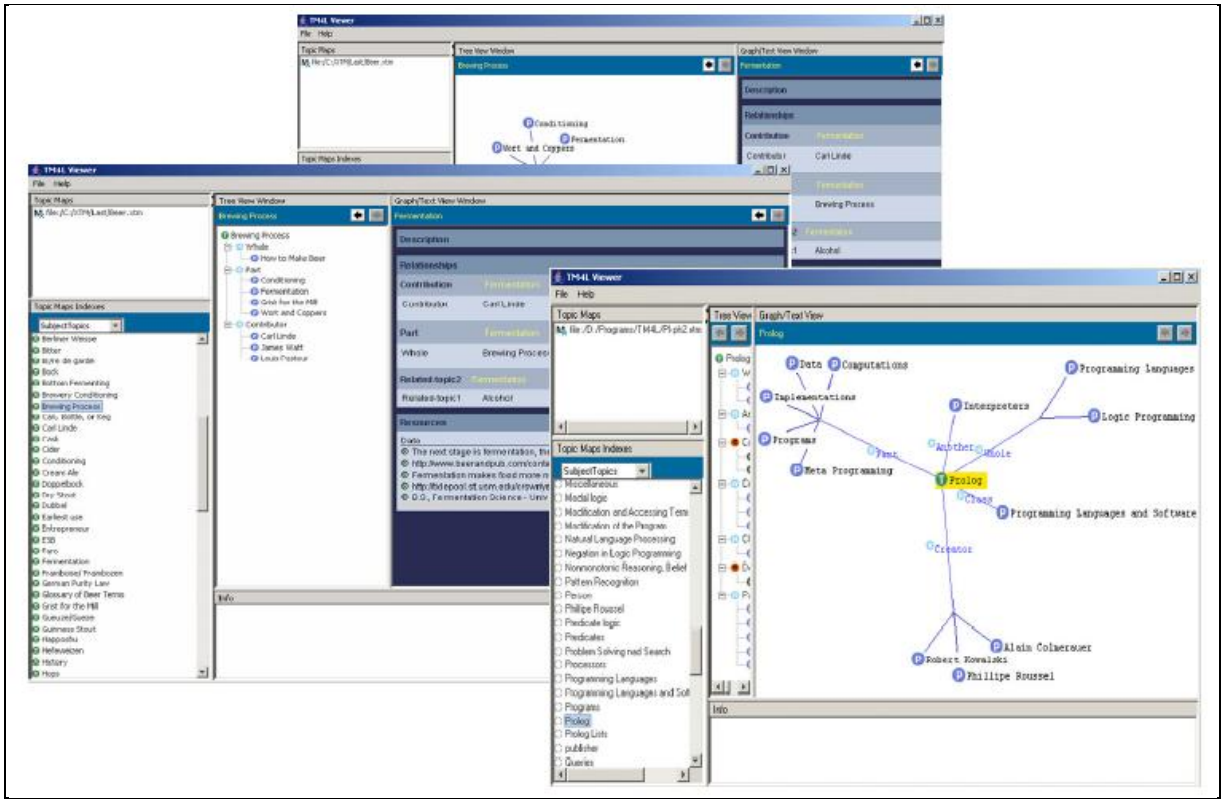


Figure 2.22 Interfaces de visualisation de TM4L Viewer

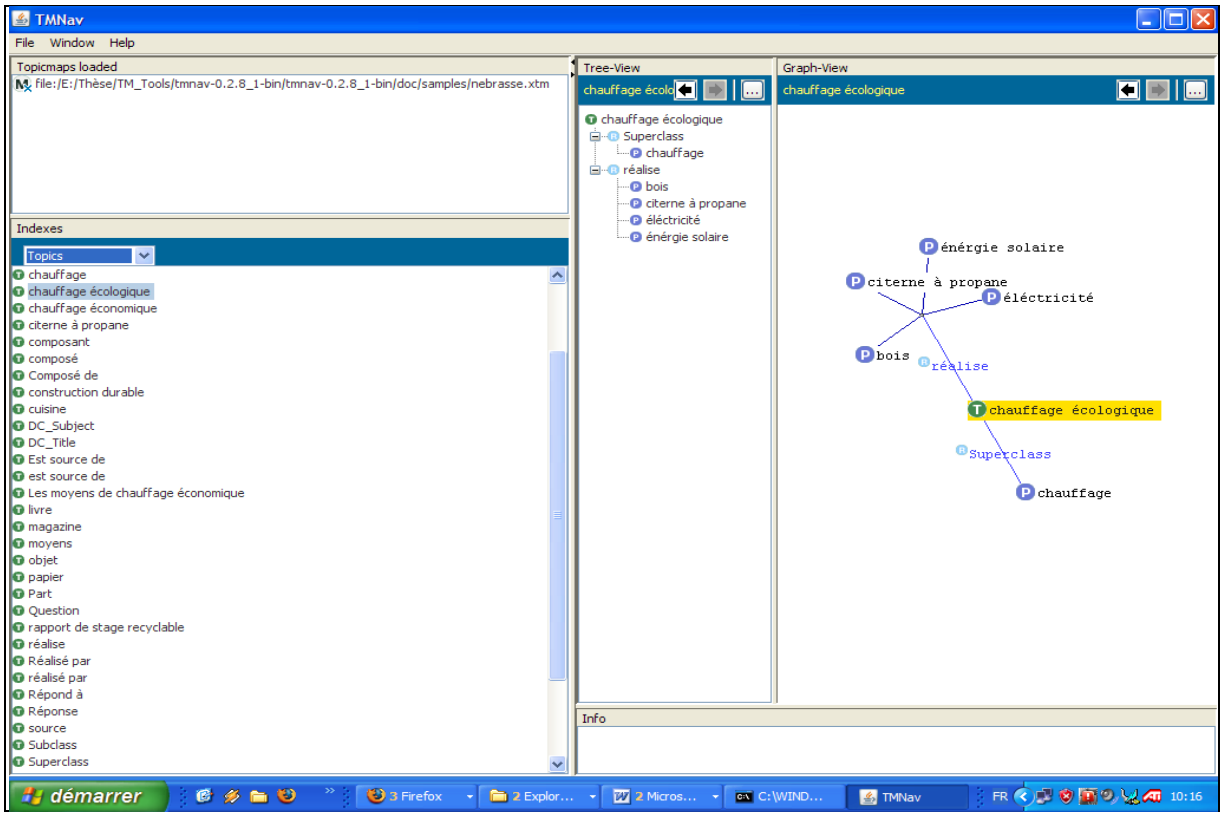


Figure 2.23 Visualisation avec l'outil TMNav

UNIVIT : L'outil UNIVIT *(Universal Interactive Visualisation Tool)* proposé par Legrand [Legrand et Soto, 1999] permet une visualisation de Topic Map sous forme de monde virtuel, permettant à l'utilisateur de se déplacer dans un environnement familier. L'objectif est de permettre à l'utilisateur de mettre en oeuvre les mêmes mécanismes de mémorisation et de localisation que dans le monde réel. Les auteurs ont choisi de représenter la Topic Map sous forme de ville virtuelle, car le milieu urbain est familier à l'utilisateur. Les Topics sont représentés par des immeubles, comme le montre la figure 2.24 ; ceux-ci sont regroupés dans des quartiers reconnaissables par la couleur du sol. Toutes les caractéristiques d'un immeuble, telles que la superficie de sa base, sa hauteur ou sa couleur, reflètent les dimensions du Topic correspondant.

Les immeubles sont répartis dans des quartiers, qui correspondent à des clusters. Deux représentations de la ville sont fournies à l'utilisateur : une représentation 3D et une carte traditionnelle en 2D, en bas à droite de la figure 2.24. La cohérence entre les deux vues est assurée en permanence.

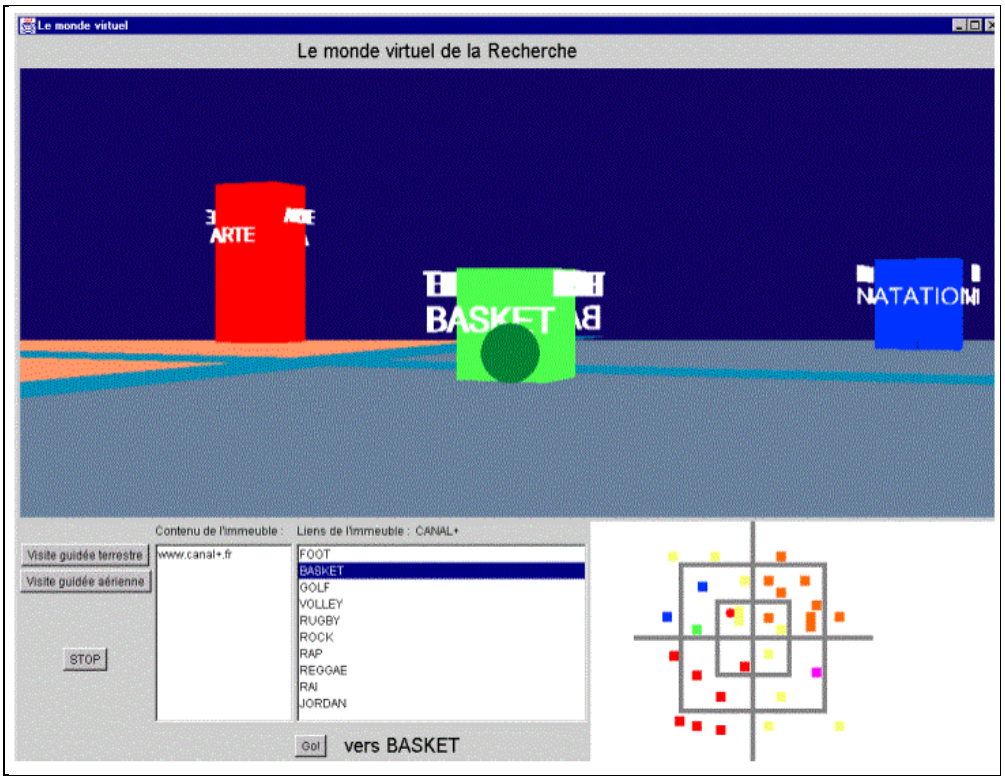


Figure 2.24 Visualisation de Topic Map sous forme de Ville virtuelle et carte 2D

En résumé, nous remarquons que les outils existants de visualisation de Topic Map permettent de représenter graphiquement les relations entre les éléments des Topic Maps. Ces visualisations facilitent la compréhension de la structure de ces systèmes. Cependant, la plupart de ces représentations n'exploitent que la sémantique exprimée explicitement dans la Topic Map. De plus, l'utilisation de l'un de ces outils nécessite la connaissance auparavant du formalisme des Topic Maps. Enfin, tous les outils de gestion de Topic Map permettent d'afficher n'importe quel fichier XTM à condition qu'il soit conforme à la DTD de la norme.
Nous proposons dans la figure 2.25 une classification des outils d'édition et de visualisation de Topic Map [Ellouze et al. 2008a].

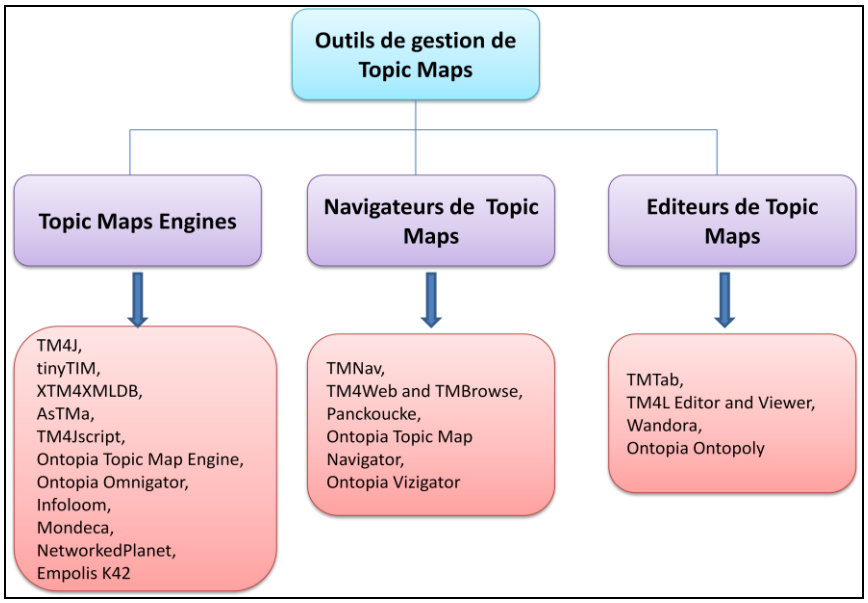


Figure 2.25 Classification des outils d'édition et de visualisation de Topic Map

2.3.7 Interrogation de Topic Maps

Des langages d'interrogation sont particulièrement dédiés au modèle des Topic Maps, l'ISO a définit la spécification TMQL[30] (*Topic Map Query Language*) [Garshol, 2005a] qui devient un standard officiel ISO 18048. TMQL n'est pas un langage mais il constitue une spécification pour les langages d'interrogation des Topic Maps. Il a pour but de simplifier le développement des applications basées sur les Topic Maps et de permettre l'interopérabilité entre les applications qui l'utilisent.

Il y a plusieurs implémentations de la spécification TMQL comme par exemple Tolog[31] [Garshol, 2001] [Garshol, 2005b], TMRQL[32] (*Topic Map Relational Query Language*) [Ahmed, 2005], AsTMa[33]. La plupart des implémentations de TMQL constituent des prototypes et il n'existe pas une implémentation stable et des outils pour les utiliser. D'après la littérature, Tolog est l'implémentation la plus stable de TMQL. Tolog est un langage destiné à l'interrogation des Topic Maps, édité par la société Ontopia (www.ontopia.net), il respecte la spécification TMQL. Tolog est inspiré de DataLog (proche de Prolog) et SQL. Certains travaux tels que [Ahmed, 2009] et [Stefanova et Risch, 2010] proposent d'adapter le langage de requête SPARQL[34] pour l'interrogation de Topic Map.

[30] http://www.isotopicmaps.org/tmql/
[31] http://www.ontopia.net/omnigator/docs/query/tutorial.html
[32] http://www.networkedplanet.com/download/TMRQL.pdf
[33] http://astma.it.bond.edu.au/querying.xsp
[34] http://www.w3.org/TR/rdf-sparql-query/

2.3.8 Comparaison des approches de construction de Topic Map

Pour évaluer les approches existantes de construction de Topic Maps, nous avons défini un ensemble de critères de comparaison, ces critères sont identifiés en analysant les différentes approches et outils de construction de Topic Maps existants dans la littérature. Comme nous l'avons déjà évoqué la majorité de ces approches s'inspirent des méthodes de construction d'ontologie puisqu'une Topic Map est composée d'une ontologie et d'autres liens sémantiques, donc, nous avons remarqué que ces approches intègrent les principales fonctionnalités incluses dans le processus de construction d'ontologies dans leurs propres méthodes.

2.3.8.1 Critères de comparaison

En se basant sur les principales fonctionnalités incluses dans le processus de construction de Topic Maps, nous avons défini un ensemble de critères qui concernent les points de vue méthodologique et pratique c'est-à-dire, la méthode adoptée et l'outil implémenté pour la construction de la Topic Map puisque certaines approches de construction sont outillées. Dans ce qui suit, nous présentons en détail ces critères.

Niveau d'automatisme : ce critère indique la manière dont l'approche extrait les éléments de la Topic Map (Topics, associations et occurrences) : automatique, semi-automatique ou manuelle.

Fusion de Topic Map : ce critère définit si l'approche ou l'outil propose des techniques pour la fusion de Topic Maps, chaque Topic Map peut être construite à partir de ressources différentes, en utilisant des techniques variées ou écrite selon une syntaxe différente.

Construction collaborative : ce critère indique si l'approche ou l'outil permet à plusieurs utilisateurs d'intervenir dans la construction.

Environnement distribué : permettre à des utilisateurs distants (connectés par un réseau) d'intervenir dans la construction de la Topic Map en introduisant de nouvelles ressources par exemple.

Aspect multilingue : Prendre en compte des ressources multilingues pour la construction de la Topic Map, c'est-à-dire un système qui permet de retrouver des documents écrits dans une langue différente de la langue de la requête.

Ressources hétérogènes : Prendre en compte des ressources hétérogènes comme entrée pour la construction de la Topic Map : documents textes, bases de données, ressources termino-ontologiques, thésaurus, ontologies, sources d'interrogation FAQ, etc.

Réutilisation de ressources existantes : si l'approche prend en entrée des ressources existantes telles que les ontologies de domaine, les thésaurus, pour créer, enrichir ou valider la Topic Map.

Visualisation de la Topic Map : si l'approche permet de visualiser la Topic Map ou si elle utilise un outil existant de visualisation, si l'interface est interactive c'est-à-dire permettant une recherche par requête ou par navigation.

Niveaux de représentation de la Topic Map : ce critère concerne l'affichage et la visualisation de la Topic Map, si l'approche permet une visualisation multi-niveaux par exemple selon les profils des utilisateurs ou selon les langues ou bien selon les différentes facettes du domaine étudié.

Gestion de l'évolution de la Topic Map : Au cours du temps, une Topic Map pourrait changer et l'utilisation de la Topic Map va certainement changer aussi, donc la Topic Map doit être modifiée pour refléter la réalité et les nouvelles utilisations. Ce critère concerne la capacité de l'approche de prendre en compte ces changements lors de l'élaboration de la Topic Map.

Importation ou exportation de Topic Map : Si l'outil permet l'exportation ou l'importation de Topic Maps externes, ce critère est lié à celui qui concerne la fusion de Topic Map.

Echange de Topic Maps ou de fragments de Topic Map entre plusieurs utilisateurs distants.

Scalabilité : ce critère indique si l'approche ou l'outil doit refaire tout le processus de construction si une nouvelle ressource est ajoutée.

Evaluation de la Topic Map : si l'approche propose des techniques d'évaluation de la qualité de la Topic Map résultante, car l'objectif d'une Topic Map est de refléter le mieux possible la réalité et permettre de répondre aux besoins de l'utilisateur.

2.3.8.2 Tableaux comparatifs

Dans les deux tableaux 2.2 et 2.3, nous proposons de comparer les différentes approches et outils de construction de Topic Maps que nous avons pu identifiés dans la littérature selon les critères déjà définis, cette évaluation a pour objectif de repérer les principales orientations dans le domaine de la construction de Topic Maps et dégager les limites des travaux existants.

Critères / Approches	Niveau d'automatisme	construction collaborative	Fusion de Topic Maps	Environnement distribué	Aspect multilingue	Ressources hétérogènes	Réutilisation de ressources existantes
[Gronmo, 2002]	- génération automatique de Topic Map à partir de métadonnées RDF	Non	Oui	Non	Non	BD relationnelles	Non
[Reynolds et Kimber, 2002]	Semi-automatique -XSLT -enrichissement manuel par les experts du domaine	Non	Oui Utilise l'utilitaire de fusion de TM4J	Non	Non	ressources XML	Ontologie du domaine
[Lin et Qin, 2002]	Semi-automatique Enrichissement de Topic Map avec des liens hiérarchiques à partir de ressources existantes	Non	Non	Non	Non	Thésaurus Ontologies Topic Maps	Non
[Legrand et Soto, 2002b]	Semi-automatique - techniques de clustering - méthodes statitiques	Non	Non	Non	Non	sites Web	Ontologie
[Folch et Habert, 2002]	Semi-automatique - techniques *NLP* - Clustering	Non	Oui	Non	Non	Textes du domaine (ressources distribuées)	Non
[Ahmed, 2003]	Semi-automatique - enrichissement de Topic Map à partir de requêtes utilisateurs - construction collaborative	Oui	Oui	Oui (application peer to peer)	Non	Topic Maps Locales	Vocabulaires partagés
[Lavik et al. 2004]	Semi-automatique BrainBank Learning : construction collaborative de Topic Map personnelles comme une stratégie pour apprendre.	Oui	Non	Non	Non	ressources Web	Non
[Zaher et al. 2006]	Semi-automatique -Co-construction collaborative de Topic Map avec different acteurs	Oui	Oui	Oui	Non	pages Web	Non
[Dicheva et Dichev, 2006]	Semi-automatique -techniques de classification en utilisant la structure d'une ontologie	Oui	Oui	Oui	Non	Ressources externes	Ontologie du domaine
[Kasler et al. 2006]	Semi-automatique -Apprentissage -Mapping à partir de métadonnées à des Topic Maps - heuristiques	Non	Non	Non	Oui (textes en Anglais et hongrois)	-Textes - métadonnées	Ontologie du domaine
[Ouziri, 2006]	Semi-automatique -Enrichissement de Topic Maps	Non	Oui	Non	Non	Ressources Web d'apprentissage	Ontologie
[Roberson et Dicheva, 2007]	-Semi-automatique -à partir d'ontologie -Crawling de sites Web en utilisant un ensemble d'heuristique	Non	Non	Non	Non	pages Web	Ontologie
[Librelotto et al. 2008]	TMBuilder : extraction automatique des éléments de Topic Map à partir de documents XML	Non	Non	Non	Non	documents XML	Non
[Jiang et al. 2009]	-Semi automatique -Fusion de Topic Maps distantes -basé sur le méta-modèle ETM	Oui	Oui	Oui	Non	Ressources pédagogiques hétérogènes et distribuées	WordNet pour le calcul de similarité sémantique entre Topics

Tableau 2.2 Comparaison des approches de construction de Topic Maps

Criteres / Approches	Visualisation de la Topic Map	Niveaux de representation	Import/Export de la Topic Map	Gestion des évolutions de la Topic Map	Echange de Topic Map ou de fragments de Topic Map	Scalabilité	Evaluation de la Topic Map	Outil associé
[Gronmo, 2002]	Outil Tmproc	Non	Non	Non	Non	Oui	Information non disponibles dans les papiers	tmproc, a TM processor written in Python
[Reynolds et Kimber, 2002]	TM4J	Non	Non	Oui	Non	Non	-Experts du domaine -Utilisateur	TM4J
[Lin et Qin, 2002]	interface Web pour édition et navigation dans des Topic Maps (Servlets Java)	Oui (plusieurs stratégies de navigation)	Oui	Non	Oui (Partage et réutilisation de Topic Maps)	Oui	Utilisateur (travaux futurs)	XML stylesheets for displaying Topic Maps
[Legrand et Soto, 2002b]	Outil de visualisation 3D	Oui	Non	Non	Oui	Oui	User	Outil de visualisation 3D développé par les auteurs
[Folch et Habert, 2002]	stylesheets XSLT	Oui (navigation selon différent points de vue)	Non	Non	Non	Non	User	-Zellig -ALCESTE Outils de text mining
[Ahmed, 2003]	TM4J	Oui	Oui	Non	Oui	Non	Utilisateur	TM4J Framework JXTA
[Lavik et al. 2004]	application Web	Oui (enseignants et étudiants)	Non	Non	Oui	Non	Utilisateur	Ontopia Knowledge Suite
[Zaher et al. 2006]	application Web	Oui	Non	Oui	Oui	Non	-User -Expert	Agorae tool
[Dicheva et Dichev, 2006]	TM4L : Environnement pour édition et navigation dans des Topic Maps comme support d'éducation	Oui (étudiants et enseignants) Représentation contextuelle	Oui (interopérabilité avec n'importe quel outil suivant le standard des Topic Maps	Oui	Oui	Oui	- Utilisateur -Expert -Tests	TM4L
[Kasler et al. 2006]	Portail Web	Non	Non	Non	Non	Non	Information non disponibles dans les papiers	framework Web Tapestry
[Ouziri, 2006]	application Web	Non	Non	Non	Non	Non	Utilisateur	Information non disponibles dans les papiers
[Roberson et Dicheva, 2007]	Editeur TM4L	Non	Non	Non	Non	Non	- Utilisateur -Tests	TM4J TM4L WebSphinx(open source Web crawler)
[Librelotto et al. 2008]	Ulisses (navigateur de TM XML)	Non	Non	Oui (Supporte des systèmes de gestion de BD)	Non	Oui	Utilisateur	TM Extractor developed by the authors
[Jiang et al. 2009]	ETM Toolkit (permet la navigation dans la Topic Map)	Oui (étudiants et enseignants)	Non	Non	Oui	En cours	-Outil testé par des étudiants	ETM Toolkit

Tableau 2.3 Comparaison des approches de construction de Topic Map (suite)

2.4 Synthèse

Dans ce chapitre, nous avons dressé un état de l'art des différents champs de recherche concernés par notre travail à savoir le Web sémantique et les modèles de représentation des connaissances dans le cadre du WS, la recherche d'information multilingue et les travaux existants pour résoudre les problèmes liés au multilinguisme dans un SRI.

Dans la troisième section de ce chapitre, nous avons élaboré un **état de l'art sur les approches et les outils de construction de Topic Maps** [Ellouze et al. 2008a], [Ellouze et al. 2008b]. Nous avons présenté les principaux axes de recherche en relation avec la construction de Topic Map en particulier les techniques d'extraction de concepts et de relations à partir de documents textes puisque certaines approches de construction de Topic Map prennent comme entrée des documents textes et utilisent ces techniques pour la création et l'enrichissement des Topic Map, nous avons également décrit brièvement les méthodes de construction d'ontologies en effet, des travaux de construction de Topic Map proposent de réutiliser ces méthodes dans leur démarche. La plupart des approches d'élaboration de Topic Map utilisent des techniques de fusion de schémas conceptuels et d'ontologies ainsi que les algorithmes de fusion offerts par le modèle des Topic Maps pour obtenir une seule Topic Map unifiée à partir de plusieurs Topic Maps par exemple dans le cadre d'intégration de ressources hétérogènes.

Une **étude comparative** entre ces approches a été également présentée à la fin du chapitre, Nous avons défini un ensemble de **critères de comparaison**, ces derniers sont identifiés suite à l'analyse des différentes approches et outils de construction de Topic Maps existants dans la littérature, nous nous sommes aussi inspirés des critères définis pour la comparaison des méthodes de construction d'ontologies puisque dans la littérature de nombreuses études ont été faites dans ce domaine. Les critères que nous avons définis concerne le niveau méthodologique et technique (c'est-à-dire pratique) de l'approche proposée.

Cette étude comparative des approches de construction de Topic Map nous a permis de dégager les constations suivantes :

- Aucune des approches, excepté celle de Kasler, ne permet de traiter **un contenu multilingue** ;
- Aucune des approches existantes n'exploitent **plusieurs sources d'informations** pour la construction d'une Topic Map ;
- Bien que les Topic Maps soient, par définition, orientées utilisation et recherche d'information, peu d'entre elles prennent en compte **les requêtes des utilisateurs** ;
- La plupart des approches existantes ne proposent pas des techniques pour **l'évaluation de la qualité** de la Topic Map résultante ;
- Les travaux existants ne permettent pas **un guidage méthodologique** de l'utilisateur lors du processus de construction.

Notre approche

C'est pourquoi, nous proposons **une approche incrémentale et évolutive de construction de Topic Map multilingues** qui, en plus de la base de documents disponibles dans **différentes langues**, elle conjugue l'utilisation de trois sources d'information : un thésaurus du domaine, deux ontologies générales (WordNet et WOLF, WordNet libre du français) ainsi que l'ensemble de toutes les **sources d'interrogations** possibles telles que les questions relatives aux documents sources qu'un expert du domaine ou un utilisateur quelconque peut poser, les foires aux questions (FAQ), les traces des discussions téléphoniques et des consultations directes avec les travailleurs du domaine.

Notre approche est fondée sur deux méta-modèles que nous définissions : un **méta-modèle de référentiel** de documents segmentés thématiquement et indexés sémantiquement et **un méta-modèle de Topic Map**. Nous définissons le méta-modèle de Topic Map comme une extension du modèle des Topic Maps existant auquel nous rajoutons : (i) la notion de **méta-propriétés de Topics** pour la mesure de la qualité de la Topic Map ; (ii) la possibilité **d'indexer un Topic par un segment** de document et ce grâce à la segmentation thématique des documents sources et (iii) la **classification de liens dans une Topic Map**, en plus des liens d'occurrences, nous définissons deux autres types de liens : les liens d'usage et les liens ontologiques.

La Topic Map finale a pour objectif de faciliter la recherche d'information pour les utilisateurs, cette recherche peut être une **recherche classique par requête** ou une **recherche par navigation**.

Notre travail fait intervenir deux étapes principales :

- La construction d'un référentiel de documents segmentés thématiquement et indexés sémantiquement en utilisant des algorithmes appropriés ;
- La construction de la Topic Map multilingue, enrichie et annotée à partir du référentiel de documents, d'un thésaurus du domaine, de deux ontologies générales, et d'un ensemble de scénarios d'usage (des questions/réponses extraites à partir de sources d'interrogation telles que les FAQ).

Dans le prochain chapitre, nous allons exposer notre solution : une approche générale de construction de Topic Map multilingue, nous allons décrire les deux méta-modèles proposés et les différents types de recherche offerts par la Topic Map.

CHAPITRE 3

Approche générale et méta-modèles

3.1 Problématique et objectifs

L'accès au contenu sémantique des documents issus du Web ou de grands corpus nécessite une phase d'enrichissement sémantique de ces documents. Il s'agit, à ce niveau, de réaliser une médiation sémantique entre la sémantique des producteurs (rédacteurs) des documents, la sémantique des indexeurs des documents et la sémantique des besoins en information des utilisateurs. Non seulement ces trois groupes d'acteurs peuvent ne pas se connaître mais en plus ils n'ont pas forcément les mêmes centres d'intérêts, les mêmes habitudes, la même culture et ne partagent pas de ce fait le même vocabulaire (figure 3.1).

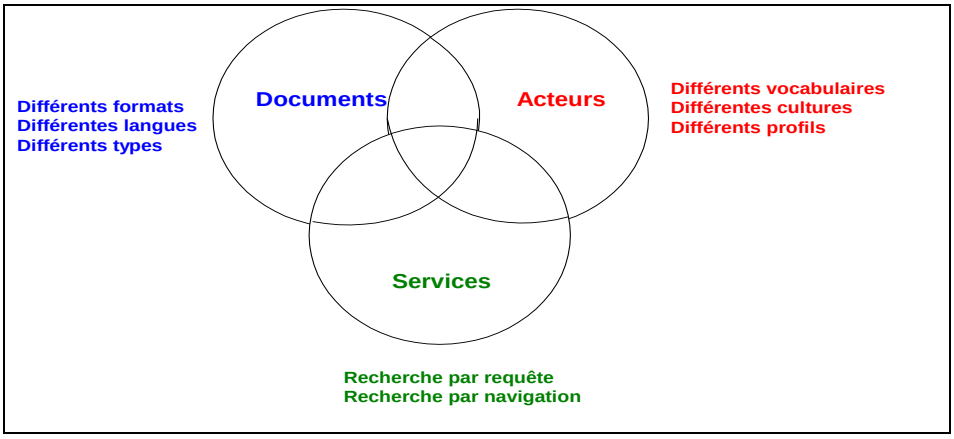


Figure 3.1 Problématique générale de la recherche d'information

Lors d'une session de recherche, l'utilisateur peut avoir recours, pour exprimer son besoin, à des termes (ou concepts) qui ne sont pas forcément les mêmes que ceux utilisés par l'indexeur ou par les auteurs. L'ambiguïté rencontrée lors de l'expression des besoins est la source principale des problèmes de bruit et silence souvent rencontrés dans les systèmes de recherche d'information. L'ambiguïté peut se faire sentir principalement à deux niveaux : l'ambiguïté du besoin de l'utilisateur par rapport à l'indexeur et l'ambiguïté de la sémantique de l'auteur par rapport à l'indexeur (figure 3.2).

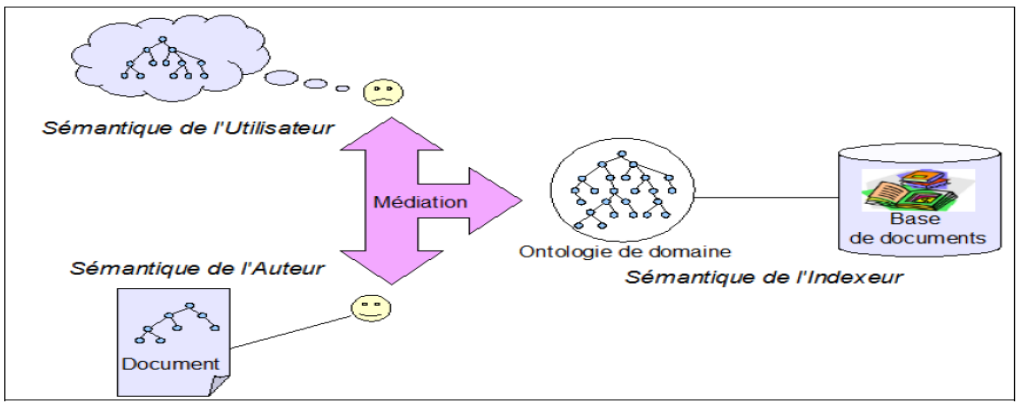


Figure 3.2 Problématique de la médiation sémantique

Nous distinguons deux types de recherche :

- **La recherche exacte** : l'utilisateur connaît exactement ce qu'il cherche, connaît le langage d'indexation utilisé, et formule sa requête en adéquation avec le format d'indexation, exemple : chercher un document sachant son titre ;
- **La recherche floue** : l'utilisateur a une idée de ce qu'il cherche sans connaître le format du langage d'indexation avec une requête à formulation variable. Pour aider l'utilisateur dans le cas de recherche floue, plusieurs techniques ont été proposées, ces techniques ont pour objectifs d'étendre les requêtes et ceci par diverses moyens : l'expansion (grâce à un thésaurus qui fournit des termes synonymes ou hyperonymes), la reformulation (grâce à des groupes de termes associées souvent rencontrés dans le même contexte) et le « relevance feedback » (par des mots clés issus des meilleurs documents résultant d'une première requête).

Dans le cas de **systèmes de recherche par exploration / navigation**, la collection de documents est accessible en naviguant de lien en lien. Les liens peuvent se situer entre les documents (système hypertexte classique) et/ou dans une structure décrivant l'organisation des informations (carte conceptuelle, index, etc.). La variété de ces outils est assez large, allant d'index peu structurés à des index fortement structurés comme divers types de systèmes de navigation. La structuration des informations guide l'utilisateur tout le long de sa recherche (choix d'un lien ou d'un autre). L'utilisation d'un tel outil ne nécessite pas de connaissances sur le fonctionnement informatique.

En revanche, la navigation conduit souvent à une désorientation des utilisateurs caractérisée par la consultation de nombreux documents et une baisse des performances dans le temps. L'utilisateur n'obtient pas une vue d'ensemble de la structure des informations et éprouve des difficultés à établir un but et sa planification.

Nous pouvons à cette étape retenir que, les systèmes par requête requièrent de l'utilisateur une bonne connaissance des modalités de fonctionnement du système et en particulier de la manière de formuler les requêtes pour parvenir à des résultats pertinents. De leur coté, les systèmes par navigation sont susceptibles d'entraîner une forme de désorientation des utilisateurs. Pour pallier ce problème, il est bénéfique de proposer des aides à la navigation sous la forme d'une structuration des informations utiles à la navigation.

C'est à ce niveau que se situent nos travaux de recherche, notre objectif est **de proposer une approche de recherche intelligente d'information** permettant, en plus de la recherche par requête, une recherche par navigation dans un contenu textuel multilingue.

Pour cela, notre approche sera fondée sur deux méta-modèles : le premier est basé sur le modèle des Topic Map et le deuxième est représenté par un référentiel de documents multilingues segmentés thématiquement et indexés sémantiquement. Notre approche a deux objectifs : d'une part, définir un modèle de représentation des connaissances à partir de l'analyse des documents, des acteurs et des services et d'autre part, réaliser un appariement entre le contenu des documents et les besoins des acteurs, en leur offrant, via des services, les documents ou les segments de documents les plus pertinents par rapport à leurs besoins et préférences.

Pour la construction et l'enrichissement de ces deux méta-modèles, en plus du contenu textuel multilingue, nous prenons comme entrées un thésaurus du domaine contenant les concepts du domaine et les relations normalisées telles que les relations hiérarchiques et les relations de synonymie, deux ontologies générales représentant la terminologie du langage commun et des scénarios d'usage construits à partir de FAQ.

Nous avons choisi ces deux méta-modèles parce que nous visons, à travers notre approche, à permettre trois modes de recherche : une recherche par navigation à travers la Topic Map, une recherche classique par requête et pour représenter l'usage dans notre Topic Map, nous proposons également un troisième mode de recherche basé sur un ensemble de scénarios de questions préparés à partir de FAQ dont on prévoit les réponses.

Parmi les raisons qui nous ont motivés à choisir le modèle des Topic Map est que ce dernier intègre cette notion d'usage, en effet, par rapport aux ontologies définies comme une conceptualisation formelle du réel, permettant de décrire le réel indépendamment de l'usage, les Topic Maps ajoutent la notion d'utilisation, elles décrivent le réel en prenant en compte son usage par exemple supposons qu'on dispose d'une ontologie sur les tulipes et les roses, si un utilisateur demande des documents sur « tulipe », le système va étendre sa requête et lui retourner des documents sur les « roses » alors que dans une Topic Map, on peut rajouter des informations sur l'usage qu'on en fait, par exemple, les fleurs servent à faire des cadeaux et le chocolat sert aussi à faire des cadeaux et donc de cette manière, nous donnons à l'utilisateur des idées de cadeaux.

Par ailleurs, nous avons choisi les Topic Maps parce qu'elles constituent un modèle de représentation des connaissances orienté navigation, c'est une carte sémantique permettant, en

plus de la recherche par requête utilisant un langage dédié aux Topic Maps (ou un autre langage tel que SPARQL [Ahmed, 2009]), une recherche exploratoire par navigation. Il a été conçu pour l'organisation d'un ensemble de documents grâce à la notion de sujets (Topics) et aux associations entre ces sujets qu'elle représente afin de faciliter la navigation dans ces documents. Si l'utilisateur ne sait pas ce qu'il cherche, il peut naviguer dans la Topic Map et découvrir des informations susceptibles de l'intéresser et qu'il ne pensait pas avoir.

De plus, par rapport à nos objectifs de construire la Topic Map à partir de documents multilingues, le modèle des Topic Maps propose la notion de scope ou contexte que nous explorerons pour la gestion du multilinguisme. Il propose également la notion de facette, un ensemble d'attributs-valeurs reliés au lien occurrence pour caractériser la ressource en question, nous explorerons aussi cette notion pour implémenter l'aspect multilingue.

Tout au long de ce chapitre nous détaillons toutes nos contributions tout en les motivant. Dans la section suivante, nous décrivons notre approche générale.

3.2 Notre approche générale

Notre approche a pour objectif d'organiser, d'indexer et d'annoter des documents sources multilingues afin d'améliorer l'accès et la recherche dans ces documents et ce grâce au modèle des Topic Maps.

Notre approche est basée sur un processus automatisé, elle prend en entrée un contenu textuel multilingue et des outils terminologiques pour modéliser le domaine. Pour l'occurrence, nous utilisons comme ressource terminologique un thésaurus bilingue de domaine et deux ontologies générales représentant la terminologie du langage commun (WordNet et WOLF). Pour la prise en compte de l'usage dans la Topic Map, nous prenons aussi comme entrée un ensemble de scénarios d'usage composé de questions types avec leurs réponses, ces scénarios sont préparés à partir de l'analyse de sources d'interrogation possibles relatives aux documents sources telles que les FAQ.

Notre approche se compose principalement de deux modules correspondant respectivement à la construction du référentiel et à la construction de la Topic Map.

Le premier module a pour objectif de construire un référentiel de documents annotés avec des méta-informations descriptives et indexés sémantiquement et thématiquement, ce module contient les étapes suivantes :

- La **première** étape concerne le prétraitement des documents sources ;

- la **deuxième** étape consiste à segmenter thématiquement les documents du corpus en utilisant l'algorithme TextTiling comme segmenteur thématique. Nous obtenons pour chaque document, la liste des segments thématiques qui le compose, chaque segment représente une thématique abordée dans le document, ces thématiques seront représentées dans la Topic Map sous forme de Topics thèmes. Nous expliquons davantage ce point dans la deuxième section du chapitre suivant ;
- La **troisième** étape consiste à extraire les termes et les concepts les plus représentatifs indexant sémantiquement les divers documents et leurs segments en utilisant *Latent Semantic Indexing Model*. Les termes et les concepts sont pondérés par leurs degrés de pertinence.

Suite à ces trois étapes, nous générons un référentiel de documents indexés thématiquement et sémantiquement en associant à chaque document, la liste de ses segments qui représentent les thématiques abordés dans le document et la liste des termes et des concepts représentatifs de son contenu, à chaque terme et à chaque concept (qui sera représenté par un Topic dans la Topic Map) est associé des mesures qui sont *tof*×*idf* (*topic frequency-inverse document frequency*) pour les Topics et *tf*×*idf* (*term frequency-inverse document frequency*) pour les termes. A ces index thématiques et sémantiques de chaque document seront ajoutées des métadonnées descriptives (titre, taille, langue, pertinence du document, URL, organisation, date, etc). Ce référentiel sert à la construction et à l'annotation de la Topic Map. Nous décrivons en détail les deux mesures *tof*×*idf* et *tf*×*idf* dans le chapitre suivant.

Le deuxième module a pour but de construire et enrichir la Topic Map multilingue à partir du référentiel de documents segmentés thématiquement et indexés sémantiquement, d'un thésaurus du domaine, de deux ontologies générales et d'un ensemble de scénarios d'usage. Notre approche de construction de la Topic Map est **incrémentale et évolutive**, pour chaque document du référentiel, nous commençons par extraire une liste de Topics et de liens ontologiques et associatifs entre ces Topics en utilisant le résultat de l'indexation avec LSI, des techniques d'analyse linguistique de documents textuels et en se basant sur le thésaurus du domaine, ensuite la Topic Map résultante est enrichie à partir des deux ontologies générales (WordNet pour l'anglais et WOLF pour le français) et des questions relatives aux documents sources. Cette Topic Map enrichie, associée au document en question est ensuite fusionnée, selon un processus semi-automatique et avec l'aide d'experts du domaine, avec la

Topic Map obtenue à partir du document précédent, ce processus est répété jusqu'à ce que nous terminions tous les documents disponibles dans le référentiel.

A la fin du processus, nous obtenons une Topic Map globale, enrichie et annotée à partir du référentiel de documents en associant à chaque Topic la liste des documents et leurs segments triés par degré de pertinence. Dans notre cas, nous avons choisi dans un premier temps de traiter les deux langues, français et anglais, et en perspectives d'autres langues seront prises en compte telles que l'arabe.

Le référentiel annoté est destiné à être la cible des procédures de recherche en mode requête, alors que la Topic Map enrichie et annotée doit servir au mode de recherche par navigation. Le modèle conceptuel de l'approche proposée est illustré par la figure 3.3 :

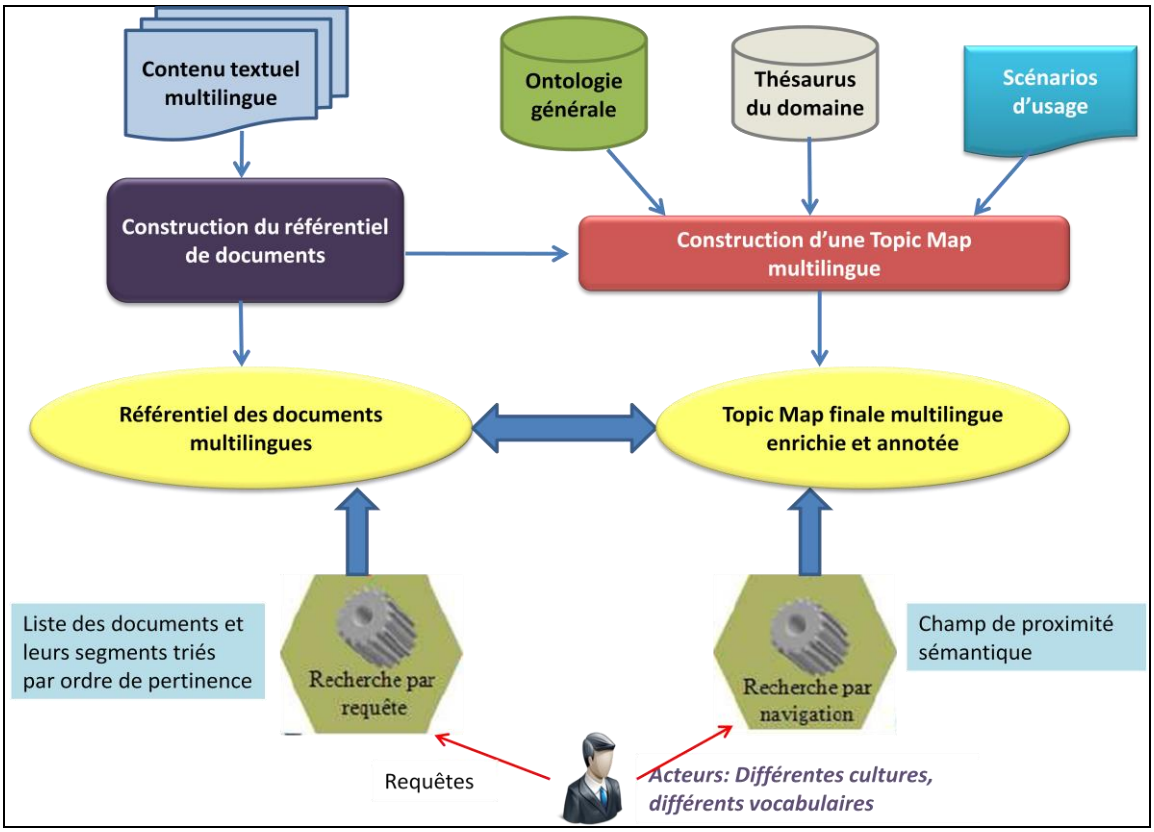


Figure 3.3 Notre approche générale

Nous définissons deux méta-modèles permettant de décrire respectivement le référentiel de documents et la Topic Map, nous détaillons ces deux méta-modèles dans les prochaines sections.

Nous commençons par présenter notre méta-modèle de Topic Map et ses différentes composantes, nous détaillons en particulier les extensions que nous avons intégrées au modèle de Topic Map existant pour répondre à nos objectifs.

3.3 Méta-modèles proposés

3.3.1 État de l'art sur les méta-modèles de Topic Map existants

Le modèle Topic Map a été formalisé en norme ISO 13250 en 2000. Il a été inclus dans l'ODM (*Ontology Meta-Model Definition*) par l'OMG[35] dans l'objectif de fournir un modèle standard TM-UML favorisant l'applicabilité des concepts du MDA (*Model Driven Architecture*) à l'ingénierie des Topic Maps. La figure 3.4 montre un extrait du méta-modèle des Topic Maps proposé dans l'ODM.

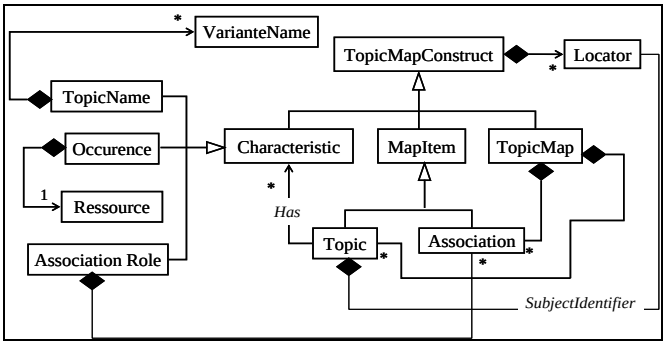


Figure 3.4 Extrait du méta-modèle des Topic Maps selon la spécification de l'OMG

Peu de travaux dans la littérature proposent d'étendre le modèle de Topic Map existant, ces travaux définissent des méta-modèles de Topic Map selon leurs objectifs de recherche et l'application sur laquelle ils travaillent. Parmi les méta-modèles de Topic Map proposés, nous citons : le méta-modèle Hypertorpic de [Zacklad et al. 2003a], le méta-modèle TMGrid de [Korthaus et al. 2006] et le méta-modèle ETM de [Lu et al. 2008].

Le méta-modèle Hypertopic [Zacklad et al. 2003a]

Hypertopic[36] est proposé par Zacklad en 2003, il a été ensuite enrichi et amélioré en 2007 par [Zaher et al. 2006] au sein du laboratoire Tech-CICO dans le cadre de leurs travaux de recherche sur le Web socio sémantique (défini comme l'expression d'une sémantique explicite d'objets métiers dans un domaine, à l'intention de divers rôles d'acteurs). Le modèle HyperTopic un méta-modèle générique inspiré des Topic Maps pour la réalisation d'applications permettant de structurer et de rechercher un ensemble de ressources présentant des similarités « fonctionnelles » pour des utilisateurs partageant des préoccupations communes.

[35]Object Management Group, Ontology Definition Metamodel Request For Proposal, OMG Document: ad/2003-03-40, http://www.omg.org/cgi-bin/doc?ad/2003-03-40
[36]http://www.hypertopic.org/

Ce modèle intègre, en plus des Topics, scopes (ou thèmes), associations et ressources qui reprennent les concepts normalisés des Topic Maps, les notions d'Entité et de Point de vue (figure 3.5). Une **Entité** représente une structure générique composée d'un certain nombre de descripteurs permettant la caractérisation d'un certain nombre de ressources. Un **Point de vue** correspond à une famille de caractéristiques de l'entité regroupés et hiérarchisés en plusieurs niveaux en fonction d'un angle de vision pour un acteur ou un ensemble d'acteurs.

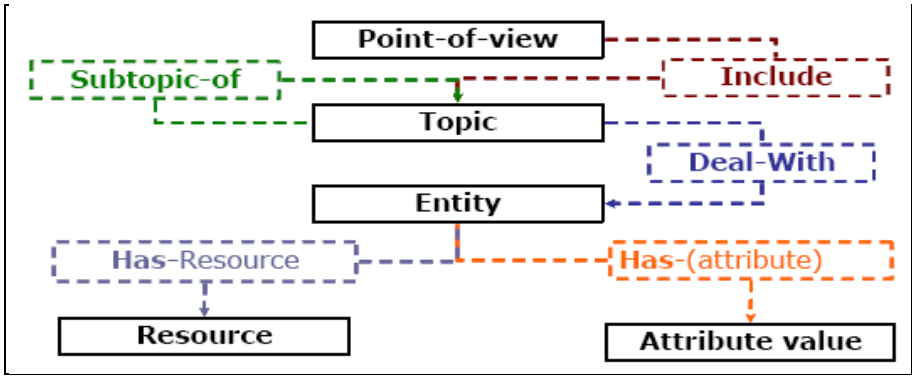


Figure 3.5 Le méta-modèle HyperTopic

Le méta-modèle TMGrid [Korthaus et al. 2006]

Dans le cadre d'applications distribuées, certains travaux proposent d'étendre le modèle des Topic Maps pour supporter la construction de Topic Map dans un environnement distribué, par exemple, nous citons le modèle **TMGrid** proposé par Korthaus [Korthaus et al. 2006], c'est un réseau où chaque nœud représente une connaissance sous forme d'une ou plusieurs Topic Maps , ces Topic Maps sont interrogeables par les autres utilisateurs situés au niveau des autres nœuds du réseau par l'intermédiaire d'un protocole spécifié, l'objectif de ces travaux est de produire une vision unifiée des connaissances contenues dans des Topic Maps distribuées comme si l'utilisateur travaillait sur une seule Topic Map. Cette idée a été inspirée de la notion du *Grid Computing*. Le méta-modèle TMGrid est illustré par la figure 3.6.

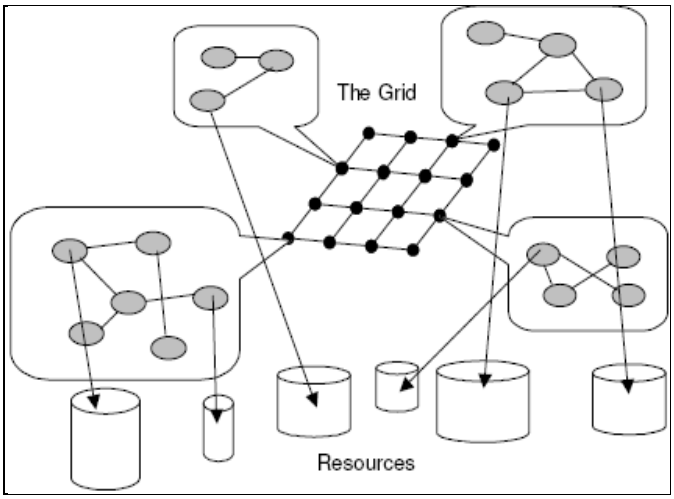


Figure 3.6 Le méta-modèle TMGrid

Le méta-modèle ETM [Lu et al. 2008]

Un autre travail récent proposé par Huimin Lu et ses collègues [Lu et al. 2008] présente une extension du modèle des Topic Maps, nommé *Extended Topic Map* (ETM). Ce modèle a été défini dans le domaine du e-learning, un outil a été implémenté en se basant sur ce modèle pour permettre à des étudiants de naviguer dans un ensemble de ressources e-learning et à des enseignants de partager des connaissances du domaine avec ces étudiants.

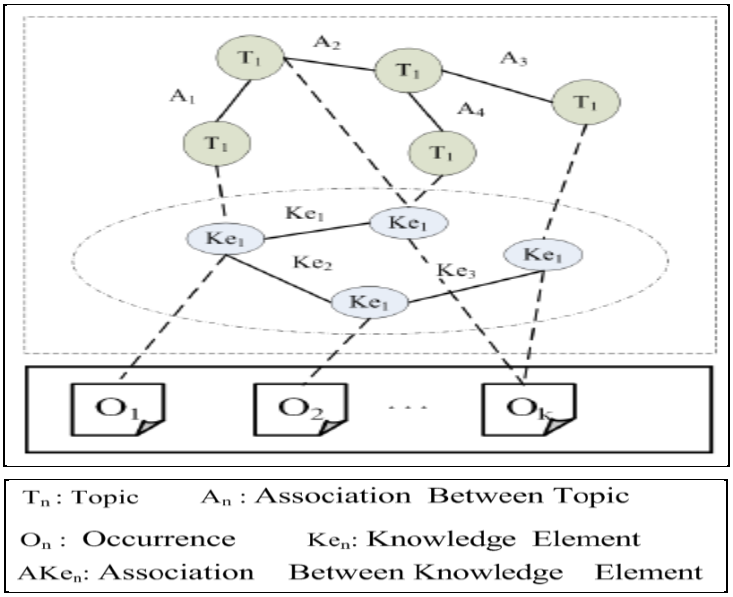


Figure 3.7 Le modèle ETM (*Extented Topic Map*)

Dans le modèle ETM, les auteurs proposent d'ajouter un autre niveau de connaissance dans la Topic Map définie par l'ISO, contenant des éléments qu'ils appellent « *Knowledge elements* », ETM définit alors de nouvelles associations entre les Topics (qui représentent les

éléments de connaissances) et les occurrences et des liens reliant les éléments de connaissances entre eux (figure 3.7).

Notons que dans la littérature, la plupart des chercheurs dans le domaine des Topic Maps proposent d'utiliser et d'appliquer le modèle tel qu'il a été défini par l'ISO pour réaliser leurs travaux, peu d'entre eux, à l'exception de ceux travaillant dans un environnement distribué tels que les applications de e-learning, se sont intéressés à la définition de méta-modèles basés sur les Topic Maps.

Dans le contexte de notre approche de recherche intelligente, nous proposons d'étendre le modèle des Topic Maps et définir un nouveau méta-modèle qui prend en compte les objectifs et les particularités de notre travail. La prochaine section est consacrée à la description de notre méta-modèle de Topic Map.

3.3.2 Notre méta-modèle de Topic Maps

Notre méta-modèle de Topic Map (figure 3.8) est une extension du modèle de Topic Map déjà défini, il présente l'originalité d'intégrer trois nouvelles notions qui n'étaient pas explicitement présentes dans le modèle des Topic Maps [Ellouze et al. 2009a] [Ellouze et al. 2009c], les principes de base de notre méta-modèle se résument comme suit :

- **Préciser et enrichir la sémantique des liens d'une Topic Map**. En plus des liens d'occurrences déjà existants dans le modèle des Topic Maps, nous séparons la définition des associations eu deux types : les liens ontologiques et les liens d'usage ;
- **Définir des méta-propriétés associées aux Topics**. Une méta-propriété qui renseigne sur l'importante du Topic au cours du temps et une méta-propriété qui indique le niveau auquel appartient un Topic. En effet, notre idée consiste à classifier et organiser la Topic Map en niveaux : en plus des deux niveaux (niveau Topic et niveau ressources) déjà défini dans le modèle, nous séparons la couche Topic en deux sous couches, la première du niveau le plus haut contenant les Topics thèmes et les Topics questions et la deuxième sous couche contenant les Topics concepts du domaine, des instances de Topics, des sous Topics, des synonymes de Topics, ou des synonymes d'instances de Topics ou de sous Topics ;
- **Indexer un Topic par un fragment de document** au lieu du document en entier pour mieux cibler les réponses aux demandes des utilisateurs et ce grâce au

méta-modèle du référentiel que nous définissons dans la section suivante et la segmentation thématique des documents sources.

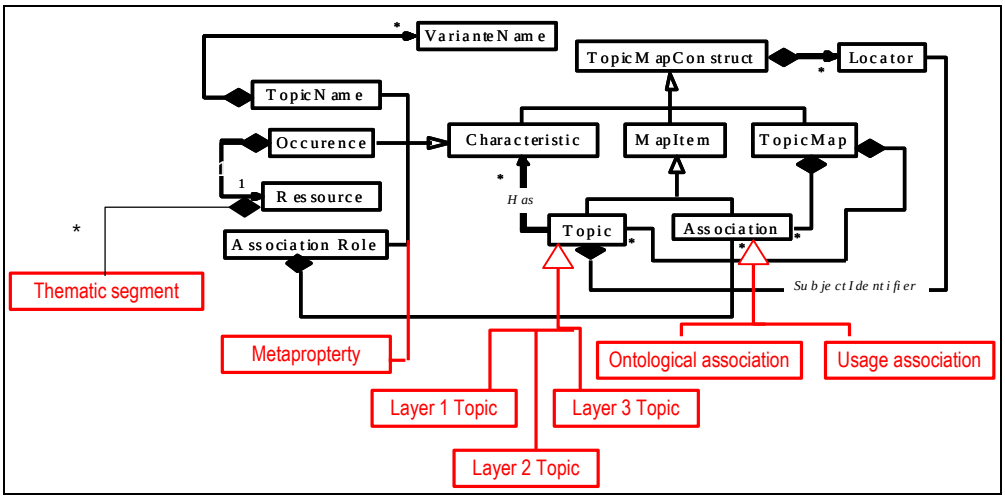


Figure 3.8 Notre méta-modèle de Topic Maps

3.3.2.1 Classification des liens dans la Topic Map

Le modèle des Topic Maps ne permet pas de faire la différence entre les liens dans une Topic Map, ils sont tous représentés comme des associations (« est un », « réalisé par », « a un effet sur », « partie de »…). Tel qu'il est mentionné dans [Pepper, 2008], il n'y a pas de limitation dans la définition des liens dans une Topic Map, ils sont spécifiés par le créateur de la Topic Map selon les besoins en information, les connaissances véhiculées par la Topic Map ainsi que l'application à laquelle elle est destinée.

Pour répondre à nos objectifs liés à l'organisation des documents pour en améliorer la recherche, nous proposons de classifier les liens dans la Topic Map.

Dans le cas général, la classification de liens sémantiques entre concepts ou termes a fait l'objet de plusieurs résultats de recherche. A titre d'exemple, nous citons la classification proposée par le ANSI (*American National Standards Institute*) dans le ANSI/NISO Z39.19-2005. Cette dernière considère trois catégories de liens : (1) les liens « d'équivalence » telles que la synonymie et la quasi-synonymie ; (2) les liens hiérarchiques telles que la spécialisation et la généralisation et (3) les autres liens (nommés «associatifs ») tels que le lien « cause/effet » et le lien « action/cible ». Comme autre exemple de classification, nous citons l'ontologie proposée dans [Storey et Sandeep, 2004] pour la catégorisation des relations représentées par les verbes dans les phrases d'un texte.

Dans notre méta-modèle de Topic Map, nous catégorisons les liens entre Topics en deux catégories [Ellouze et al. 2009a] :

a) **les liens ontologiques et structurels** qui regroupent les liens de spécialisation, les liens de composition ainsi que les liens associatifs, tels que ceux définis dans le standard (ANSI/NISO Z39.19-2005), que nous identifions suite à l'analyse des documents à organiser ;

b) **les liens d'usage** définis comme des hyper liens de type « répond à » (hyper lien questions/réponses) entre la question représentée comme un Topic et les réponses associées, c'est-à-dire les Topics référençant les documents qui permettent de répondre à la question. Nous proposons dans ce contexte de relier la question à chacun des mots clés la constituant via un hyper lien de type « est composé de » qui lui aussi est considéré comme un lien d'usage.

3.3.2.2 Organisation de la Topic Map en niveaux

Dans le modèle des Topic Maps, tel qu'il a été conçu, tout peut être représenté comme un Topic (un concept du domaine, un terme, une instance d'un concept). Ces Topics sont très souvent mélangés et mal organisés, le modèle des Topic Maps propose d'organiser la Topic Map en deux couches, une couche Topic qui contient les connaissances du domaine et une couche ressources, des travaux tels que [Dicheva et Dichev, 2006] proposent en plus de la couche ressource, de différencier le niveau Topic en deux sous niveaux, les Topics représentant les concepts du domaine et les Topics représentant les instances.

Dans notre méta-modèle de Topic Map, nous proposons d'organiser la Topic Map en trois niveaux :

- **Le niveau 1** contient les Topics thèmes représentant les thèmes du domaine et les Topics questions représentant les questions types et éventuellement les réponses adéquates identifiées à partir de l'analyse d'un ensemble de sources d'interrogation, pour notre cas, nous avons choisi les FAQ ;
- **Le niveau 2** englobe les Topics qui représentent les concepts du domaine, les Topics instances de ces concepts, les sous Topics ainsi que les Topics synonymes d'un concept ou synonyme d'une instance d'un concept ;
- **Le niveau 3** contient les ressources c'est-à-dire l'ensemble des documents textuels disponibles en différentes langues ainsi que les sources d'interrogation représentées par les FAQ.

La figure 3.9 présente l'architecture générale de la Topic Map selon notre méta-modèle.

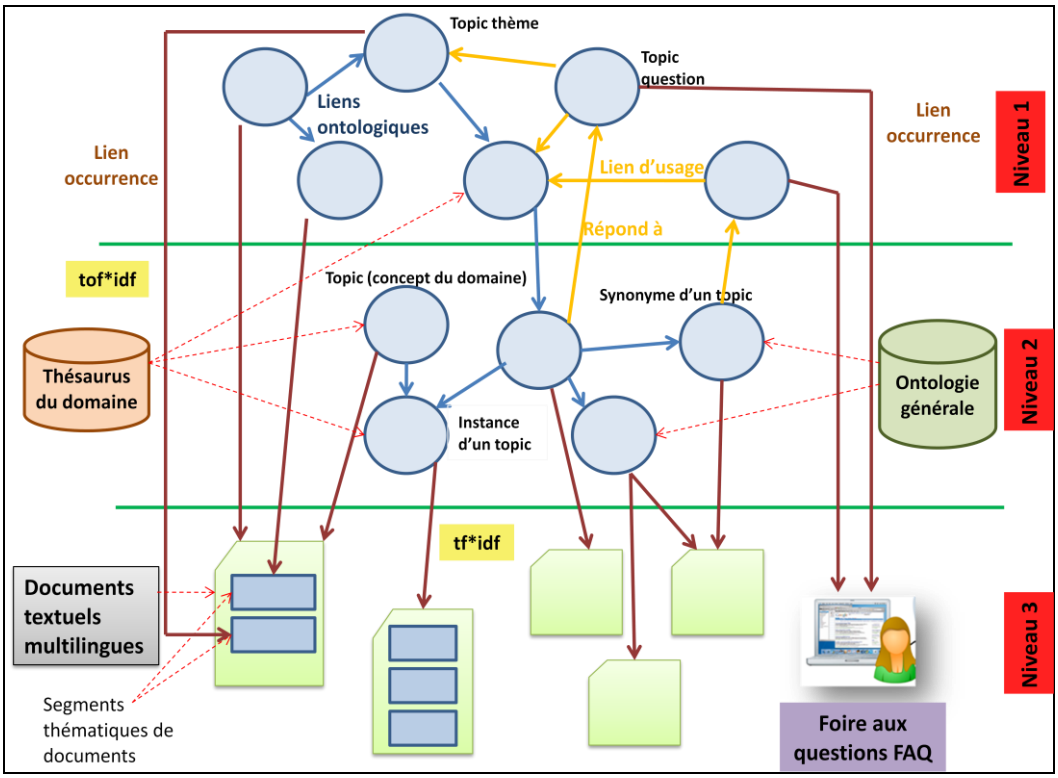


Figure 3.9 Architecture générale de la Topic Map selon notre méta-modèle

3.3.2.3 Les métadonnées d'un Topic

Pour l'implémention des différentes couches (*layers*) de la Topic Map, nous proposons de définir une **première métadonnée** attribuée aux Topics qui renseigne sur **le niveau auquel appartient le Topic** en question, c'est-à-dire si le Topic est un thème ou une question, alors ce dernier appartient au premier niveau sinon, si le Topic est un concept du domaine, une instance ou un synonyme d'un Topic alors ce dernier appartient au deuxième niveau de la Topic Map.

Nous définissions également dans notre méta-modèle une **deuxième métadonnée** qui renseigne sur **le degré de popularité ou l'usage du Topic** c'est-à-dire si le Topic est peu demandé, moyennement demandé ou trop demandé par les utilisateurs, cette métadonnée nous servira particulièrement dans le processus d'élagage évolutif de la Topic Map et entre autres pour la mesure de la qualité de la Topic Map, elle est définie comme la note du Topic qui peut augmenter ou diminuer au cours du temps. Nous explicitons cette idée dans le chapitre 5 consacré à la qualité de la Topic Map.

Actuellement, dans nos travaux de recherche, nous avons défini deux types de métadonnées, dans nos travaux futurs, nous étudierons la possibilité de généraliser la notion de métadonnée aux autres concepts tels que le concept d'association.

3.3.2.4 Indexation de Topic par un fragment de document

Le modèle des Topic Maps tel qu'il a été défini ne permet pas d'indexer un Topic par un fragment de document. Or, dans le cas de document de grande taille, lorsque l'utilisateur accède à ce document, il doit alors parcourir la totalité du document pour retrouver le sujet ou le thème qu'il recherche. Pour cela, nous proposons un méta-modèle pour la représentation du contenu textuel multilingue sous la forme d'un référentiel de documents pour compléter le méta-modèle de Topic Map. En effet, les documents du référentiel sont segmentés thématiquement et puis indexés par une liste de termes et de concepts représentatifs de leur contenu. Nous aurons alors, en plus des documents, leurs segments thématiques et grâce au méta-modèle du référentiel, un Topic peut être relié à un fragment du document au lieu du document en entier.

En résumé, comme le montre la figure 3.9, les principaux éléments de notre Topic Map fondée sur le méta-modèle de Topic Map que nous avons défini sont :

- Les nœuds qui représentent les Topics (thèmes, questions, concepts du domaine, instances de Topics, sous Topics), les documents et les segments de documents ;
- Les arcs typés et pondérés qui représentent les liens ou associations, ces liens sont de différents types :
    - **Des liens ontologiques** qui regroupent : (i) Des liens entre deux Topics, ces liens peuvent être de type « est un », « partie de » ou un lien associatif tel que le lien « réalisé par » entre « chauffage solaire » et « capteur solaire » identifié suite à l'analyse linguistique des documents sources ; (ii) Des liens entre un Topic et ses instances ; et des liens de similarité entre les Topics ;
    - **Des liens d'usage** qui regroupent : L'hyper lien « répond à » entre le Topic qui représente la question et les Topics réponses à cette question, extraites à partir des FAQ et l'hyper lien « composé de » entre le Topic qui représente une question et les Topics qui représentent les mots clés qui la compose ;
    - **Des liens d'occurrences** qui regroupent : (i) Des liens reliant un Topic aux documents ou aux segments de documents qu'il indexe. Ces liens sont étiquetés par le degré de pertinence **(*tof*×*idf*)** du Topic dans le document (ou segment de document) ; (ii) Des liens entre termes et documents ou

segments de documents. Ces liens sont étiquetés par le degré de pertinence **(*tf*×*idf*)** de chaque terme dans le document (ou segment de document).

Dans le chapitre 4, nous expliquons en détail les démarches et les techniques que nous proposons pour implémenter les différents composants de notre méta-modèle en particulier les mesures (*tof*×*idf*) et (*tf*×*idf*). Dans la section suivante, nous décrivons notre méta-modèle de référentiel de documents.

3.3.3 Notre méta-modèle du référentiel de documents

Le référentiel de documents est construit à partir du contenu textuel multilingue en se basant sur trois étapes : (a) une étape de prétraitement (élimination de mots vides, lemmatisation) ; (b) une étape de segmentation thématique des documents ; et (c) une étape d'indexation sémantique pour l'extraction des termes et des concepts représentatifs du contenu des documents et de leurs segments. Avant de détailler ces trois étapes qui feront l'objet de la deuxième section du chapitre suivant, nous nous concentrons, dans ce paragraphe, sur la description de notre méta-modèle de référentiel de documents (figure 3.10).

Par référentiel sémantique, nous désignons un fichier direct qui associe aux Topics, les documents et les segments qui les contiennent. En d'autres termes, le référentiel contient les documents avec leurs fiches descriptives c'est-à-dire pour chaque document :

- ses méta-informations descriptives ;
- les segments thématiques qui le composent dans le cas où le document est volumineux et qu'il nécessite une segmentation, obtenus en appliquant un algorithme de segmentation thématique sur le corpus, ces segments décrivent les thèmes abordés dans le document ;
- la liste des concepts pondérés par les *tof*×*idf* et les termes pondérés par les *tf*×*idf.*

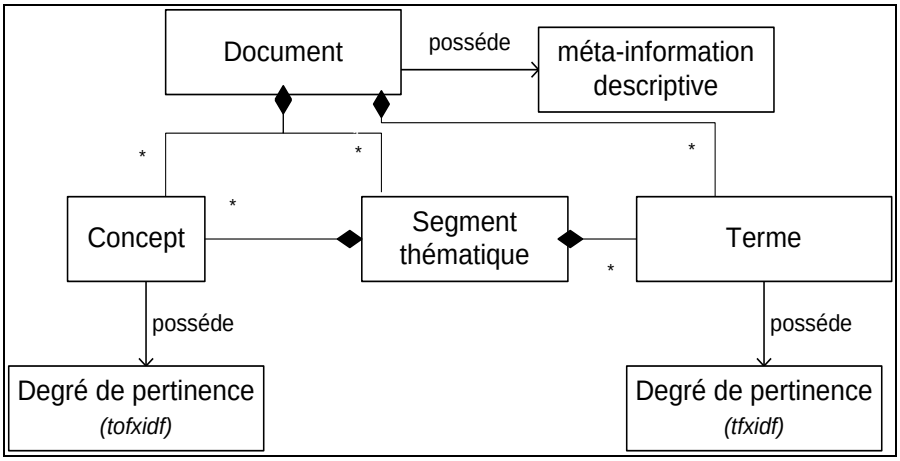


Figure 3.10 Notre méta-modèle du référentiel

La génération des méta-informations descriptives permettent d'annoter les documents par des annotations descriptives. Ces méta-informations peuvent englober des renseignements (des précisions) concernant les auteurs, des informations relatives au titre, la période de création, la langue, la spécialité, la taille du document, le territoire géographique, le format de la ressource, etc. Dans le cadre de notre travail, ces méta-informations jouent un rôle très important puisqu'elles permettent de filtrer et sélectionner les ressources les plus pertinentes correspondant aux besoins de l'utilisateur selon des critères bien précis. Nous explicitons ce point dans la deuxième section du chapitre suivant consacrée à la description des étapes de création du référentiel.

Pour notre cas, nous avons spécifié les méta-informations descriptives suivantes : titre, format, type, langue, taille, public cible, auteur, date, organisation, pertinence, objet ou le scope, URL et nous avons défini un modèle de document annoté via un schéma XML comme le montre la figure 3.11.

```
<?xml version="1.0" encoding="UTF-8" ?>
<!_ edited with XMLSpy v2006 rel. 3 sp1 (http ://www.altova.com) by www.serials.ws
(www.serials.ws) _>
<xs :schema xmlns :xs="http ://www.w3.org/2001/XMLSchema" elementFormDefault="qualified"
attributeFormDefault="unqualified">
<xs :element name="Document_annoté">
<xs :annotation>
<xs :documentation>Comment describing your root element</xs :documentation>
<xs :documentation/>
<xs :documentation/>
<xs :documentation/>
<xs :documentation source="doc_annoté"/>
</xs :annotation>
<xs :complexType>
<xs :sequence>
<xs :element name="En-tête">
<xs :complexType>
<xs :sequence>
<xs :element name="MD_descriptives">
<xs :complexType>
<xs :sequence>
<xs :element name="Identifiant" type="xs :anyURI"/>
<xs :element name="Auteur">
<xs :complexType>
<xs :sequence>
<xs :element name="Titre"/>
<xs :element name="Scope"/>
<xs :element name="Organisation" maxOccurs="unbounded">
<xs :complexType>
<xs :sequence>
<xs :element name="Editeur"/>
<xs :element name="Date">
<xs :complexType>
<xs :sequence>
<xs :element name="Date_création" type="xs :date"/>
<xs :element name="Date_modification" type="xs :date"/>
</xs :sequence>
</xs :complexType>
</xs :element>
<xs :element name="Format"/>
```

```
<xs :element name="Taille" type="xs :integer"/>
<xs :element name="Langue"/>
<xs :element name="Pertinence"/>
</xs :sequence>
</xs :complexType>
</xs :element>
<xs :element name="Public_cible">
<xs :complexType>
<xs :sequence>
<xs :element name="Type_ressource"/>
<xs :element name="URL"/>
</xs :sequence>
</xs :complexType>
</xs :element>
</xs :sequence>
</xs :complexType>
</xs :element>
<xs :element name="Fragment" maxOccurs="unbounded">
<xs :complexType>
<xs :sequence>
<xs :element name="Titre" minOccurs="0"/>
<xs :element name="Contenu">
<xs :complexType>
<xs :sequence>
<xs :element name="Annotations_conceptuelles">
<xs :complexType>
<xs :sequence>
<xs :element name="Concepts_clés" maxOccurs="unbounded">
<xs :complexType>
<xs :sequence>
<xs :attribute name="tof×idf "/>
<xs :element name="Termes_clés" maxOccurs="unbounded">
<xs :complexType>
<xs :sequence>
<xs :attribute name="tf×idf "/>
</xs :complexType>
</xs :element>
</xs :sequence>
</xs :schema>
```

Figure 3.11 Le Schéma XML du référentiel sémantique

3.3.4 Combinaison des méta-modèles du référentiel et de Topic Map pour la recherche d'information

Notre choix de proposer deux méta-modèles pour notre approche de recherche intelligente présente plusieurs avantages. En effet, le méta-modèle des Topic Maps et celui du référentiel sont complémentaires et pour bénéficier des particularités de ces deux méta-modèles, nous proposons une démarche qui permet de les combiner de la manière suivante comme le montre la figure 3.12 :

- Le **niveau interne** décrit le modèle de stockage, il est représenté par la Topic Map sous le **format XTM** et par le référentiel de documents que nous avons choisi de représenter sous le **format XML** ;

- le **niveau externe** décrit le modèle de présentation, c'est-à-dire la **visualisation** graphique de la Topic Map à partir du fichier XTM qui va servir pour la recherche par navigation et la recherche par **requête** effectuée à travers le référentiel de documents, ce référentiel est construit à partir du fichier XML.

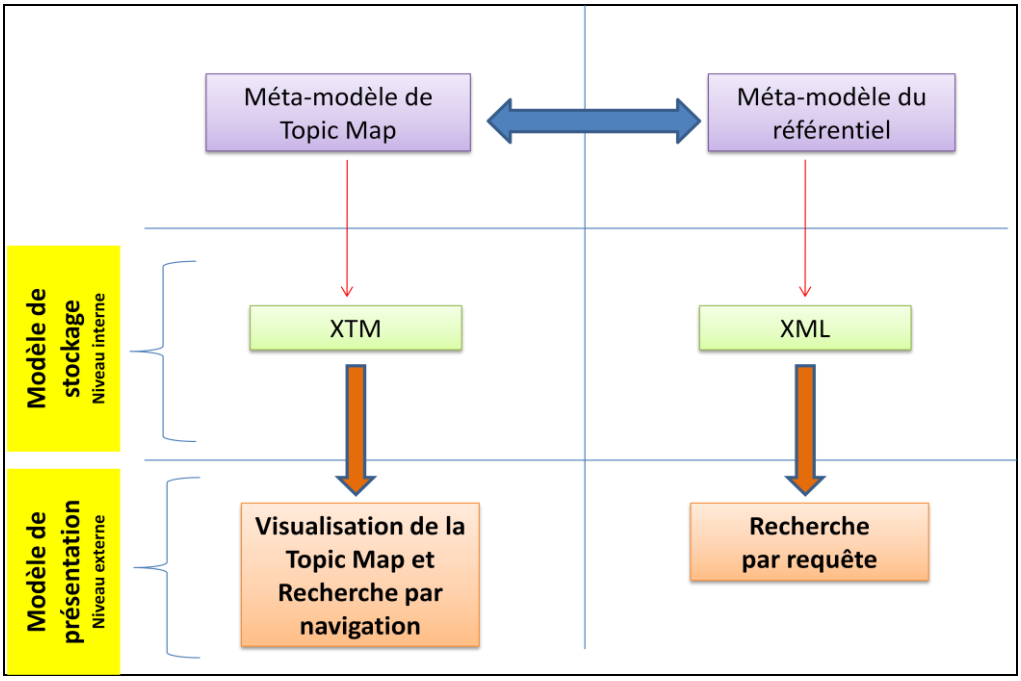


Figure 3.12 Combinaison des méta-modèles du référentiel et de Topic Maps

La cohabitation entre le méta-modèle des Topic Maps et celui du référentiel est l'une des originalités de notre approche, elle présente plusieurs avantages pour notre travail comme par exemple l'indexation d'un Topic par un segment de document et non pas par le document en entier. En effet, la notion de segment n'existe pas dans le modèle des Topic Maps, selon ce modèle un Topic ne peut pas être relié à un segment de document, la notion d'occurrence telle que présentée dans le modèle permet de relier un Topic à tout le document. Or, dans le cas de document de très grande taille, l'utilisateur sera amené à examiner la totalité du document pour retrouver le paragraphe ou la partie qui l'intéresse. Pour cette raison, l'utilisation du référentiel dans la construction de la Topic Map nous a permis d'intégrer la notion de segments dans notre méta-modèle de Topic Map.

De plus nous exploitons les pondérations obtenues dans le référentiel ($tf \times idf$ et $tof \times idf$) pour l'annotation des liens occurrences reliant les Topics aux documents et segments de documents qui en parlent. Les liens entre Topics thèmes et segments de documents seront pondérés par les $tof \times idf$ et les liens entre Topics (instances, synonymes de Topics) et documents ou segments seront étiquetés par les $tf \times idf$.

Plus concrètement, les pondérations des arcs occurrences seront représentées par la notion de facettes, un ensemble d'attributs valeurs reliés aux ressources, nous proposons **un attribut degré de pertinence** qui reflète l'importance du document par rapport au Topic et **un attribut langue** qui contient la langue du document. L'attribut degré de pertinence servira pour le filtrage des ressources selon leur importance pour n'afficher à l'utilisateur que les documents et les segments pertinents triés par degré de pertinence. L'attribut langue va servir à filtrer les documents selon la langue dans le cas où l'utilisateur souhaite afficher que les documents en anglais ou en français.

Dans notre Topic Map, les Topics sont typés et organisés par des liens ontologiques et des liens d'usage, c'est à travers ces liens que l'utilisateur pourra naviguer dans la Topic Map, cette notion de typage et d'organisation n'existe pas dans le référentiel d'où l'avantage du modèle des Topic Maps par rapport à celui du référentiel.

A long terme, nous prévoyons d'inclure le méta-modèle du référentiel dans la norme XTM pour l'intégration de la notion de segments et pourvoir indexer un Topic par un segment de document surtout dans le cas de documents de très grande taille (par exemple des thèses de 250 pages). Ce point sera traité comme perspectives de nos travaux de recherche.

En résumé, notre démarche est fondée sur la combinaison des deux méta-modèles du référentiel et de Topic Map, elle profite de leurs avantages pour améliorer l'accès aux documents et faciliter la recherche dans leur contenu. Dans la section suivante, nous décrivons les trois modes de recherche pris en compte dans notre approche.

3.4 Types de recherche offerts par notre approche

Cette section présente les trois modes de recherche offerts par notre approche : une **recherche par navigation** à travers la Topic Map multilingue enrichie à partir du thésaurus du domaine des ontologies générales et des scénarios d'usage et annotée par les documents et leurs segments thématiques, **une recherche par requête à partir de scénarios de questions** intégrés dans la Topic Map dont on prévoit les réponses adéquates à partir de FAQ et **une recherche classique par requête** dans le référentiel de documents annotés descriptivement et indexés sémantiquement et thématiquement.

Nous détaillerons tous ces points dans ce qui suit, nous nous concentrons en particulier sur la recherche par navigation et la notion de champ de proximité sémantique d'un Topic.

3.4.1 Recherche par navigation

3.4.1.1 La Topic Map comme espace sémantique de navigation et cible de recherche

La Topic Map représente une carte sémantique qui permet d'associer à chaque Topic les documents ainsi que les segments de documents les plus représentatifs de ce Topic. Cette Topic Map est utilisée pour organiser les documents et représenter leurs contenus sémantiques grâce aux réseaux de liens sémantiques entre les sujets qu'elles représentent, nous rappelons que la Topic Map est composée de nœuds qui représentent les Topics, les documents et les segments de documents et d'arcs typés et pondérés qui représentent les liens ou associations, ces liens sont de différents types : des liens entre deux ou plusieurs Topics, des liens entre un Topic et ses instances, des liens de synonymie entre Topics, des liens que nous avons appelés liens d'usage entre Topics questions et les Topics qui permettent de répondre à ces questions, des liens d'occurrences reliant un Topic aux documents et/ou aux segments qu'il indexe étiquetés par le degré de pertinence (*tof*×*idf* ou *tf*×*idf*) de chaque Topic dans le document (et/ou dans le segment).

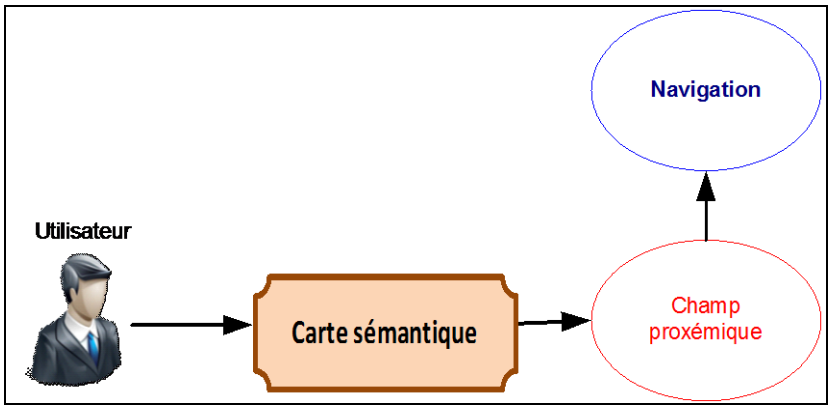


Figure 3.13 Recherche par navigation

La Topic Map affiche au centre le Topic sélectionné par l'utilisateur. Elle permet de représenter et d'organiser de manière graphique le **champ de proximité sémantique** d'un Topic. Ce champ contient les documents et les segments de documents (les plus pertinents) associés aux différents Topics qui sont liés sémantiquement au Topic cherché par l'utilisateur. Ce qui permet de réaliser une recherche connotative, exploratoire et thématique.

La recherche par navigation présente plusieurs avantages, en effet, l'utilisateur a accès à tous les documents répondant à une recherche précise, exploratoire et connotative à travers la Topic Map. De ce fait, il n'y aurait pas besoin de reformuler ou étendre ses requêtes (figure 3.13). De plus, la recherche par navigation permet de couvrir l'ensemble du champ de proximité sémantique correspondant aux Topics du besoin formulé. Les similarités

sémantiques, les synonymes, les liens hiérarchiques (généralisation/spécialisation, composition) et associatifs entre Topics sont couverts par ce champ permettant à l'utilisateur de découvrir l'ensemble de l'arbre sémantique sous-jacent au besoin formulé.

3.4.1.2 Champ de proximité sémantique d'un Topic

Le champ de proximité sémantique représente la proximité sémantique qui peut exister entre deux Topics. La notion de proximité sémantique est différente de celle relative à la similarité sémantique, Haifa Zargayouna a travaillé sur ces deux notions dans le cadre de sa thèse [Zargayouna, 2005].

La notion de proximité sémantique (*semantic relatedness*) est une notion plus large que la similarité sémantique, elle prend en considération tout type de lien entre les Topics. Ainsi deux Topics peuvent être proches sémantiquement par leur similarité de point de vue sens, par exemple « voiture » et « automobile » mais aussi par d'autres liens comme par exemple le lien *partie de*, « réservoir » est une partie de « chaudière » ou contraire entre « chauffage» et « climatisation ».

La figure 3.14 présente un exemple de Topic Map illustrant le champ de proximité sémantique du Topic « chauffage » dans les deux langues traitées. Ce champ recouvre les liens de subsomption, les liens de synonymie et les liens associatifs relatifs au domaine d'étude et identifiés suite à l'analyse linguistique des documents source et avec l'aide d'experts du domaine.

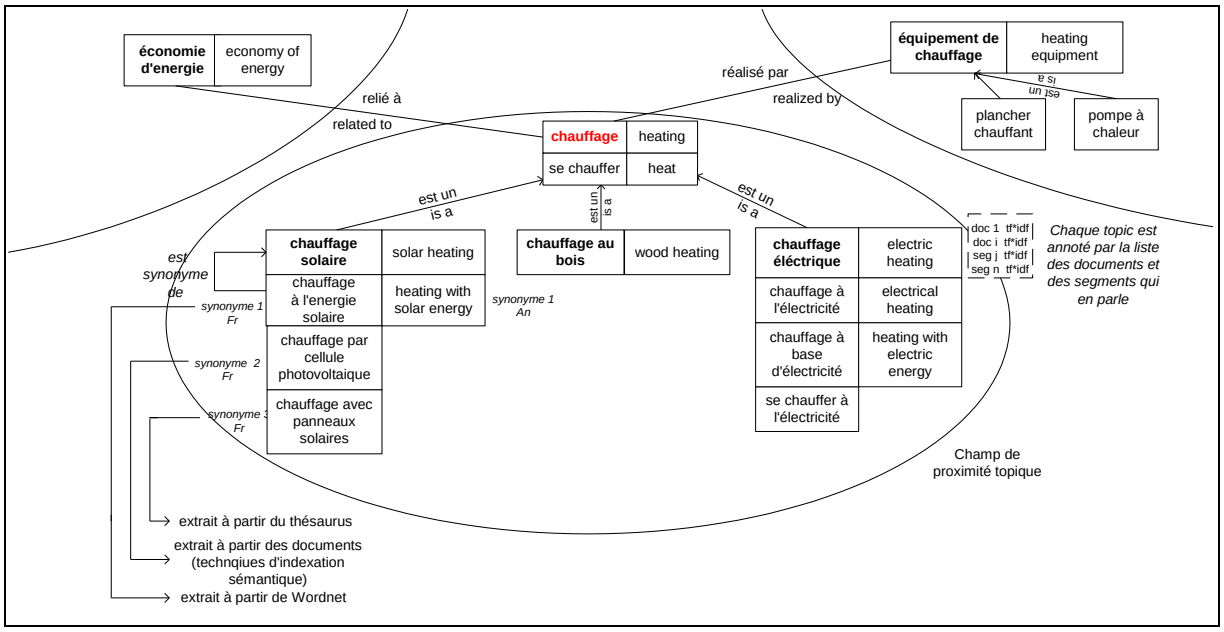


Figure 3.14 Champ de proximité sémantique du Topic « chauffage »

La recherche par navigation est destinée à des utilisateurs qui ont des difficultés pour exprimer leurs besoins en information et qui ne connaissent pas le vocabulaire utilisé par les auteurs des documents. Ce type de recherche fournit à l'utilisateur le champ de proximité sémantique qui correspond au Topic cherché. De cette manière, nous évitons la reformulation ou l'expansion de la requête, puisque l'utilisateur aura une vision globale des différents Topics et les instances de son domaine ainsi que les liens qui les relient. Il peut affiner sa requête en choisissant son parcours de recherche et en sélectionnant les segments de documents qui répondent à son besoin.

Par exemple, un utilisateur, cherchant les documents ou les segments de documents qui parlent du Topic « chauffage », aura le sous arbre de la Topic Map relatif à ce Topic. Ce sous arbre est schématisé dans la figure 3.14. L'utilisateur peut alors trouver les différentes instances de ce concept ainsi que les différents documents attachés à chaque instance pondérées avec leur $tf \times idf$ et les segments de chaque document qui contiennent cette instance.

Un utilisateur qui cherche, par exemple, des documents qui parlent du « chauffage au bois » comme solution pour l'économie d'énergie, il s'aperçoit, lors de sa navigation dans le champ de proximité sémantique de ce Topic, que le chauffage au bois a des effets négatifs sur la santé (bronchite, irritations du système respiratoire) d'où l'intérêt des Topic Maps.

3.4.2 Recherche basée sur des scénarios de questions préparés à partir de FAQ

Pour représenter l'usage dans la Topic Map, nous proposons **d'intégrer les questions des utilisateurs dans la Topic Map**. Pour cela, nous élaborons un ensemble de scénarios de questions recensés et collectés à partir de FAQ et nous proposons de représenter ces scénarios ainsi que leurs réponses dans la Topic Map, c'est à dire une question type est représentée par un Topic reliés aux termes qui composent la question qui sont eux mêmes représentés comme des Topics, ces liens sont nommés « est composé de » qui font partie de la classe liens d'usage. Ensuite ce Topic question est relié aux Topics qui permettent de répondre à cette question, ces liens nommés « répond à » font aussi partie des liens d'usage. Les réponses à ces questions sont extraites à partir de FAQ.

Comme la FAQ est un Topic, l'utilisateur peut le voir en navigant. Cette démarche permet de **faciliter la tache de recherche** pour l'utilisateur et lui permettre **de gagner du temps**, dans la mesure où il trouve directement, lors de sa navigation dans la Topic Map, le Topic qui correspond à la question qu'il voulait poser et dans ce cas, il n'aura pas à formuler

sa requête, il pourra visualiser la réponse à sa question directement sans avoir besoin à la poser en parcourant les hyper liens « répond à » entre le Topic question et les Topics réponses.

Nous utilisons les liens « est composé de » reliant le Topic de la question aux termes qui la constituent pour **la recherche automatique de question similaires**, en effet, chaque question est représentée sous forme d'un vecteur de Salton [Salton, 1989] contenant les termes qui la composent, lorsqu'un utilisateur pose sa question, cette dernière est, elle aussi, représentée sous forme d'un vecteur de Salton, il s'agit ensuite de rechercher, par le calcul de distance entre le vecteur de la requête FAQ et le vecteur de la requête de l'utilisateur, le vecteur de la FAQ qui se rapproche le plus du vecteur des mots de la requête parmi la liste des requêtes appartenant aux scénarios d'usage déjà préparées et représentées dans la Topic Map. Enfin, nous renvoyons à l'utilisateur les réponses adéquates puisque celles-ci sont déjà prévues et recensées à partir de l'analyse des FAQ. Nous détaillerons l'étape de recherche des FAQ les plus proches de la requête de l'utilisateur dans le chapitre suivant.

3.4.3 Recherche par requête en utilisant un langage de requêtes

En plus de la recherche par navigation à travers la Topic Map, pour les utilisateurs qui préfèrent le mode de recherche classique par requête (figure 3.15), le référentiel est utilisé pour répondre à ce type de recherche. Après appariement entre documents et requêtes, le système retrouve la liste des documents et des segments qui répondent au mieux au besoin de l'utilisateur. Ce dernier exprime sa requête en langage naturel, cette requête est ensuite traduite en un langage de requête dédié au modèle des Topic Maps par exemple Tolog ou bien un langage tel que SPARQL, le système retrouvera, tous les documents et leurs segments qui répondent à cette requête.

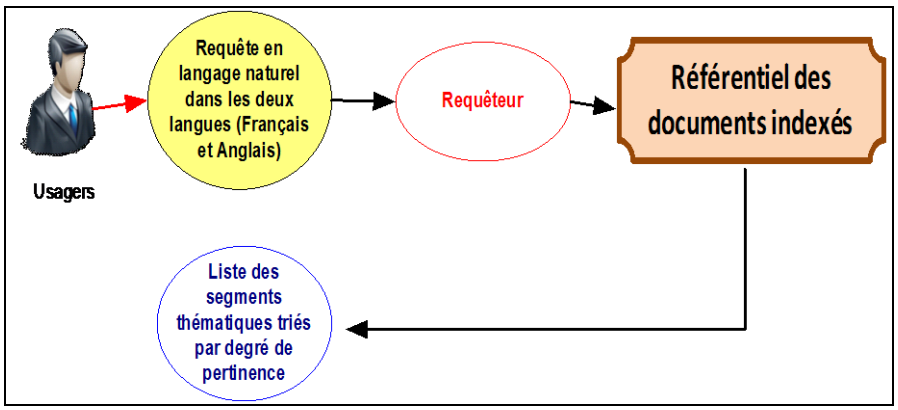


Figure 3.15 Recherche par Requête

SPARQL est le langage développé par le W3C pour interroger des descriptions RDF. Tolog [Garshol, 2001] est un langage destiné à l'interrogation des Topic Maps, édité par la société Ontopia, il respecte la spécification TMQL. Tolog est inspiré de DataLog et SQL. Tolog permet d'exécuter différents types de requêtes, il permet de rechercher :

- Tous les Topics dans une Topic Map ;
- Les Topics ayant un certain type dans un certain contexte ;
- Tous les Topic utilisés dans une association donnée, etc.

En Tolog, une requête constitue un ensemble de prédicats. Un prédicat peut avoir un ou plusieurs paramètres. Le résultat d'une requête est constitué de l'ensemble des valeurs des variables (paramètres) qui vérifient les prédicats. Dans la majorité des cas, il existe des relations implicites entre les Topics qui ne sont pas représentées d'une manière explicite sous forme d'associations. Ces relations peuvent être déduites à partir d'autres relations explicites. Tolog offre la possibilité d'exprimer ces règles implicites en utilisant les règles d'inférences.

Tolog permet l'interrogation d'un fichier XTM en se basant sur deux types de liens : le lien de spécialisation/généralisation et le lien d'instanciation. Le voisinage d'un Topic est défini par les parents, les frères et les descendants directs (fils) de ce Topic.

Pour exécuter une requête donnée il faut tout d'abord, suivant le principe de Tolog, charger la Topic Map en question, ensuite instancier un nouveau évaluateur de requêtes tolog, puis ajouter la ou les règle(s) d'inférence nécessaires pour l'interprétation de la requête et à la fin exécuter la requête (Figure 3.16). La récupération du résultat de la requête est possible soit dans des variables, soit le stocker dans un fichier XTM.

```
#chargement de la topic map
tm=readTopicMap(fileIn);

#création d'un nouveau évaluateur
QueryEvaluator queryEvaluator = QueryEvaluatorFactory.newQueryEvaluator(tm);

#ajout de la règle d'inférence
queryEvaluator.addRule("descendant-of($A, $B) :- { direct-descendant-of($A, $B) |
direct-descendant-of($A, $C), descendant-of($C, $B) }.");

#execution de la requête
TologResultsSet  results = queryEvaluator.execute("instance-of($A, chauffage) ?");
```

Figure 3.16 Exemple de schéma d'exécution d'une requête Tolog.

3.5 Conclusion

Dans ce chapitre, nous avons présenté notre approche générale de recherche intelligente fondée sur deux méta-modèles : le méta-modèle du référentiel de documents indexés sémantiquement et thématiquement et le méta-modèle des Topic Maps. Nous avons commencé par rappeler la problématique et les objectifs de notre approche ainsi que les raisons pour lesquelles nous avons choisi le modèle des Topic Maps pour la réalisation de notre travail. Ensuite, nous avons présenté notre approche générale et décrit en détail les deux méta-modèles proposés. Enfin, nous avons mis en exergue les avantages de notre approche puisqu'elle offre à l'utilisateur, en plus de la recherche classique par requête qui s'appuie sur le référentiel de documents, un mode de recherche par navigation à travers la Topic Map.

L'idée de notre approche est de proposer **une méthode de construction et d'enrichissement de Topic Map multilingue** à partir **d'un référentiel sémantique** de documents textuels et multilingues segmentés thématiquement et indexés sémantiquement, d'un thésaurus bilingue du domaine, de deux ontologies générales et d'un ensemble de scénarios d'usage. Les **deux** produits de la démarche constructive de la Topic Map et annotative du corpus serviront aux deux modes de **recherche par navigation et par requête**.

Dans notre approche, nous avons proposé d'**étendre** le modèle des Topic Maps déjà existant et proposé **un nouveau méta-modèle de Topic Map** en définissant les liens d'usage et une liste de **méta-propriétés** associés à chaque Topic, ces méta-propriétés seront utilisées dans le processus d'évolution et d'élagage de la Topic Map afin de supporter les éventuels changements relatifs aussi bien au contenu de la Topic Map qu'à son utilisation au cours du temps .

Dans notre méta-modèle, nous précisons la sémantique des liens dans les Topic Maps, en effet, comme nous l'avons déjà mentionné le modèle des Topic Maps ne fait pas la différence entre les liens dans une Topic Map, ils sont tous représentés comme des associations à part les liens d'occurrences qui reliant les Topics aux ressources.

Nous avons classé ces liens en deux classes différentes; ceux que nous avons appelés **les liens ontologiques** et ceux nommés **les liens d'usage**. En effet, un des principaux fondements de notre approche est la prise en compte de l'usage c'est-à-dire toutes les questions potentielles relatives aux documents sources. Dans notre cas, nous avons sélectionné ces questions et leurs réponses à partir de FAQ.

En plus de méta-modèle de Topic Map, nous avons proposé un méta-modèle pour la représentation et la description sémantique des documents du corpus sous la forme d'un

référentiel de documents annotés avec des méta-informations descriptives (telles que le type, le public cible, la taille, l'URL de la ressource,…), segmentés thématiquement et indexés sémantiquement.

La **combinaison entre le méta-modèle des Topic Maps et celui du référentiel** est l'une des originalités de notre approche de recherche intelligente, en effet le méta-modèle du référentiel permet de compléter celui des Topic Map surtout par rapport à la possibilité d'indexer un Topic par un segment de document. Et inversement, le méta-modèle des Topic Map permet d'organiser, typer les Topics et définir des liens entre eux, ce qui n'est pas possible dans le méta-modèle du référentiel.

Au final, le référentiel annoté est destiné à être la cible des procédures de recherche en mode requête, alors que la Topic Map doit servir au mode de recherche par navigation.

Recherche par requête

Ce mode de recherche s'appuie sur le référentiel annoté thématiquement et sémantiquement, ce référentiel est destiné à servir d'espace de recherche à base d'un requêteur de type Tolog ou SPARQL par exemple. Concernant les requêtes, nous avons proposé de les formuler en langage quasi-naturel dans une des deux langues du corpus avec une traduction en langage de requête pour fournir à la sortie une liste de documents et de leurs segments les plus significatifs dans les deux langues triée par degré de pertinence.

Recherche par navigation

Ce type de recherche s'effectue à travers les liens de la Topic Map. La navigation permet de réduire la charge cognitive de l'utilisateur ; ce dernier n'est plus obligé de reformuler sa requête ou parcourir la totalité de la ressource retournée afin de trouver le segment qui répond le mieux à son besoin. La recherche par navigation permet à l'utilisateur de choisir son parcours de recherche à partir de la Topic Map. Cette dernière propose alors le **champ de proximité sémantique** qui répond au besoin de l'utilisateur. Ce dernier peut affiner, par la suite, ses choix en navigant dans la Topic Map et en utilisant les liens sémantiques dans le champ de proximité du Topic concerné. A travers **l'intégration des scénarios d'usage (FAQ)** dans la Topic Map, l'utilisateur, lors de sa navigation, pourra accéder directement au Topic qui correspond à sa question ce qui lui permet de gagner du temps dans sa recherche et afficher les réponses en utilisant l'hyper lien « répond à » reliant la question aux Topics réponses.

CHAPITRE 4

Description détaillée de l'approche proposée

Nous nous concentrons dans ce chapitre sur la description détaillée des deux principaux modules de notre approche : Le premier concerne **la construction du référentiel de documents** segmentés thématiquement et indexés sémantiquement et le deuxième module concerne **la construction incrémentale et l'enrichissement de la Topic Map multilingue** à partir de quatre sources d'information : (a) le référentiel de documents avec pour chaque document, ses segments thématiques et la liste des termes représentatifs indexant sémantiquement le document et ses segments, (b) un thésaurus bilingue (Français/Anglais) du domaine ; (c) deux ontologies générales (WordNet pour l'anglais et WOLF pour le français) et (d) un ensemble de scénarios d'usage sous forme de questions/réponses que nous recensons à partir de sources d'interrogations possibles telles que les questions relatives aux documents sources qu'un expert du domaine ou un utilisateur quelconque peut poser, les foires aux questions (FAQ), les traces des discussions téléphoniques et des consultations directes avec les travailleurs du domaine.

Dans ce chapitre, nous commençons tout d'abord par la description des étapes de construction du référentiel de documents. Ensuite, nous détaillons les différentes phases de notre approche de construction de Topic Map multilingue que nous avons appelée **ACTOM** pour **A**pproche de **C**onstruction d'une **TO**pic Map **M**ultilingue. Enfin, avant de conclure, nous évoquons les problèmes liés au multilinguisme rencontrés lors du processus de construction de la Topic Map et les démarches proposées pour les résoudre.

4.1 Construction du référentiel de documents

L'objectif de cette étape est de représenter le contenu des documents dans un référentiel sémantique avec pour chaque document la liste des segments thématiques qui le compose et pour chaque segment et documents la liste des termes représentatifs de leur contenu. Ce référentiel sera utilisé d'une part pour la construction de la Topic Map et d'autre part pour effectuer des recherches par requête dans le cas où l'utilisateur choisit ce mode d'interrogation. La construction du référentiel se divise en trois phases (figure 4.1) :

- Le **prétraitement** des documents sources ;
- La **segmentation thématique des documents** du corpus afin d'extraire les segments thématiquement homogènes à partir de chaque document. Nous proposons dans cette phase d'utiliser le segmenteur thématique TextTiling ;
- L'**indexation sémantique des documents** du corpus en utilisant l'algorithme d'indexation sémantique LSI pour l'extraction des termes et des concepts

représentatifs du contenu de chaque document et de ses segments thématiques, chaque terme est pondéré par son degré de pertinence, dans notre cas nous utilisons la mesure *tf*×*idf*, et chaque concept est pondéré par une mesure appelée *tof*×*idf* (*topic frequency inverse document frequency*) que nous avons défini dans le contexte de notre travail.

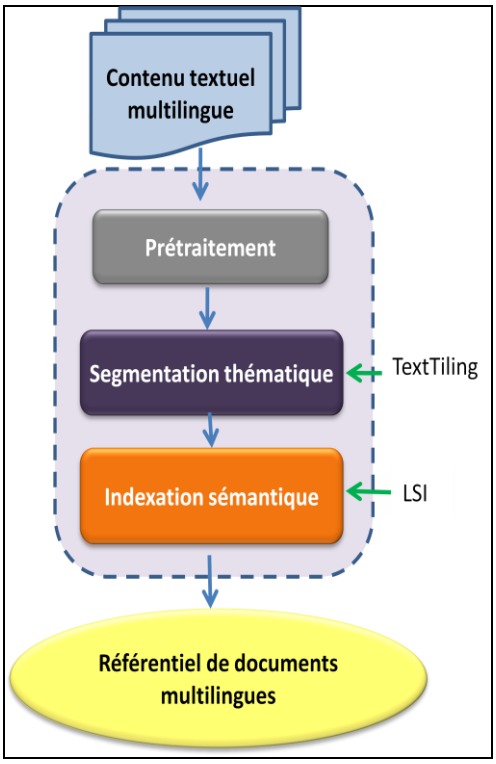


Figure 4.1 Etapes de construction du référentiel

Les résultats obtenus suite aux précédentes phases serviront à la **génération d'un référentiel de documents** multilingues annotés sémantiquement et thématiquement. Ils participent ainsi à la satisfaction du type de recherche par requête notamment la recherche précise (chercher les documents qui traitent un ou plusieurs Topics donnés), la recherche exploratoire et la recherche connotative (chercher les documents parlant de Topics similaires à un Topic donné).

4.1.1 Prétraitement des documents

Lors de cette étape nous commençons tout d'abord par l'élimination des mots vides. Ceux sont les termes d'usage qui ont une fréquence souvent élevée et qui n'apportent aucune information pour la description du document. Ensuite nous effectuons un filtrage statistique, c'est-à-dire nous éliminons les termes ayant une fréquence inférieure à un seuil. En effet, la loi de Zipf [Zipf, 1949] montre que les termes peu fréquents sont les plus nombreux, en

conséquence le nombre des termes d'indexation peut augmenter de manière très importante avec la taille du corpus. L'élimination de ces termes est donc justifiée surtout si on connait que la plupart d'eux sont peu représentatifs du contenu d'un document et ne sont donc pas discriminant pour sa description. Finalement, nous procédons à la lemmatisation des termes restants afin de regrouper en un seul attribut les multiples formes morphologiques de mots qui ont une sémantique commune.

4.1.2 Segmentation thématique des documents textuels

Pour améliorer les performances de leurs systèmes de recherche d'information, certains travaux proposent de commencer par une étape de segmentation thématique des documents source, en effet, dans le cas de la navigation à l'intérieur de longs documents textuels, avoir des documents thématiquement segmentés peut résulter en la récupération des segments de texte courts et pertinents qui correspondent directement à la requête d'un utilisateur au lieu de longs documents examinés avec soin par l'utilisateur pour trouver l'objet de son intérêt.

La segmentation thématique de textes a pour objectif de localiser les changements de thème dans les documents [Ferret, 2006]. Ce type d'information peut permettre l'amélioration de la recherche et la navigation à l'intérieur de longs documents, qui est le cas de notre travail. Selon la littérature, les approches de détection des changements de thème dans un texte se divisent en deux classes majeures [Ferret, 2006] : les approches linguistiques et les approches dirigées par le contenu.

Approches linguistiques

Ces approches se basent sur l'identification de marques, en particulier linguistiques, caractéristiques des changements de thème. Certaines des marques discursives, en l'occurrence les introducteurs de cadres thématiques, permettent de réaliser une forme d'identification des thèmes d'un document. Dans une phrase commençant par *En ce qui concerne les problèmes du chauffage au bois...*, l'introducteur de cadre *en ce qui concerne* est une marque de segmentation et le groupe nominal qui la suit caractérise le thème du nouveau segment [Grosz et Sidner, 1986], [Passonneau et Litman, 1993].

Approches dirigées par le contenu

Ces approches se basent sur la détection des changements du contenu du discours. Les approches dirigées par le contenu ont emprunté deux voies. La première ne fait pas appel à

des connaissances externes, pour ceci ces méthodes sont aussi appelées méthodes endogènes. La deuxième piste se base sur l'exploitation des connaissances sur les relations de cohésion lexicale. Ces méthodes sont appelées aussi méthodes exogènes [Choi et al. 2001], [Utiyama et Isahara, 2001], [Ferret, 2002], [Galley et al. 2003], [Caillet et al. 2004], [Labadié et Chauché, 2007].

Dans cette deuxième approche, les algorithmes de segmentation thématique des documents se décomposent dans la majorité des cas en trois étapes distinctes :

- La première étape permet de **diviser le document en entrée en blocs élémentaires**. Un bloc élémentaire décrit un thème et peut être constitué d'un ensemble de phrases, de paragraphes ou encore avoir une taille arbitraire ;
- La seconde étape consiste à **mesurer les similarités entre les blocs élémentaires**. Cette évaluation s'appuie sur la notion de cohésion lexicale qui caractérise le fait qu'un texte se présente comme un tout cohésif et non comme la simple juxtaposition d'un ensemble de phrases. Différentes techniques ont été définies pour évaluer la cohésion lexicale existante entre des blocs de texte dans le cadre de la segmentation thématique. La plupart d'entre elles s'appuient sur la simple notion de **répétition lexicale** [Choi, 2000] mais certaines cherchent à mettre en évidence des relations moins directes, soit par l'intermédiaire de **cooccurrences** lexicales [Choi, 2000] [Ferret, 2002], soit par l'utilisation de relations explicitement typées (**hyperonymie, synonymie**, …) provenant de thésaurus ou de dictionnaires [Morris et Hirst, 1991] ;
- La dernière étape **définit les blocs thématiques homogènes entre eux** grâce à une technique de regroupement. Cette technique fusionne les éléments élémentaires constituant un même thème. Néanmoins la principale difficulté de ces algorithmes demeure dans la détermination du nombre de blocs thématiques par document, c'est-à-dire l'identification du nombre de thèmes traités.

Il faut remarquer que la première approche a été globalement moins étudiée que la seconde. Une explication possible de cette situation, selon [Ferret, 2006] est que s'attacher au contenu du discours est en apparence toujours réalisable alors que la seconde approche est limitée aux discours présentant des changements de thème caractérisés par des marques spécifiques. Or ces marques ne sont en pratique pas très fréquentes et sont par ailleurs assez dépendantes du type de discours considéré. Lorsqu'elles sont présentes, elles ont néanmoins l'avantage de localiser les changements de thème avec précision ce qui n'est, à l'inverse, pas

toujours le cas des méthodes fondées sur le contenu du discours. Les deux approches sont donc plus complémentaires que concurrentes. C'est d'ailleurs ce que l'on observe dans les travaux existants puisqu'une part importante de ceux exploitant des marques linguistiques de segmentation le font en conjonction avec des indices relatifs au contenu du discours.

Dans notre travail, nous nous intéressons à l'approche dirigée par le contenu. Les algorithmes les plus connus représentant cette classe sont l'algorithme TextTiling de Marti Hearst [Hearst, 1997], l'algorithme Segmenter de [Kan et al. 1998], l'algorithme C99 de Choi [Choi, 2000] et l'algorithme DotPlotting de [Reynar, 2000].

L'étape de segmentation thématique des documents textuels a pour objectif de segmenter les documents du corpus en plusieurs blocs thématiquement cohérents. Nous disposons de documents assez volumineux et la segmentation permet d'améliorer l'accès à ces documents. Nous proposons pour cette étape d'utiliser l'algorithme TextTiling puisqu'il supporte en entrée plusieurs formats de documents : .pdf, .html, .rtf, .doc. En plus, la version dont nous disposons permet de traiter les textes en anglais aussi bien qu'en français et un autre critère très important est celui du temps d'exécution, effectivement l'algorithme TextTiling que nous utilisons permet de traiter des documents volumineux (thèse 230 pages) en des temps raisonnables (30s) ce qui fait de cet algorithme un bon compromis entre complexité et efficacité [Chaar, 2003]. Par ailleurs, l'algorithme TextTiling est robuste et ne demande pas beaucoup de ressources pour être exécuté. D'après la littérature, son utilisation dans le cadre des systèmes d'extraction et de recherche d'information donne de bons résultats. L'unité de base pour cet algorithme est un segment de texte défini par un nombre fixe de phrases. La figure 4.2 illustre les principales étapes de l'algorithme de TextTiling.

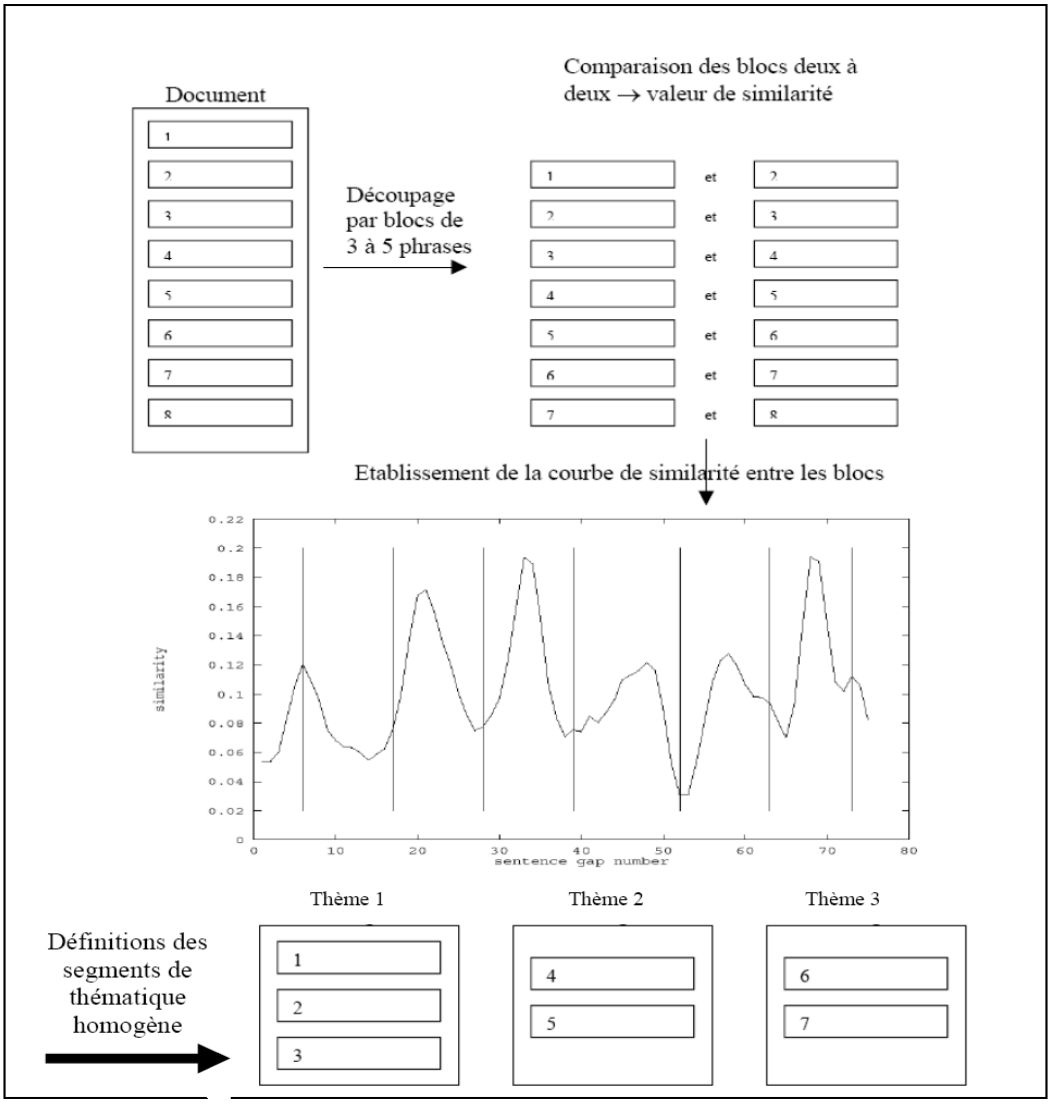


Figure 4.2 Illustration des étapes de l'algorithme de TextTiling

4.1.2.1 La segmentation physique

Après le prétraitement des documents textuels, la première étape est la segmentation physique du document, elle se base sur une mesure de similarité lexicale. Les lemmes issus des textes prétraités sont groupés en pseudo phrases, c'est-à-dire un ensemble de lemmes adjacents, qui sont eux mêmes regroupés en blocs de taille fixée au départ (Figure 4.3). Cette taille des segments est variable, elle peut aller de 3 à 5 pseudo phrases à un paragraphe. En général, on prend la moyenne de la longueur des paragraphes. Les paragraphes réels ainsi que les phrases ne sont pas pris en compte car leur longueur peut être fortement irrégulière conduisant à des comparaisons déséquilibrées.

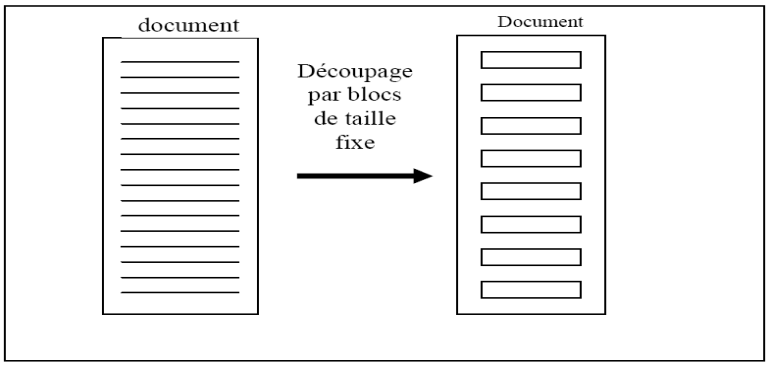


Figure 4.3 Mise en place de blocs de pseudo phrases

4.1.2.2 Calcul de la similarité entre blocs adjacents

La deuxième étape, « Calcul de la similarité entre blocs adjacents » (figure 4.4), consiste à déplacer une fenêtre de taille fixe et calculer pour chaque position de celle-ci la similarité entre ses deux parties.

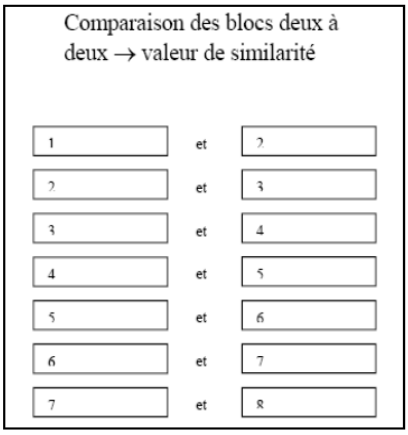


Figure 4.4 Calcul de la similarité entre deux blocs adjacents de pseudo phrases

La similarité est calculée par la mesure cosinus. (Equation 4.1 : étant donné des blocs de textes b1 et b2).

$$score(i) = \frac{\sum_t W_{t,b1} W_{t,b2}}{\sqrt{\sum_t W_{t,b1}{}^2 \sum_t W_{t,b2}{}^2}}$$

Equation 4.1

Où *t* s'étend à l'ensemble des termes dans le document et $w_{t,b1}$ est le poids *tf*×*idf* assigné au terme *t* dans le bloc *b1*. *tf*×*idf* correspond au nombre de lemmes communs et au nombre de fois qu'ils apparaissent dans le texte tout entier. Donc, si le score de la similarité entre deux blocs est élevé, alors non seulement les blocs ont des termes en commun, mais les termes qu'ils ont en commun sont relativement rares en ce qui concerne le reste du document.

Les valeurs de similarité sont représentées graphiquement par une courbe. Une valeur élevée de similarité indique une cohérence entre les blocs et est représentée sur la courbe par des « *peaks* ». Par contre, une valeur faible de similarité indique une frontière potentielle entre les blocs et est représentée par une vallée. L'algorithme TextTiling calcule ainsi un score pour chaque unité élémentaire en fonction de l'unité qui la suit. Le score pour chaque paire de phrase est calculé en tenant compte de l'un des critères suivants : (1) les mots communs comptabilisés par le produit scalaire entre les vecteurs représentatifs des deux phrases ; (2) le nombre de mots nouveaux dans ces phrases ; (3) le nombre de chaînes lexicales actives. La similarité entre des blocs de pseudo phrase adjacents est calculée par une mesure du cosinus.

4.1.2.3 Extraction des zones thématiques

La troisième étape concerne le processus de segmentation qui se base sur la détection des frontières dans la courbe de similarité. Le principal problème qu'on peut rencontrer avec cette méthode est l'existence de minimums locaux n'indiquant pas forcément un changement de thématique et ne représentant pas une bonne frontière. Pour résoudre ce problème un algorithme de lissage pourrait être appliqué pour lisser la courbe et éliminer les petits minimums locaux (Figure 4.5). Le score d'un segment de texte est le produit normalisé des scores de chaque paire de phrases qu'il contient (*lexical scores*). Si l'écart entre le score d'un segment et les scores de celui qui le précède et de celui qui le suit est grand, une frontière thématique est définie pour ce segment. La rupture entre deux unités documentaires est située dans une zone du texte entourée de zones présentant des valeurs de cohésion très différentes de la sienne.

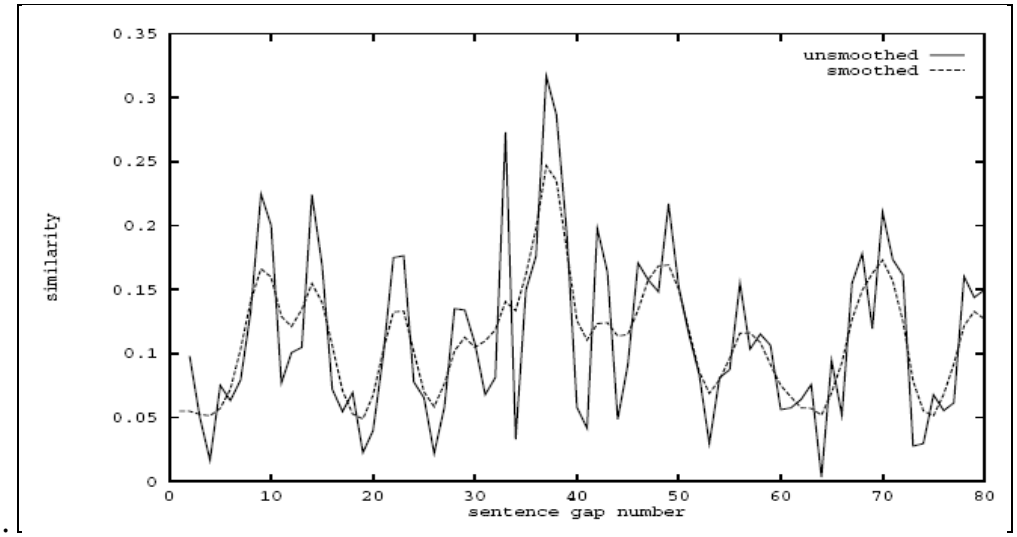


Figure 4.5 Lissage de la courbe

Si l'algorithme décrit permet en effet une segmentation thématique pertinente à l'échelle de groupes de paragraphes, rien ne garantit que deux zones non contiguës traitant d'un même thème soient caractérisées par des listes de mots identiques, ce qui rend difficile la détection de proximité thématique entre segments non consécutifs.

4.1.2.4 Exemple de segmentation

La figure 4.6 illustre un exemple de l'utilisation de TextTiling sur un document de taille 4535 Ko, intitulé « Guide sur le chauffage au bois résidentiel » extrait à partir du site : http://www.pechabot.com/documentsPDF/GuideChauffageBois.pdf, sur les ressources naturelles au Canada. La figure 4.6 montre trois exemples de segments thématiques identifiés dans ce document.

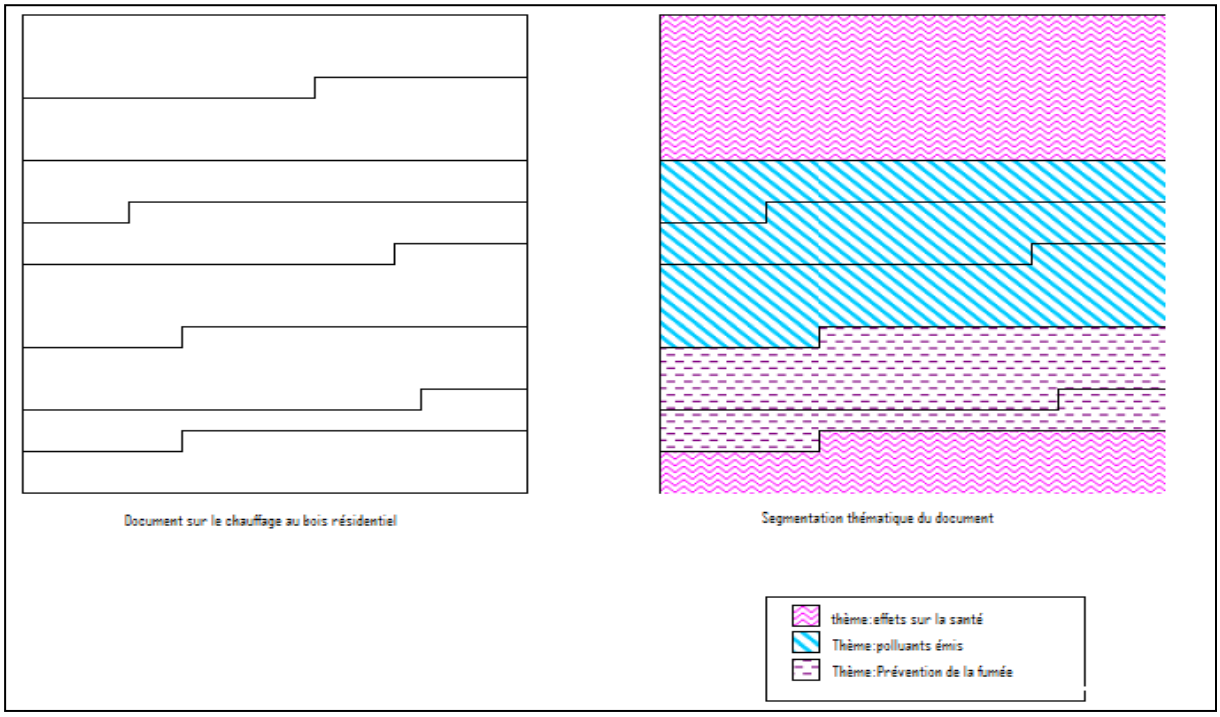


Figure 4.6 Exemple d'application de TextTiling sur un document

4.1.3 Indexation sémantique des documents sources

L'objectif de cette phase est d'extraire les termes et les concepts pertinents représentatifs du contenu de chaque document et de ses segments. Les méthodes telles que le *tf*×*idf* [Salton, 1983], [Salton, 1988] et l'analyse sémantique latente [Deerwester et al. 1990], [Landauer et al. 1998] sont fréquemment utilisées pour extraire des termes significatifs à partir d'un corpus. L'idée de ces méthodes dites discriminantes, est qu'un terme est d'autant plus important qu'il est fréquent dans un texte d'un corpus et peu présent dans les autres.

Dans le cadre de notre travail, le but est d'indexer sémantiquement les documents sources et leurs segments thématiques, ces index, qui peuvent être des termes ou des concepts, seront par la suite utilisés pour la construction de la Topic Map associée à chaque document et la génération de la Topic Map globale.

Plusieurs travaux se sont intéressés à l'indexation sémantique de documents textes, ces travaux sont le plus souvent intégrés dans le cadre de systèmes de recherche d'information, Certaines approches s'appuient sur l'utilisation de ressource termino-ontologiques comme support du processus d'indexation. [Khan, 2000] propose une méthode d'indexation des documents prenant en considération les cooccurrences des termes ainsi que la proximité sémantique de ces dernières par rapport à une ressource termino-ontologiques. [Baziz, 2005] propose d'intégrer une ontologie selon deux principes : dans l'expansion des requêtes et dans le processus de recherche d'information. Dans ce deuxième principe, il présente deux méthodes d'indentification de concepts : projection du document (et de la requête) sur l'ontologie ou l'inverse.

Dans le cadre de notre travail, nous avons choisi d'utiliser la méthode LSI *Latent Semantic Indexing Model* basée sur l'indexation sémantique latente [Landauer et al. 1998]. Nous avons également proposé d'intégrer le thésaurus dans le processus d'indexation avec LSI comme une base de connaissance pour l'identification des concepts dans les documents et leurs segments. Notre choix de LSI est justifié d'une part, par le fait que cette technique soit indépendante des langues [Dumais et al. 1996] et d'autre part, par le fait que son code source soit disponible et flexible puisque dans notre cas une étape de segmentation des documents a précédé l'étape d'indexation qui a pris en considération les documents mais aussi leurs segments thématiques. Le package LSI nous a permis cette combinaison avec le segmenteur thématique TextTiling, il nous a également donné la possibilité d'intégrer le thésaurus du domaine pour pouvoir indexer les documents et leurs segments non seulement par les termes mais aussi par les concepts.

4.1.3.1 Présentation de LSI

L'indexation sémantique latente (LSI) [Landauer et al. 1998] est une méthode factorielle, le fait que des mots apparaissent dans le même contexte veut dire qu'ils sémantiquement proches. Il existe une sémantique sous-jacente à un corpus (collection de textes). Cette méthode opère par la projection de la représentation d'un document basée sur les termes en une représentation basée sur des concepts abstraits. Le corpus est représenté sous forme matricielle. Les lignes sont relatives aux mots et les colonnes représentent les différents

contextes choisis (un document, un paragraphe, une phrase). Chaque cellule de la matrice représente le nombre d'occurrences des mots dans chacun des contextes du corpus. Deux mots proches au niveau sémantique sont représentés par des vecteurs proches. La mesure de proximité est généralement définie par le cosinus de l'angle entre les deux vecteurs. La théorie sur laquelle s'appuie LSI est la décomposition en valeurs singulières (SVD). Une matrice $A = [a_{ij}]$ où a_{ij} est la fréquence d'apparition du mot *i* dans le contexte *j*, se décompose en un produit de trois matrices USV^T.

U et *V* sont des matrices orthogonales et S une matrice diagonale. La figure 4.7 représente le schéma bien connu d'une telle décomposition où *r* est le rang de la matrice *A* qui représente le corpus d'origine de **m** lignes (mots) et **n** colonnes (contextes).

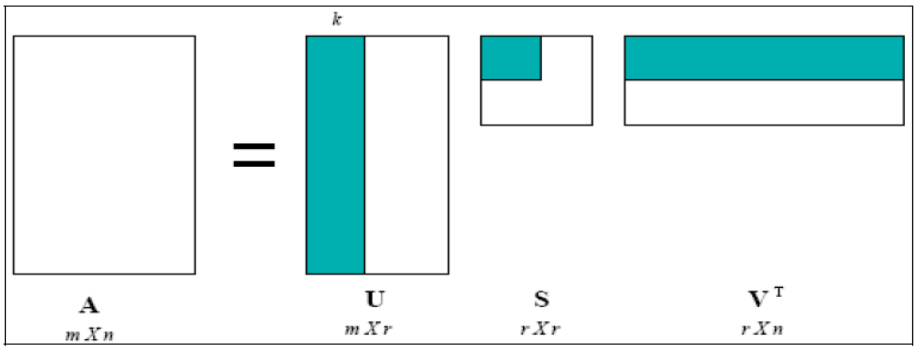


Figure 4.7 Décomposition en valeurs singulières

Soit S_k où $k < r$ la matrice produite en enlevant de *S* les *r* - *k* colonnes qui ont les plus petites valeurs singulières. Soit U_k et V_k les matrices obtenues en enlevant les colonnes correspondantes des matrices *U* et *V*. La matrice $U_k S_k V_k^T$ peut alors être considérée comme une version compressée de la matrice originale *A*. Notons que les valeurs de la matrice représentent généralement l'occurrence des termes dans les documents. Mais ces valeurs peuvent être normalisées en utilisant des pondérations comme la mesure *tf*×*idf* ou l'entropie.
En général, le module LSI pondère chaque terme candidat en fonctions de sa densité dans le document et ceci selon la formule suivante (Equation 4.2) :

$$densité(t_i) = \frac{fr(t_i)}{N} \times k \times 100$$

Equation 4.2

où $fr(t_i)$ représente la fréquence de t_i dans le document, *N* représente le nombre total de termes dans le document, et *k* représente le nombre de mots constituants le terme (si le terme est simple *k=1*; sinon - le terme est composé de plusieurs mots - alors *k* est égal à ce nombre de mots).

L'espace sémantique construit, il faut choisir une mesure appropriée afin de déterminer la proximité entre deux éléments. Les tests empiriques ont privilégié la méthode du cosinus [Zampa, 2005]. La proximité entre deux vecteurs est le cosinus de leur angle. La proximité sémantique entre deux termes, entre deux paragraphes ou entre un terme et un paragraphe est donc une valeur entre –1 et 1, où 1 indique une très forte proximité sémantique. La proximité entre deux documents d_1 et d_2 serait alors mesurée par la formule suivante (Equation 4.3) :

$$\cos(d_1, d_2) = \frac{\sum_{t \in d_1 \cap d_2} tf.idf_{t,d_1} \times tf.idf_{t,d2}}{\sqrt{\sum_{t \in d_1} tf.idf_{t,d_1}{}^2 \sum_{t \in d_2} tf.idf_{t,d_2}{}^2}}$$

Equation 4.3

4.1.3.2 Application de LSI dans notre contexte

Dans le cadre de notre travail, nous avons appliqué la méthode LSI sur les documents et les segments résultants de la segmentation avec TextTiling. Le module LSI reçoit en entrée les documents du corpus, leurs segments et le thésaurus du domaine. Il génère une matrice où les lignes sont relatives aux termes ou aux concepts du domaine et les colonnes sont représentées par les contextes, c'est-à-dire les documents et les segments. Chaque cellule de la matrice représente la pondération des termes et des concepts, pour notre cas, nous avons adopté la mesure *tf×idf* au lieu de la densité (fréquence), en effet, un terme ou un concept sera d'autant meilleur pour représenter le contenu d'un document (ou segment) s'il est à la fois fréquent dans ce document et rare dans l'ensemble des documents à analyser. La fréquence inverse du document, idf = log(N/n), où N est le nombre total de documents et n est le nombre de documents contenant le terme, vient donc modérer ou accentuer l'importance de la fréquence de chaque terme. Ainsi, ce calcul est utilisé, dans le cadre de notre analyse, afin d'extraire les termes les plus représentatifs des documents et de leurs segments.

Ainsi une valeur élevée de *tf×idf* implique que le terme est important dans le document et qu'il apparaît peu dans les autres. En général, l'utilisation des formules *tf×idf* donne de meilleurs résultats que l'utilisation de la fréquence d'occurrence seule.

Au final, les termes indexant les documents et leurs segments sont donc pondérés par leurs degrés de pertinence représentés par la mesure *tf×idf*. Pour faciliter l'élaboration de la Topic Map, nous avons choisi de pondérer les concepts par une mesure, qui suit le même principe que la mesure *tf×idf*, appelée *tof×idf* (*topic frequency, inverse document frequency*) puisque les concepts seront par la suite transformés en Topics dans la Topic Map globale.

En effet, comme nous l'avons déjà précisé dans notre approche générale (chapitre 3), les degrés de pertinence des termes et des concepts résultant de la phase d'indexation des documents seront utilisés dans le processus de construction de la TM en particulier pour pondérer les liens occurrences reliant les Topics aux documents et/ou aux segments qui leur font référence. L'idée de proposer une pondération pour les concepts, (qui représentent des Topics) et une autre pour les termes (qui peuvent être des Topics, des instances de Topics, des synonymes de Topics, des sous Topics) permet de faciliter l'affectation des poids aux liens occurrences et faciliter ainsi l'élaboration de la Topic Map. Dans ce qui suit, nous présentons en détail les deux mesures *tf*×*idf* et *tof*×*idf.*

4.1.3.3 Calcul de *tf*×*idf*

Plusieurs formules pour *tf* (Equation 4.4 et Equation 4.5) et *idf* (Equation 4.6) ont été proposées dont voici les plus souvent utilisées [Salton, 1988]:

$$tf_{ik} = \frac{f_{ik}}{\sqrt{\sum_{j=1}^{n} (f_{ij}^{\ 2})}}$$

Equation 4.4

tf_{ik} représente le poids de pertinence du terme T_k dans le document D_i

f_{ik} *représente la fréquence d'occurrence d'un terme T_k dans le document D_i*

T correspond à l'univers des termes descripteurs des documents d'un corpus D et $n = |T|$

La somme du dénominateur (normalisation) dans cette expression de *tf* (Equation 4.4) sert pour ne pas privilégier les documents longs par rapport à ceux qui sont courts mais qui sont pertinents par l'information qu'ils contiennent.

$$tf_{ik} = 0{,}5 + 0{,}5\frac{f_{ik}}{f_i}$$

Equation 4.5

f_i représente le maximum des tf_{ik} sur l'ensemble du document D_i

Cette deuxième expression de *tf* (Equation 4.5) définit également une fréquence normalisée augmentée permettant de diminuer l'effet de la variation de la taille du document en attribuant les poids les plus importants aux descripteurs les plus fréquents dans le document.

$$idf_k = \log \frac{N}{N_k}$$

Equation 4.6

idf_k représente une mesure générale de la pertinence du terme T_j pour tout le corpus
N correspond au nombre de documents dans tout le corpus
N_k correspond au nombre de documents indexés par le terme T_k

Cette dernière expression (Equation 4.6) exprime le fait que l'importance d'un terme est inversement proportionnelle à sa distribution dans tout le corpus, elle a été introduite par Sparck-Jones [Sparck-Jones, 1972].

Dans cette expression, un descripteur T_k indexant tous les documents ne permet pas à un utilisateur de retrouver spécifiquement un document, ce qui se traduit par une valeur nulle de son poids idf_k.

Dans nos travaux de thèse, nous avons adopté les formules 4.5 et 4.6 pour calculer respectivement *tf* et *idf*. Finalement, le poids s'obtient en multipliant les deux mesures (*tf*×*idf*), comme l'indique les équations suivante (Equation 4.7 et Equation 4.8) :

$$w_{ik} = \frac{f_{ik} \log \frac{N}{N_k}}{\sqrt{\sum_{j=1}^{n} (f_{ij})^2 (\log \frac{N}{N_k})^2}}$$

Equation 4.7

$$w_{ik} = (0{,}5 + 0{,}5 \frac{f_{ik}}{f_i}) \log \frac{N}{N_k}$$

Equation 4.8

Le calcul de *tf*×*idf* permet donc le filtrage et la pondération puisqu'il permet selon un certain seuil de choisir des termes qui décrivent le mieux le contenu des documents et de leurs segments.

4.1.3.4 Calcul de *tof*×*idf*

tof×*idf* est le nombre d'occurrences de chaque concept qui est calculé par la somme des occurrences de ses instances présents dans le document (ou dans le segment). Pour le calcul du poids de chaque concept, nous avons utilisé une mesure appelée *tof*×*idf*, qui est une extension de la mesure du *tf*×*idf*. La mesure du *tof*×*idf* suit la même logique que celle du

tf×*idf*. Elle combine une mesure locale, la fréquence du concept (Topic) (*tof*), et une mesure globale (*idf*) qui permet de relativiser l'importance du Topic dans l'ensemble du document à traiter.

$$w_{ij} = tof_{ij} \times idf_i$$

Equation 4.9

Où $tof_{ij} = \sum freq(T_k)$ est la fréquence du Topic i dans le document j et $freq(T_k)$ est la fréquence du kième terme du Topic i dans le document j. $idf_i = \log \frac{N+1}{df + 0{,}5}$ dénote l'importance relative du Topic dans le document. N est le nombre total de documents et df est le nombre total de documents où le Topic i apparaît. D'autres mesures ont été proposées dans la littérature. On peut citer Névéol et ses collègues [Névéol et Ozdowska, 2005] qui intègrent dans le calcul de la fréquence du concept C, la fréquence de ses sous-concepts, ou Baziz [Baziz, 2005] qui intègre le nombre de mots du terme dénotant le concept (un concept est équivalent à un nœud du thésaurus WordNet, et dans ce cas il est dénoté par un seul terme).

4.1.4 Génération du référentiel de documents

Suite aux étapes précédentes, nous avons généré le référentiel de documents, dans ce référentiel, chaque document est annoté :

- Avec des méta-informations descriptives ;
- Avec des annotations thématiques, chaque document est composé d'un ensemble de segments thématiquement homogènes, chaque segment étant annoté par un thème majeur et éventuellement des thèmes mineurs ;
- Avec des annotations sémantiques indexant les documents et leurs segments. Ces annotations sont le résultat de l'étape d'extraction des termes et des concepts pertinents avec la méthode LSI appliquée aux documents et à leurs segments.

Cette **triple annotation** de chaque document servira à la construction de la Topic Map globale multilingue qui servira d'interface médiant les besoins d'une large typologie d'utilisateurs. Les documents de notre corpus sont textuels, multilingues (initialement le français et anglais, et en perspective d'autres langues) et multi-formats (pdf, doc, txt, html,...).

4.1.4.1 Indexation descriptive

Dans cette étape, les documents sont indexés par les méta-informations descriptives (figure 4.8) regroupant le titre du document, la taille, le format, l'objet ou le scope, la pertinence, la langue, l'auteur, l'organisation, la date, l'URL, le public cible et le type du document par exemple guide, cours, article, dossier de presse, etc.

Méta-informations descriptives

Nom du document:Recherche_developpement_sur_procédés_photovoltaiques
Taille:1183KO
Type:Article
Scope:Energie solaire photovoltaique
Format:PDF
Langue:Français
Organisation: CSTB
Date de création:Mai 2008
Source/URL:http://www.cstb.fr/fileadmin/documents/telechargements/energie_solaire/Recherche_developpement_sur_procedes_photovoltaiques-mai-2008.pdf
Public cible: Consommateur (Homme du monde)

Figure 4.8 Exemple de fiche descriptive d'un document avec les méta-informations descriptives

4.1.4.2 Indexation thématique

L'indexation thématique revient à identifier le thème majeur et/ou le(s) thème(s) mineur(s) évoqué(s) dans les documents et par la suite attribuer aux documents une ou plusieurs étiquettes ou catégories thématiques en associant à chacune d'elles un degré de pertinence ou de représentativité du contenu informationnel du document. Chaque thème est défini par un label.

L'identification des thèmes dans les documents fait intervenir plusieurs domaines de recherche en particulier ceux relatifs à l'analyse thématique, à la classification et la catégorisation automatique des documents.

Ainsi, l'identification des thèmes est bien distincte de l'analyse thématique. En effet, bien qu'elle soit en partie composée de ce dernier processus, l'analyse thématique ne saurait s'y réduire. Comme l'a souligné [Chaar, 2003], « *l'identification thématique est la partie de l'analyse thématique visant à déterminer le thème d'une unité textuelle* ». L'analyse thématique peut englober en effet plusieurs autres processus. Par exemple, pour [Chaar, 2003] « *L'analyse thématique des documents consiste à segmenter les documents en régions thématiquement homogènes* ». La spécificité qui semble caractériser le processus d'analyse

thématique (par opposition à l'identification des thèmes d'un corpus) réside principalement dans l'identification de la structure et des liens possibles entre les différents thèmes.

Dans le cadre de notre travail, l'objectif est d'identifier les thèmes des différents documents du corpus et annoter ces documents en fonction de ces thèmes. Pour l'étape d'identification des thèmes d'un corpus, nous nous appuyons sur les résultats des processus de segmentation thématique de TextTiling et d'indexation sémantique avec LSI.

En effet, le processus de segmentation a permis de découper chaque document du corpus en un ensemble de segments (blocs thématiquement homogènes). Pour l'identification des principaux thèmes présents dans un document correspondant aux segments identifiés, nous nous appuyons sur la méthode d'indexation LSI qui prend entrée les documents et les segments et genre une matrice dont chaque cellule représente les $tf \times idf$ (pour les termes) et les $tof \times idf$ (pour les Topics). Ces calculs statistiques permettant de faire émerger les catégories thématiques des documents analysés. Les termes ou les concepts ayant les $tf \times idf$ ou les $tof \times idf$ les plus eleves (supérieur à un seuil fixé) par rapport à un segment de document sont attribués comme « étiquette thématique » à ce segment. Au final chaque document sera indexé thématiquement à partir des différentes thématiques associées à chaque segment dans le document en question.

4.1.4.3 Indexation sémantique

Les annotations sémantiques consistent à construire une représentation du contenu des documents du corpus par un ensemble de termes et de concepts pondérés en fonction de leur pertinence dans le document et ses segments (pour notre cas $tf \times idf$ pour les termes *et* $tof \times idf$ *pour les concepts*). Pour cette étape, nous nous appuyons sur la méthode d'indexation LSI qui prend entrée les documents et les segments et le thésaurus du domaine et genere pour chaque document et ses segments, les termes et les concepts les plus représentatifs du contenu des documents ainsi que de leurs segments pondérés avec leur degré de pertinence respectifs.

Ces termes et ces concepts serviront à construire et enrichir la Topic Map par des nouveaux Topics, instances de Topics, synonymes de Topics ou sous Topics. Les $tf \times idf$ et les $tof \times idf$ calculés sont utilisés pour étiqueter les liens entre Topics et documents ou segments et trier par la suite ces ressources selon leur importance.

Les annotations sémantiques participent à la satisfaction de plusieurs types de demandes notamment les recherches précises (chercher les documents qui évoquent un ou plusieurs Topics donnés), les recherches exploratoires (naviguer dans la hiérarchie des Topics), les

recherches connotatives (chercher les documents parlant de Topics similaires à un Topic donné).

En conclusion, le modèle d'annotation proposé s'intéresse à toutes les facettes représentatives d'un document dans l'objectif de répondre à tous les types de besoins d'information. En effet, l'annotation descriptive vise à répondre aux recherches précises et exploratoires; l'annotation conceptuelle permet de satisfaire les recherches précises, connotatives et exploratoire et finalement l'analyse thématique prend en charge les recherches thématiques et exploratoires (figure 4.9).

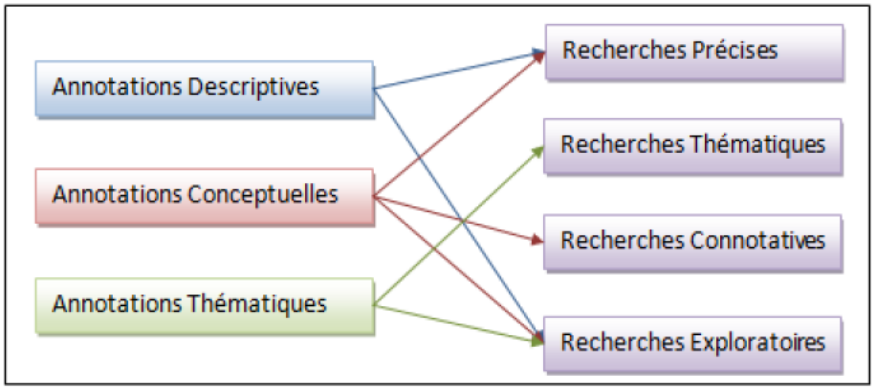


Figure 4.9 Correspondances entre les types d'annotations et les types de recherches

4.2 Construction incrémentale de la Topic Map

Dans cette section, nous décrivons l'approche ACTOM que nous avons proposée pour la construction de la Topic Map, cette Topic Map est fondée sur le méta-modèle de Topic Map que nous avons déjà défini dans la section 3 du chapitre précédent.

Comme nous l'avons déjà détaillé dans l'état de l'art, plusieurs travaux de recherche se sont penchés sur la problématique de création et de gestion d'une Topic Map. Beaucoup de propositions ont concerné la construction de Topic Maps à partir de documents textuels. Cependant, aucune d'elles ne permet de traiter un contenu multilingue. De plus, bien que les Topic Maps soient, par définition, orientées utilisation (recherche d'information), peu d'entre elles prennent en compte les requêtes des utilisateurs. Par ailleurs, aucune des approches existantes n'exploitent plusieurs sources d'informations pour la construction d'une Topic Map.

Notre approche ACTOM est **incrémentale** [Ellouze et al. 2009b], elle est basée sur un **processus automatisé** et **évolutif** qui prend en compte des documents **multilingues** et l'évolution de la Topic Map selon **le contenu et l'usage** [Ellouze et al. 2009a]. Nous prenons comme entrée un référentiel de documents textuels, dans ce référentiel, les documents sont

segmentés thématiquement et indexés sémantiquement avec pour chaque document la liste de ses segments, chaque segment représente une thématique abordée dans le document et chaque document ainsi que ses segments sont indexés par une liste de termes et de concepts représentatifs de leur contenu. La Topic Map résultante est ensuite enrichie à partir des liens ontologiques présents dans le thésaurus du domaine et à partir de deux ontologies générales qui serviront à rajouter de nouveaux synonymes du langage commun aux Topics existants et de nouveaux liens entre ces derniers.

Pour représenter l'usage dans la Topic Map, nous proposons aussi de prendre en compte un ensemble de scénarios d'usage à travers la mise en œuvre de liens d'usage entre les questions potentielles extraites des sources d'interrogations disponibles relatives aux documents sources telles que les FAQ et les réponses associées.

Intuitivement, il s'agit, dans ACTOM, de construire de façon incrémentale une Topic Map TM_i correspondante à un ensemble de documents $D= \{d_1, d_2, \ldots., d_i\}$ en fusionnant la Topic Map TM_{i-1} correspondante à l'ensemble de documents $D-\{d_i\}$ avec la Topic Map associée au document d_i. Chaque phase permettant de construire la Topic Map correspondante à un document d_i utilise comme source aussi bien le document lui-même mais aussi le thésaurus, les deux ontologies générales WordNet et WOLF, et un ensemble de questions correspondantes à ce document et extraites de sources d'interrogations.

ACTOM trouve aussi son utilité toutes les fois où le contenu à organiser est enrichi de un ou plusieurs documents, ou encore lorsque l'on souhaite introduire d'autres questions fréquemment posées. Elle contribue dans le processus d'évolution d'une Topic Map.

La figure 4.10 présente notre approche ACTOM de construction incrémentale de Topic Map.

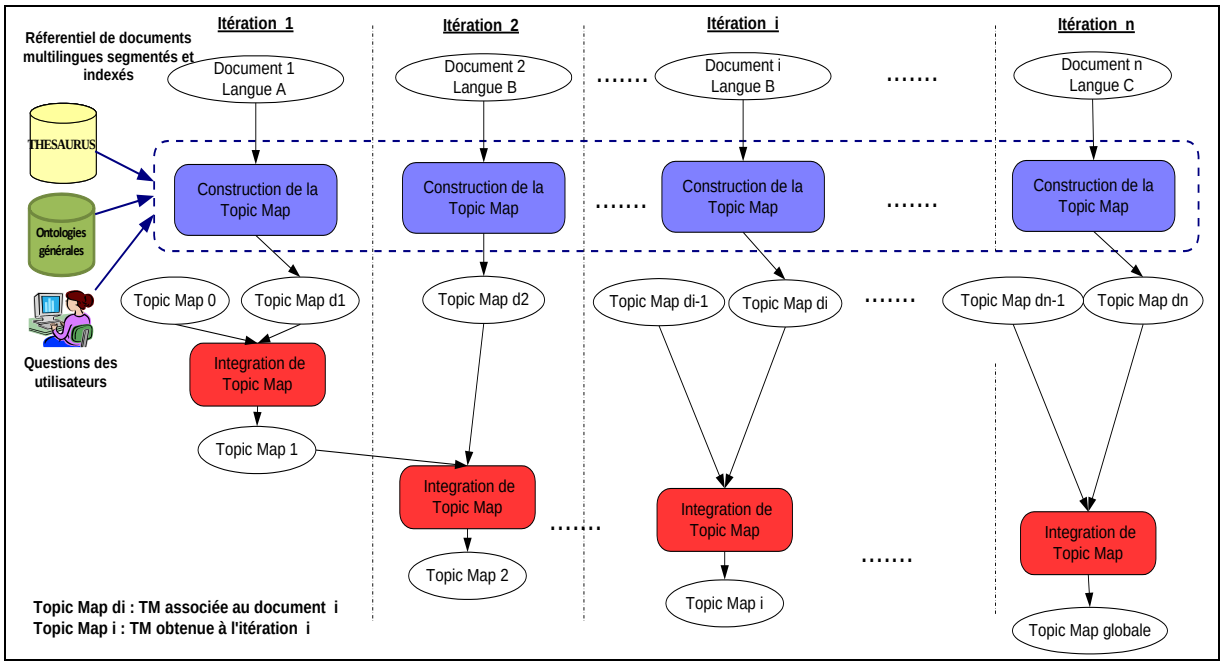


Figure 4.10 Approche ACTOM de construction incrémentale de Topic Map

L'algorithme général de l'approche ACTOM est le suivant :

Entrée *: Un référentiel de documents multilingues, un thésaurus du domaine, deux ontologies linguistiques générales, et des scénarios d'usage (questions des experts, requêtes des utilisateurs, FAQ, historique, discussions téléphoniques, consultations directes avec les travailleurs du domaine, ...)*

Sortie *: une Topic Map globale, multilingue, enrichie et annotée.*

Action 1. *Construire la racine de la Topic Map globale Topic Map 0. La racine correspond au Topic portant le nom du domaine sous les différentes langues.*

Action 2. *Traiter les sources d'interrogations et constituer, pour chaque document, un ensemble de questions potentielles.*

Pour *chaque document i du référentiel* ***faire***

Action 3. *Extraire les Topics et les associations entre ces Topics à partir du document i.*

Action 4. *Enrichir la Topic Map en ajoutant d'autres liens ontologiques et structurels à partir du thésaurus*

Action 5. *Enrichir la Topic Map par des synonymes de Topics et de nouveaux liens structurels entre ces derniers en utilisant deux ontologies générales WordNet pour l'anglais et WOLF pour le français.*

Action 6. *Enrichir la Topic Map à partir des questions potentielles correspondantes à ce document.*

Action 7. *Validation de la Topic Map d_i résultante par les experts*

Action 8.*Intégrer la Topic Map d_i associée au document i avec la Topic Map globale pour obtenir une Topic Map i correspondante à l'itération i.*

Fin pour

Notre approche de construction de Topic Map est semi-automatique, en effet, la validation de la Topic Map résultante suite à chaque action consiste à préciser la sémantique de certains liens ou encore à supprimer ou à ajouter des liens et/ou des Topics. Elle nécessite la collaboration d'un ou plusieurs experts du domaine.

Pour **la construction de la Topic Map d_i** associée au document i (figure 4.11), notre approche est basée sur quatre phases principales : **La première phase consiste en la construction d'une Topic Map** à partir d'un document du référentiel, cette phase consiste à ajouter des nouveaux Topics que nous avons appelés (dans notre méta-modèle de Topic Map) des Topics thèmes correspondants aux segments thématiques résultants de l'application d'un algorithme de segmentation thématique sur les documents sources.

Dans un deuxième temps, il s'agit d'enrichir cette Topic Map par les termes et les concepts représentatifs de chaque document et de ses segments thématiques résultant de l'indexation sémantique des documents, les concepts sont ajoutés comme des Topics (concepts du domaine) et les termes peuvent être rajoutés comme des instances de Topics, des sous Topics ou des synonymes de ces derniers. Cette étape inclut aussi l'identification des associations entre les Topics. La Topic Map résultante est ensuite **enrichie avec d'autres liens ontologiques et des synonymes de Topics extraits à partir du thésaurus** du domaine, le résultat de cette étape est une Topic Map du domaine représentant les concepts du domaine avec leurs synonymes extraits du thésaurus contenant une première hiérarchie de Topics (Topics reliés par des liens ontologiques et structurels).

La troisième phase s'intéresse à **l'enrichissement de la Topic Map par les synsets et les liens présents** dans WordNet (pour l'anglais) et WOLF (WordNet libre du français)) en ajoutant des synonymes aux Topics déjà présents dans la Topic Map et des liens structurels entre eux.

La dernière phase consiste à enrichir la Topic Map par les liens d'usage c'est-à-dire toute question potentielle (i.e. phrase en langage naturel) représentée sous forme de Topic (que nous avons appelé Topic question dans notre méta-modèle) est reliée à chacun des mots clés la constituant par un lien d'usage de type « est composé de » et aux Topics qui permettent d'y répondre par un lien d'usage de type « répond à ». Le stockage des liens « est-composé-

de » d'une question vers les termes qui la composant permet d'une part une recherche par navigation et d'autre part une recherche automatique de « question proche ». La Topic Map obtenue **associée au document d_i** est ensuite fusionnée avec la Topic Map globale.

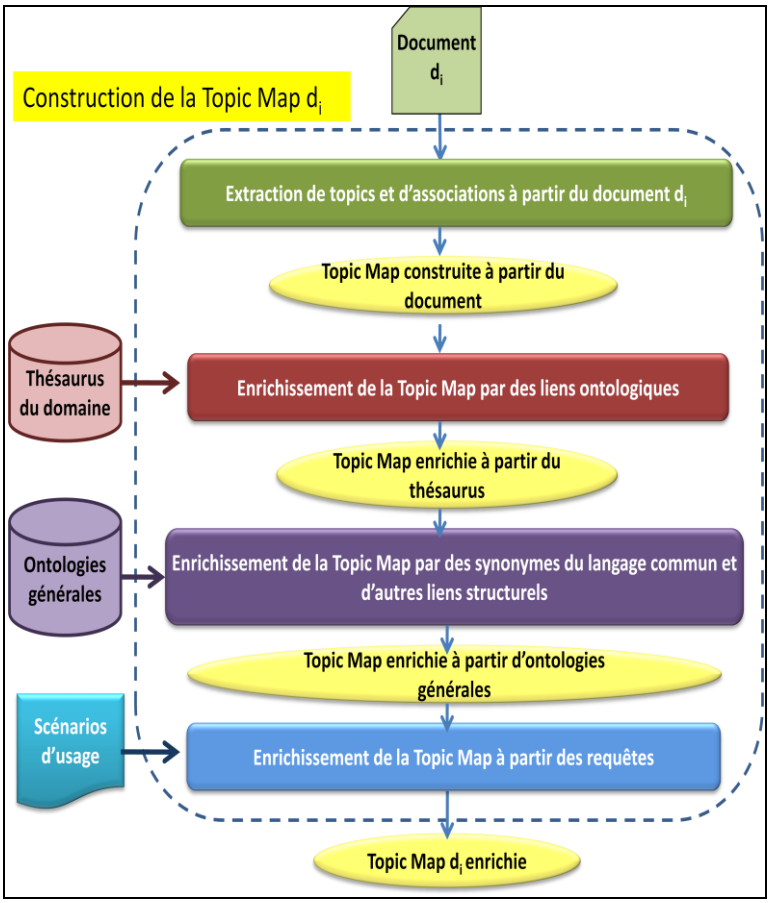


Figure 4.11 Les étapes de construction de la Topic Map associée à un document du référentiel

Au final, nous procédons à la génération **de la Topic Map globale, enrichie et annotée** par les documents et leurs segments thématiques en listant en face de chaque Topic, la liste des documents et des segments traitant de ce Topic triée par degré de pertinence décroissante, de même pour les instances de chaque Topic et les sous Topics, constituant ainsi **un espace sémantique de navigation** pour la visualisation des résultats et l'interaction avec les utilisateurs.

ACTOM exploite des documents de différentes langues. Elle permet ainsi de produire une Topic Map où les concepts sont décrits dans plusieurs langues tout en prenant en charge une des spécificités du multilinguisme qui est l'absence éventuelle de termes sémantiquement équivalents d'une langue à une autre ; ce qui est assez fréquent lorsque les contenus sont issus de différentes cultures. Toutes les étapes citées précédemment seront appliquées pour toutes les facettes linguistiques.

4.2.1 Extraction de Topics et d'associations à partir d'un document

L'objectif de cette phase est d'extraire des Topics et des associations entre ces Topics à partir d'un document du référentiel. Comme mentionné dans notre méta-modèle de Topic Map, les associations que nous souhaitons extraire correspondent aux liens ontologiques et structurels (lien « est un », lien « partie de ») et aux liens sémantiques associés au domaine de la Topic Map. Pour cela, nous nous appuyons, dans un premier temps, sur les résultats de la segmentation thématique avec l'algorithme TextTiling et de l'indexation sémantique avec la méthode LSI appliqués aux documents du référentiel pour l'extraction de Topics. Dans un second temps, nous proposons d'ajouter des liens ontologiques et associatifs entre les Topics extraits en utilisant plusieurs techniques que nous explicitons dans ce qui suit.

Nous commençons tout d'abord par présenter l'algorithme général que nous proposons permettant d'identifier les Topics et les liens entre eux, ensuite nous détaillons ses différentes étapes.

Cet algorithme se résume en ce qui suit :

Entrées : *Document d_i indexé et segmenté du référentiel,* *Matrice LSI termes et concepts/contextes pondérée avec contexte document ou segment.*

Sortie : *Topic Map construite à partir du document d_i*

Pour *chaque terme t_i et concept c_i indexant le document et ses segments résultant de la matrice LSI* ***faire***

1) ***Ajouter un nouveau Topic T_i*** *correspondant à ce terme (ou concept) dans les deux langues.*
2) ***Placer T_i*** *dans la Topic Map*

   {ajouter un lien ontologique entre Ti et T avec Ti ∈ {Topics extraits de documents}

   Et T ∈ {Topics thèmes des segments}
3) ***Ajouter les liens occurrences*** *pondérés avec les tf×idf (dans le cas où le Topic correspond à un terme) et tof×idf (dans le cas où le Topic correspond à un concept) reliant chaque Topic T_i aux documents et aux segments à partir desquels il est extrait*

Fin pour

Valider la Topic Map par des experts du domaine

1) *Préciser la sémantique de certains liens*
2) *Ajouter des liens et/ou des Topics*

4.2.1.1 Extraction de Topics

Le référentiel de documents contient pour chaque document, la liste des segments thématiques présents dans ce document et pour chaque segment et document, la liste des termes et des concepts représentatifs de leurs contenus.

La segmentation thématique des documents avec TextTiling produit pour chaque document, un ensemble de segments thématiquement homogènes et l'application de la méthode LSI sur les documents et leurs segments produit une matrice terme et concepts/contexte, le contexte peut être un document ou un segment et donne pour chaque document et pour chaque segment la liste des termes et des concepts pertinents avec leurs poids respectifs *tf×idf et tof×idf.*

La figure 4.12 illustre l'utilisation de la matrice termes et concepts/contextes de LSI pour l'identification des Topics avec leurs poids respectifs.

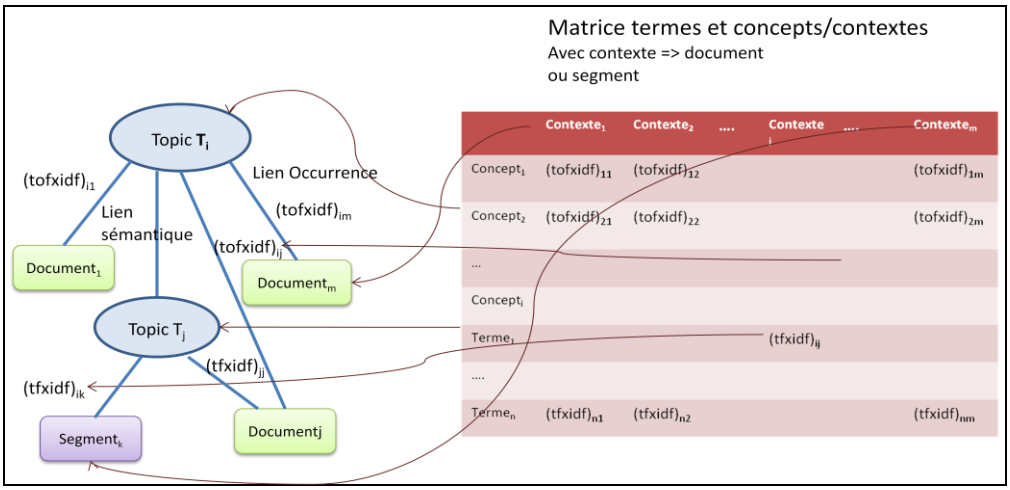


Figure 4.12 Utilisation de la matrice termes et concepts/contexte de LSI pour la construction de la Topic Map

A ce stade, la Topic Map résultante ne contient que des Topics correspondant aux concepts et aux termes indexant le document et ses segments, en se basant sur le méta-modèle de Topic Map que nous avons définit dans le chapitre précédent, ces Topics ajoutés peuvent être des **Topics thèmes**, des **Topics concepts** du domaine, des **sous Topics**, des **instances de Topics**. Il s'agit maintenant de relier ces Topics entre eux et de les organiser, ce ci revient à ajouter **des liens ontologiques** (est un, partie de) et des liens associatifs (relatifs au domaine) entre ces Topics.

4.2.1.2 Ajout des liens ontologiques et associatifs

L'objectif de cette étape est d'ajouter les liens structurels et associatifs entre les Topics présents dans la Topic Map.

Le résultat de la segmentation de l'algorithme TextTiling appliquée à notre corpus, produit, **pour chaque document**, un ensemble de segments thématiques où chacun représente un thème majeur et éventuellement un ou plusieurs thèmes mineurs. Par conséquent, notre problème revient à trouver quel type de lien il y a entre un **Topic thème d'un segment et un Topic du document**.

Plus précisément, nous définissons les Topics thèmes comme l'ensemble des concepts les plus représentatifs du contenu d'un segment (ayant les *tof*×*idf* le plus élevés), pour cela nous avons fixé **un seuil** pour la valeur de *tof*×*idf* indiquant les concepts choisis comme Topics thèmes. Par exemple les concepts ayant un ***tof*×*idf* > 0,8** dans le segment sont considérés comme des Topics thèmes puisqu'ils représentent le plus le contenu du segment.

Ces Topics thèmes sont, bien évidement, reliés aux Topics extraits à partir du document (à partir duquel on a obtenu le segment), ces liens peuvent être des liens de type « est un » (figure 4.13), « partie de » (figure 4.14) ou bien d'autres liens sémantiques tels que « est traité par », « est relié à », etc.

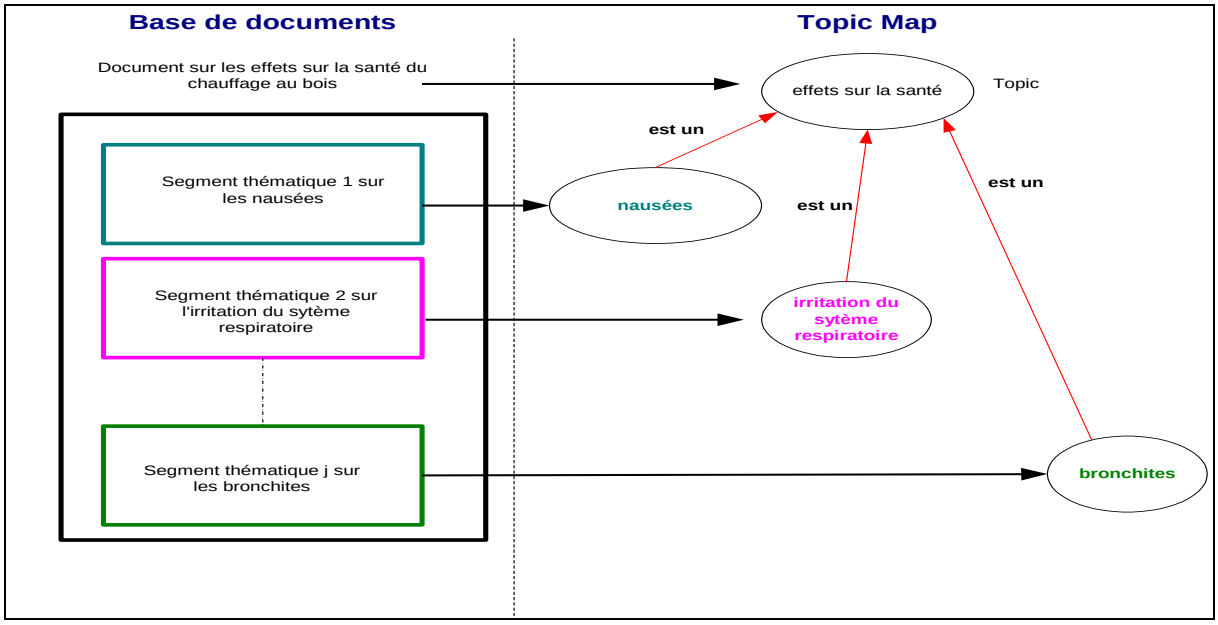


Figure 4.13 Ajout d'un lien « est un » entre le Topic extrait du document et les Topics thèmes dans le document

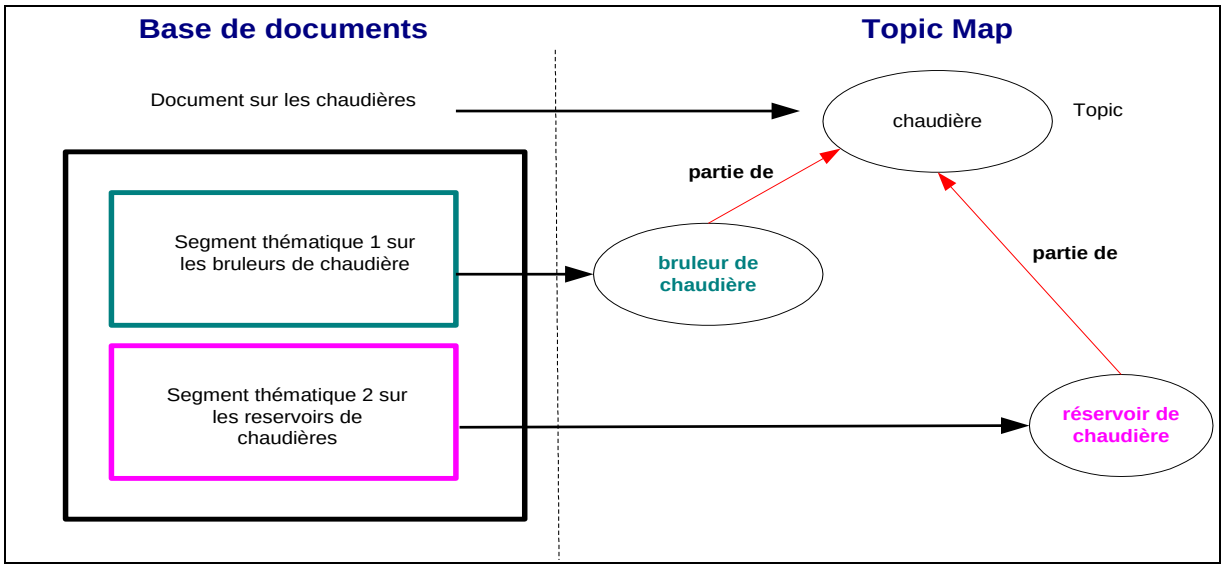


Figure 4.14 Ajout d'un lien « partie de » entre le Topic extrait du document et les Topics thèmes dans le document

Pour résoudre ce problème d'ajout de liens entre Topics, plusieurs solutions ont été proposées dans la littérature, en effet, c'est un problème très ancien, il a été longuement traité en particulier dans le contexte de construction et d'enrichissement d'ontologies à partir de textes où il existe plusieurs méthodes de détection de relations entre les concepts d'une ontologie.

Dans le cadre de notre travail, nous proposons de **combiner certaines des solutions existantes** pour trouver le lien entre un Topic thème d'un segment et un Topic extrait du document. Notre démarche inclut **les alternatives qui seront détaillées dans la suite.**

Nous procédons, en premier lieu, à une **analyse du langage naturel** pour chercher s'il y a un lien de type « est un » ou un lien « partie de » entre le Topic du segment et le Topic du document. Nous pouvons parfois le trouver dans les phrases du document, il s'agit de chercher s'il y a des phrases avec le verbe être par exemple : « une chaudière à bois est une chaudière … », ou des phrases avec « avoir », ou « est composé de », par exemple « une chaudière est composée d'un bruleur et d'un réservoir … » ;

Il existe plusieurs outils et plateformes pour l'extraction des termes et des relations à partir de documents texte, Dans le cadre de notre travail, notre choix s'est porté sur la plate-forme GATE qui a l'avantage de proposer une solution générique pour le traitement linguistique des documents textuels à travers un ensemble de modules paramétrables. Ces modules peuvent être combinés, enrichis et adaptés selon nos besoins. GATE propose un module appelé *gazetter*, pour la reconnaissance d'entités nommées à partir de dictionnaires préétablis. Ces dictionnaires peuvent être enrichis avec les termes du domaine étudié. De

plus, GATE nous donne la possibilité d'intégrer une ressource externe représentée par le thésaurus du domaine pour construire la hiérarchie de Topics et ajouter d'autres liens ontologiques dans la Topic Map. Cette étape d'enrichissement à partir du thésaurus sera discutée dans la prochaine section ;

La **deuxième solution** consiste à rechercher une première structure de liens « est un » par un **moteur de recherche** tel que Google, par exemple, si on lance une recherche automatique sur Internet et on saisit la phrase suivante « la pompe à chaleur est un équipement de chauffage » alors on aura de nombreuses réponses et d'autres renseignements tels que les types de pompe à chaleur, les performances, le coût d'installation, ou bien des adjectifs (exemple : la pompe à chaleur est un équipement de chauffage géothermique) alors que si on lance la recherche de cette phrase : « un panneau solaire est une maladie », dans ce cas, on n'aura aucun résultat.

Nous pouvons également rechercher d'autres liens, par moteur de recherche, tels que les liens « réalisé par », « est composé de ».

Une dernière alternative consiste à utiliser notre algorithme [Lammari et Métais, 2004] fondé sur les contraintes d'existence mutuelles, exclusives et conditionnelles.

Le principe de cet algorithme est le suivant : compte tenu d'un ensemble de contraintes entre les mots clés décrivant des concepts, nous déduisons une hiérarchie de concepts. Trois types de contraintes sont définis : contraintes d'existence mutuelles, exclusives et conditionnelles [Lammari et al. 2008]. Une contrainte d'existence mutuelle, définie entre deux mots clés x et y d'un concept c, dénotée par $x \leftrightarrow y$, décrit le fait que chaque instance associée au concept c peut être décrite simultanément par les mots clés x et y. La contrainte d'existence exclusive, définie entre deux mots clés x et y d'un concept c, dénotée $x \nleftrightarrow y$, décrit de fait que chaque instance associée au concept c décrite par le mot clé x, ne peut pas être décrite par le mot clé y et inversement. Enfin, la contrainte d'existence conditionnelle, dénotée $x \mapsto y$, introduit le fait que chaque instance du concept c décrite par le mot clé x doit être aussi décrite par le mot clé y (l'inverse n'est pas toujours vrai). Nous utiliserons également ces contraintes dans l'étape d'enrichissement de la Topic Map à partir du thésaurus qui sera détaillée dans la prochaine section.

Nous commençons dans un premier temps par découvrir les mots clés caractérisant les deux Topics et par la suite regarder s'il y a inclusion entre eux. Pour la recherche de mots clés caractérisant un Topic, plusieurs solutions peuvent être adoptées :

a) Effectuer une recherche automatique en utilisant un moteur de recherche sur internet ;

b) Utiliser une méthode statistique des mots associés dans les documents par exemple : Recherche des documents sur « chauffage », génération de la matrice d'occurrence des termes de ces documents, et sélection des termes les plus fréquents (pompe à chaleur, pollution, économie d'énergie, etc.). Ces termes peuvent être des mots clés ou des Topics reliés sémantiquement au Topic « chauffage » ;

c) Compléter cette recherche par une analyse linguistique des documents avec les phrases « a », « possède », etc, en utilisant un outil ou une plateforme d'analyse du langage naturel tel que GATE.

A partir d'une liste de mots clés, il existe des techniques pour créer la hiérarchie « Is-A » entre les Topics par exemple les treillis de Galois qui se basent sur un regroupement conceptuel des données ayant des propriétés communes. Pour notre cas, nous utilisons les contraintes d'existence mutuelles, exclusives et conditionnelles citées précédemment pour voir s'il y a inclusion entre les Topics et les algorithmes d'extraction, de normalisation et de translation proposés par [Lammari et Métais, 2004] pour la création des hiérarchies « Is-A », la figure 4.15 illustre un exemple d'application de ces algorithmes :

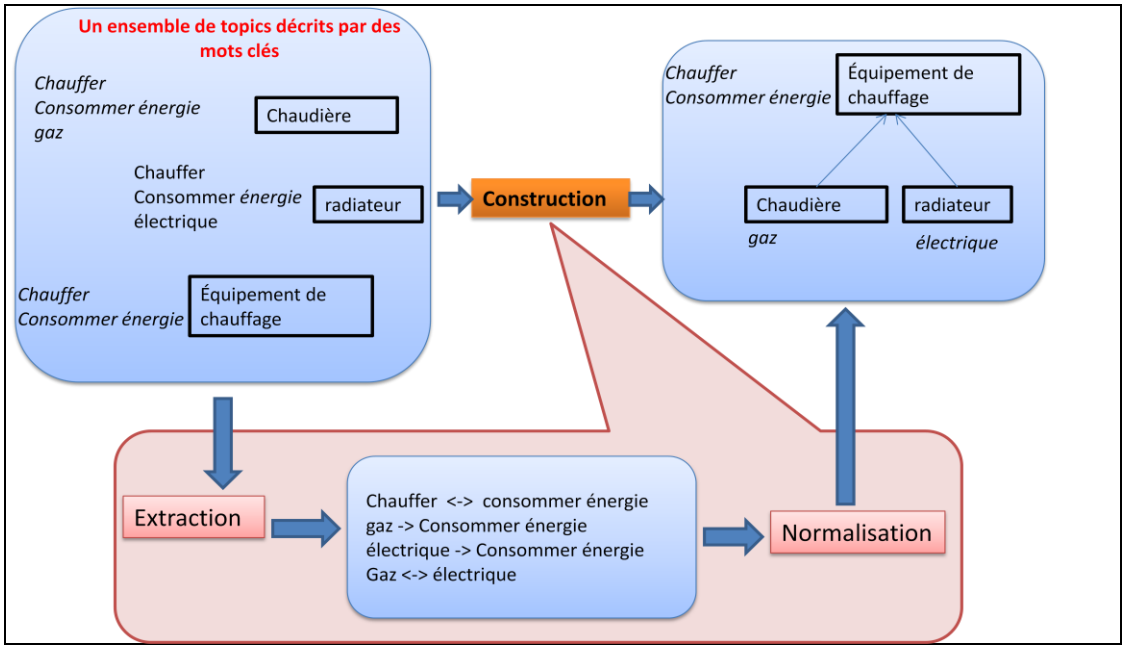


Figure 4.15 Exemple de construction d'une hiérarchie « est un » en utilisant des mots clés communs entre Topics

Si après toutes les alternatives citées précédemment, nous ne trouvons pas le type de lien recherché alors dans ce cas nous décidons d'appeler le lien **« relié à »** entre le Topic thème du segment et le Topic du document puisque cette notation est générale et pourrait convenir à n'importe quelle situation (figure 4.16).

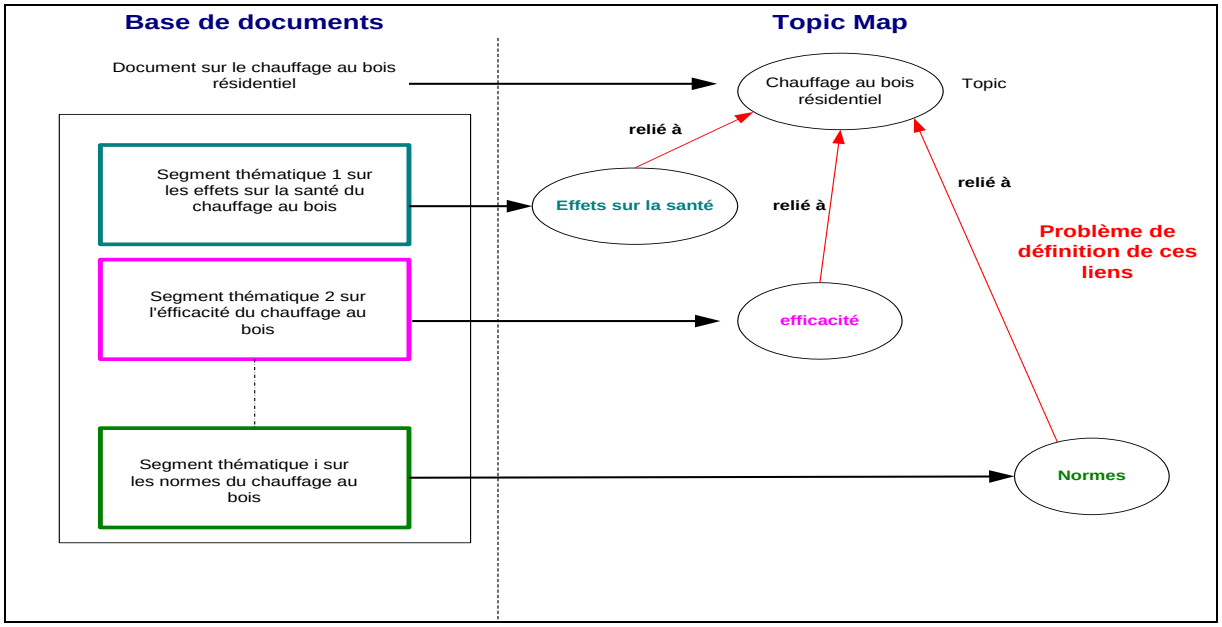


Figure 4.16 Ajout d'un lien « relié à » entre le Topic extrait du document et les Topics thèmes dans le document

Dans la Topic Map générée à partir du document, les Topics peuvent être des Topics de premier niveau, par exemple les Topics thèmes, ou bien des Topics de deuxième niveau c'est-à-dire un sous Topic ou une instance de Topic ou bien un synonyme d'un Topic existant.

Comme nous l'avons présenté dans notre méta-modèle, cette information est sauvegardée dans une **méta-propriété** du Topic, indiquant **le niveau** auquel il appartient dans la Topic Map, nous avons également défini une autre méta-propriété du Topic, qui est sa note, qui renseigne sur son usage et son importance dans la Topic Map. Ces deux méta-propriétés seront utilisées dans l'étape d'élagage de la Topic Map qui fera l'objet du chapitre suivant.

4.2.1.3 Ajout des liens d'occurrences

Le but de cette phase est **l'ajout des liens occurrences** entre les Topics et leurs ressources. Pour cela, nous explorons la matrice générée par la combinaison de LSI et TextTiling comme le montre la figure 4.12 et nous **relions automatiquement** chaque nouveau Topic ajouté aux documents et aux segments à partir desquels il est extrait par **des liens occurrences** pondérés par les *tf*×*idf* ou *tof*×*idf*. Cette mesure nous permet de classifier

les documents et les segments par ordre de pertinence dans le cas où plusieurs documents sont reliés à un même Topic.

Concrètement, cette mesure est un **attribut dans la facette,** en plus des autres attributs langue, type, taille et date du document, liée à l'occurrence et permettant de décrire le document, nous détaillerons la notion de facette et son utilisation pour notre cas dans le dernier paragraphe de cette section « Annotation de la Topic Map enrichie à partir des documents et de leurs segments ».

La figure 4.17 illustre un exemple de construction d'une Topic Map à partir d'un document du référentiel :

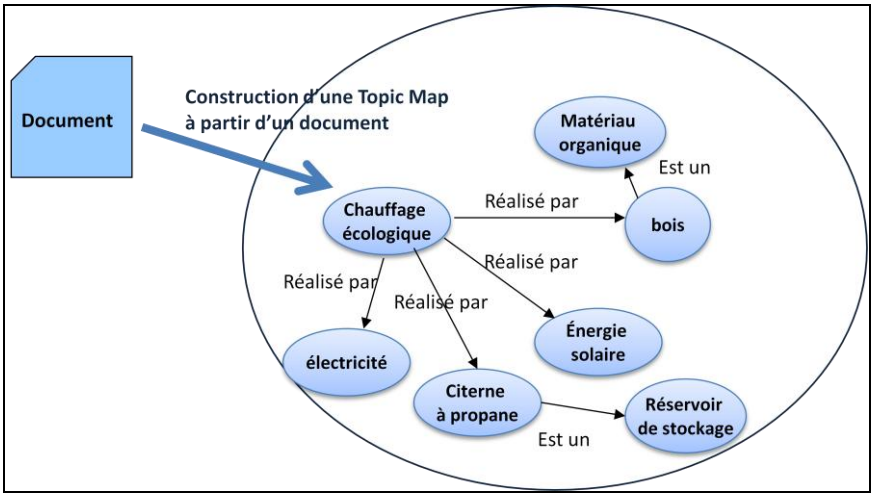


Figure 4.17 Exemple de Topic Map construite à partir d'un document du référentiel

Pour enrichir la Topic Map, nous proposons d'utiliser le **thésaurus** du domaine pour ajouter d'autres liens ontologiques, de type « est un » ou « partie de » ou bien des liens de synonymie entre deux Topics ou encore des synonymes de Topics dans une langue différente que celle qui existe déjà dans la Topic Map.

4.2.2 Enrichissement de la Topic Map par des liens ontologiques à partir du thésaurus

L'objectif de cette étape est d'enrichir les Topics issus des documents en leur ajoutant d'autres liens ontologiques et structurels. Pour cela, nous proposons d'utiliser les liens entre les termes qui existent dans le thésaurus [Ellouze et al. 2009a].

Les thésaurus sont des vocabulaires contrôlés de termes représentant généralement un domaine particulier gérant des relations hiérarchiques, associatives et d'équivalence. Dans la littérature, plusieurs travaux de recherche décrivent des approches de construction d'ontologies à partir de thésaurus. Parmi elles, on peut citer celle d'Hernandez [Hernandez,

2005] qui propose de transformer un thésaurus en une ontologie de domaine. Les auteurs définissent une méthode qui permet d'extraire les éléments du schéma conceptuel de l'ontologie à partir d'un thésaurus et de documents textuels. Leur approche est fondée sur un ensemble de règles de transformation. Ces règles exploitent les liens «est plus spécifique que», (EPS) «est plus générique que» (EPG), « Utiliser plutôt » (UP) et « Utiliser pour désigner » (UPD) d'un thésaurus pour générer les concepts de l'ontologie, les labels associés à chacun de ces concepts et la hiérarchie de concepts.

Notre démarche d'enrichissement de la Topic Map par d'autres liens ontologiques est proche de celle des travaux d'Hernandez [Hernandez, 2005] dans le sens où nous exploitons aussi les liens existant dans un thésaurus pour produire des liens ontologiques. En effet, les Topics sont organisés hiérarchiquement à partir de la relation « est un ». Ces liens hiérarchiques entre Topics sont directement issus des liens «est plus spécifique que» (EPS) et «est plus générique que» (EPG) explicitement présents dans le thésaurus. Nous utilisons aussi les relations de type « Utiliser plutôt » (UP) et « Utiliser pour désigner » (UPD) pour ajouter un nouveau nom au Topic ou encore regrouper des Topics en un seul Topic.

Cependant la démarche que nous utilisons est algorithmique. Elle s'exécute en deux temps. Dans un premier temps nous exploitons les liens UP et UPD entre termes du thésaurus pour regrouper des Topics. Dans un second temps, nous organisons les Topics en hiérarchies. La figure 4.18 illustre notre démarche d'enrichissement de la Topic Map à partir du thésaurus :

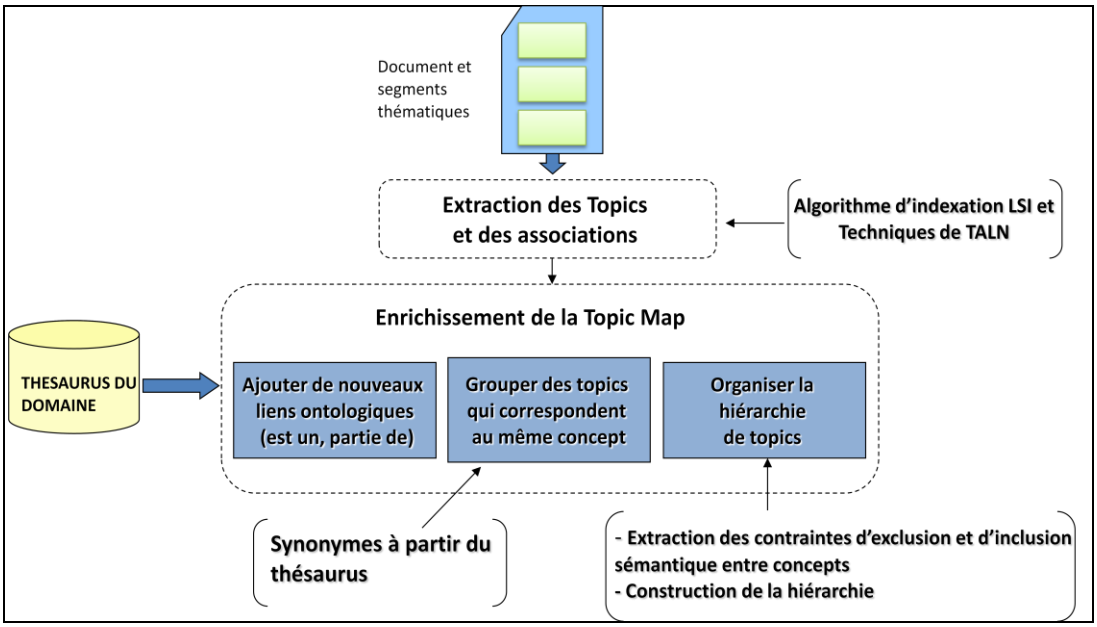


Figure 4.18 Enrichissement de la Topic Map à partir du thésaurus

Soit *SYN(Terme$_i$)* l'ensemble constitué du Terme$_i$ et de tous les termes auquel est lié Terme$_i$ via le lien UP ou UPD dans le thésaurus. Notons par TP_i le terme préféré dans SYN(Terme$_i$). L'algorithme permettant de regrouper des Topics ou d'attribuer plusieurs noms à un Topic est le suivant :

Pour** tout Topic T_i **faire

***Si** T_i est dans le thésaurus*

***Alors** construire SYN(T_i), TP_i et FILS(TP_i)*

T_i aura pour nom de base TP_i et pour autres noms les noms se trouvant dans SYN(T_i)

Fin Pour

Pour** tout couple T1 et T2 de Topic **faire

***Si** SYN(T1) = SYN(T2)*

***Alors** regrouper T1 et T2 en T3 tel que le nom de base de T3 soit TP_1, et les autres noms soient ceux faisant partie de SYN(T1). Ce regroupement implique bien sûr le regroupement de toutes les autres caractéristiques de T1 et T2 (Rôles et occurrences).*

Fin Pour

Les Topics doivent maintenant être organisés hiérarchiquement à partir des liens « est un ». Afin d'extraire ce type de lien du thésaurus, les relations « est plus spécifique que » et « plus générique que » du thésaurus sont prises en compte. L'ensemble de ces relations définies pour les termes, devenus maintenant labels d'un Topic, est retenu comme liens candidats pour représenter des liens « est un » entre le Topic et le Topic auquel se rapporte le terme lié dans le thésaurus. Les liens candidats doivent ensuite être analysées avec précaution car ils peuvent englober des relations de type « partie de » ou « instance de ». Nos travaux ne proposent pas de méthode automatique pour réaliser cette désambiguïsation. Les experts du domaine doivent intervenir pour spécifier ces types de relations manuellement.

Pour l'organisation des Topics, nous utilisons deux techniques proposées dans [Lammari et Métais, 2004] pour la construction et la maintenance d'ontologie. La première technique, nommé «translation», permet de traduire une hiérarchie de concepts en contraintes entre concepts. La seconde technique, nommée «normalisation», permet, compte tenu d'un ensemble de contraintes entre concepts, de déduire une hiérarchie de concepts.

Dans le contexte particulier de la construction de Topic Maps, nous considérons deux types de contraintes : (a) la contrainte d'exclusion de sémantique, notée $\nleftrightarrow$, qui définie entre

deux termes T1 et T2, exprime le fait que T1 et T2 n'ont pas la même sémantique et la contrainte d'inclusion de sémantique, notée $\rightarrow$, qui définie d'un terme T1 vers un terme T2 (T1 $\rightarrow$ T2) exprime le fait que la sémantique de T1 contient celle de T2 (l'inverse n'étant pas vrai).

Pour extraire les contraintes entre termes, nous appliquons la technique de translation sur le thésaurus. Nous nous basons pour cela sur les liens «est plus générique que» explicitement présents dans le thésaurus. Cette technique s'appuie sur les trois règles de translation suivantes :

- *R1* : si, dans le thésaurus, un terme T2 est plus générique qu'un terme T1, alors (T1 $\rightarrow$ T2) ;
- *R2* : si, dans le thésaurus, il n'existe pas un terme T3 comme nœud de départ de deux chemins l'un allant vers T1 et l'autre allant vers T2 tel que ces chemins soient constitués uniquement de lien «est plus générique que » alors T1 $\nleftrightarrow$ T2 ;
- *R3* : si, dans le thésaurus, deux termes T1 et T2 sont tous les deux reliés à un terme T3 par le lien «est plus générique que» alors T1, T2$\rightarrow$T3.

Soit *FILS(T)* l'ensemble de termes du thésaurus récoltés lors du parcours de tous les chemins constitués uniquement de liens EPG et ayant pour nœud de départ T, l'algorithme permettant d'extraire les contraintes entre Topics à partir du thésaurus se résume en ce qui suit :

Pour *tout Topic T1* ***faire***

Pour *tout T* $\in$ *FILS(T1)*

T$\rightarrow$*T1*

Fin Pour

Fin Pour

Pour *tout couple T1 et T2 dans la Topic Map* ***faire***

Si *FILS(T1)*$\neq$*FILS(T2)* ***alors*** *T1*$\nleftrightarrow$*T2*

Sinon *si FILS(T1)*$\cap$*FILS(T2) = {T3}* ***alors*** *T1, T2*$\rightarrow$*T3*

Fin Pour

Une fois les contraintes déduites, nous appliquons sur notre Topic Map la technique de normalisation afin d'organiser sous forme d'hiérarchie l'ensemble des Topics de notre Topic Map. Intuitivement, il s'agit :

- dans un premier temps, de construire un graphe complet non orienté ayant pour nœuds l'ensemble des Topics ;

- dans un second temps, d'éliminer tout lien entre deux Topics T1 et T2 si T1 ↮ T2 ;
- dans un troisième temps, de déduire toutes les cliques possibles respectant l'ensemble des contraintes d'inclusion et d'organiser ces cliques en un graphe d'inclusion ;
- et enfin d'inverser les liens d'inclusion pour obtenir les liens hiérarchiques entre Topics et d'éliminer les redondances en supprimant tout Topic d'une sous-classe se trouvant dans sa super-classe.

La figure 4.19 montre un exemple de déroulement de l'algorithme de normalisation :

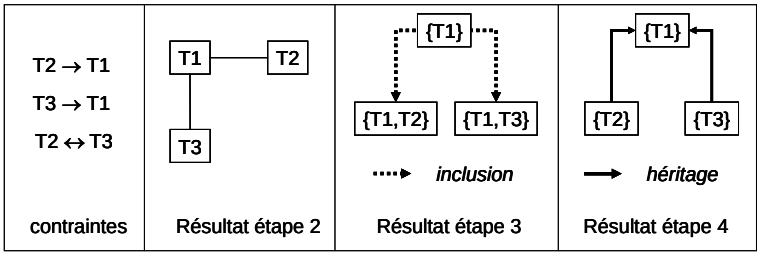


Figure 4.19 Exemple de déroulement de l'algorithme de normalisation

La figure 4.20 illustre un exemple d'enrichissement de la Topic Map à partir du thésaurus dans le domaine de la construction durable :

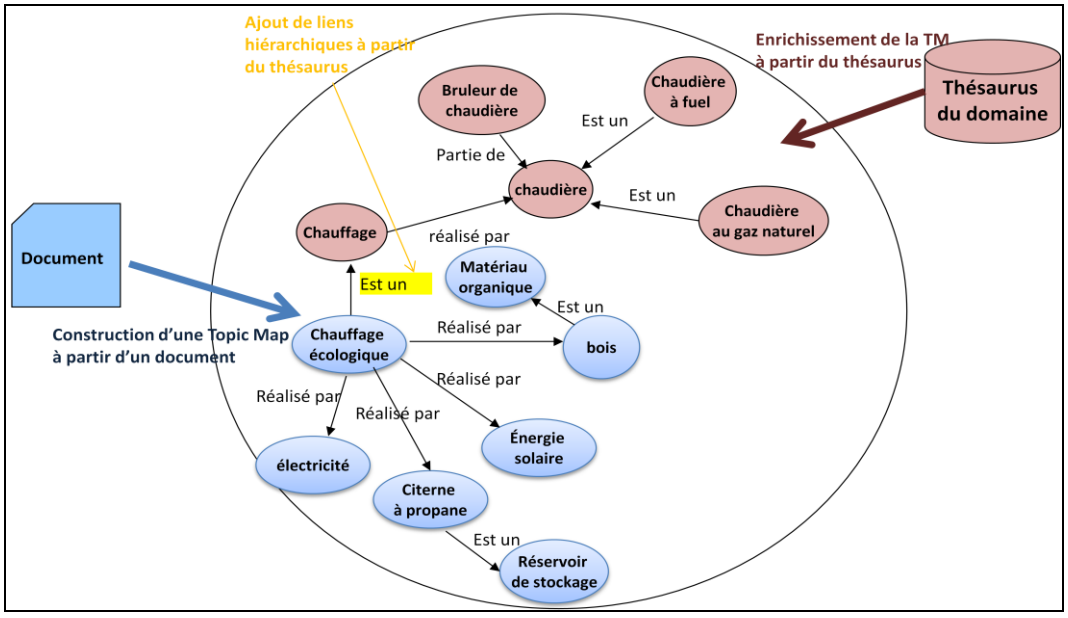


Figure 4.20 Enrichissement de la Topic Map à partir du thésaurus

A la fin des deux précédentes étapes, nous obtenons une Topic Map noyau multilingue du domaine puisqu'elle est construite à partir d'un document du domaine et enrichie par des liens ontologiques et des synonymes de Topics extraits à partir d'un thésaurus du domaine. La

richesse d'une Topic Map provient du fait qu'elle doit tout représenter, pour cela nous proposons dans la prochaine étape d'enrichir cette Topic Map par d'autres synonymes de Topics et d'autres liens structurels entre ces derniers, nous utilisons deux ressources externes, WordNet et WOLF pour les deux langues traitées anglais et français à travers l'exploitation de leur structure et des relations sémantiques présentes dans ces deux ressources.

4.2.3 Enrichissement de la Topic Map par les synsets et les liens de WordNet et de WOLF

L'objectif de cette étape est d'enrichir la Topic Map du domaine construite à partir du document en se basant sur le thésaurus. Pour cela, nous proposons d'utiliser deux ressources externes, WordNet pour l'anglais et WOLF (WordNet libre du français) pour le français, et d'exploiter leur structure et leurs relations sémantiques pour trouver les synonymes, hypéronymes, hyponymes des Topics existants dans la Topic Map. L'enrichissement de la Topic Map inclut l'ajout **des synsets** comme **synonymes de Topics** et les **liens possibles** qu'on peut trouver entre les Topics déjà présents dans la Topic Map. Par exemple si on a dans la Topic Map, les Topics « chaudière » et « réservoir » alors on pourrait rajouter un lien « partie de » entre eux puisque le réservoir est une partie de la chaudière.

WordNet[37] est une ressource lexicale de langue anglaise, disponible sur internet, qui regroupe des termes (noms, verbes, adjectifs et adverbes) en ensembles de synonymes appelés *synsets*. Un synset regroupe tous les termes dénotant un concept donné. Le terme associé à un concept est représenté sous une forme lexicalisée, sans marque de féminin ni de pluriel. Les synsets sont reliés entre eux par des relations sémantiques : relation de généralisation/spécialisation (is a kind of), relation composant/composé (is a part of). Une interface d'interrogation permet à un utilisateur de rechercher un terme dans la base de WordNet et renvoie une définition en langue naturelle, ainsi que ses généralisants, ses spécialisations et les termes auxquels il est lié par une relation de composition, pour les différents sens de ce terme (les différents synsets auxquels il appartient). WordNet possède l'avantage d'être une ressource assez fournie, le grand nombre de synsets ainsi que sa facilité d'accès ont en fait une ressource très utilisée en RI et en traitement automatique des langues.

L'initiative d'EuroWordNet a développé des dictionnaires semblables pour d'autres langues, mais celui pour l'anglais reste toujours le plus complet. WordNet utilise deux moyens pour définir le sens d'un mot : les synsets et les relations lexicales. Le sens d'un mot

[37] http://wordnet.princeton.edu/

est représenté par (i) l'ensemble des mots utilisés pour exprimer ce sens c'est à dire, par un ensemble de synonymes (les synsets) et (ii) une définition, par exemple :

heat
Synsets: heat, hotness, warmth
Definition: the sensation caused by heat energy

Le sens d'un mot est aussi déterminé par ses relations sémantiques avec d'autres sens. WordNet est structuré par des relations entre synsets et entre mots; les relations suivantes sont utilisées : synonymie, hyperonymie, hyponymie, méronymie et holonymie (Figure 4.21).

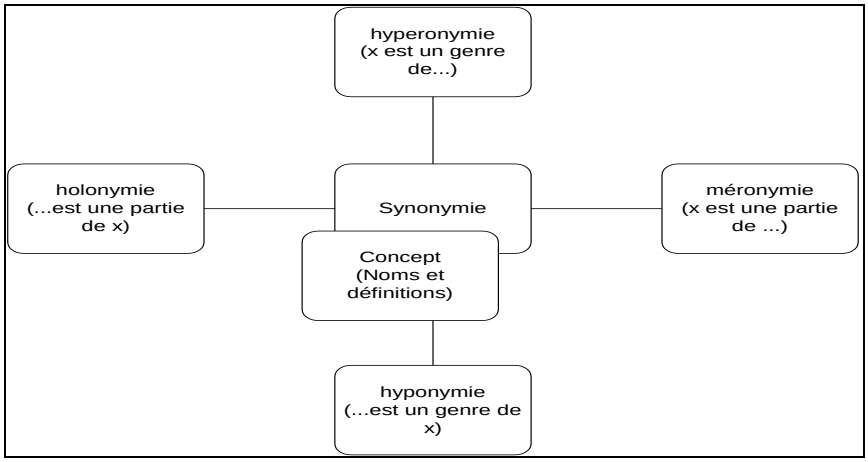


Figure 4.21 Principales relations dans WordNet

La relation **Synonymie** : le synset (synonym set), représente un ensemble de mots qui sont interchangeables dans un contexte donné.

La relation **Hyperonymie** : c'est le terme générique utilisé pour désigner une classe englobant des instances de classes plus spécifiques (c'est la relation type de)

La relation **Hyponymie** : c'est le terme spécifique utilisé pour désigner un membre d'une classe (relation inverse de Hyperonymie).

La relation **Méronymie** : Le nom d'une partie constituante (part of), substance de (substance of) ou membre (member of) d'une autre classe (relation inverse de l'holonymie). **X** est un méronyme de **Y** si **X** est une partie de **Y**. Par exemple : « voiture » a pour méronymes « porte » et « moteur ».

La relation **Holonymie** : le nom de la classe globale dont les noms méronymes font partie. Y est un holonyme de X si X est une partie de (is a part of) Y.

Des hiérarchies sont construites à l'aide de ces relations. Dans l'exemple ci-dessous, nous présentons une sous hiérarchie de WordNet correspondant au concept « *heat* ».

WOLF est une ressource lexicale sémantique libre pour le français. Elle a été construite à partir du Princeton WordNet (PWN) et de diverses ressources multilingues [Sagot et Fišer 2008]. Comme tout WordNet, WOLF est une base de données lexicales dans laquelle les mots (lexèmes) sont répartis en catégories et organisés en une hiérarchie de noeuds. Chaque nœud a un identifiant unique, et représente un concept, ou synset (ensemble de synonymes). Il regroupe un certain nombre de lexèmes synonymes dénotant ce concept.

Dans le cadre de notre travail, selon la langue, nous interrogeons WordNet et WOLF pour rajouter aux Topics existants, les termes qui y sont reliés par des relations sémantiques, pour notre cas, nous nous sommes contentés des relations de synonymie, de généralisation et de spécialisation. L'idée est donc d'enrichir les Topics existants avec la proximité sémantique en particulier les liens de généralisation, les liens de spécialisation et les liens de synonymie et ce jusqu'à un certain niveau paramétrable par le concepteur de la Topic Map (par exemple **le niveau 1** ce qui est le cas de notre travail).

Nous prenons par exemple le résultat de la recherche sur le Topic « *furnace* » (chaudière en français) qui donne :

Noun furnace has 1 sense

furnace - an enclosed chamber in which heat is produced to heat buildings, destroy refuse, smelt or refine ores, etc.
is a kind of chamber
has parts: grate, grating; register
has particulars: athanor; blast furnace; crematory, crematorium, cremation chamber; cupola; electric furnace; firebox; forge; gas furnace; incinerator; oil burner, oil furnace; tank furnace

A partir de WordNet, nous rajoutons par exemple, à la Topic Map, les Topics « *electric furnace* », « *gas furnace* » et « *oil furnace* » et nous les relions au Topic « *furnace* » par le lien ontologique « est un » (figure 4.22).

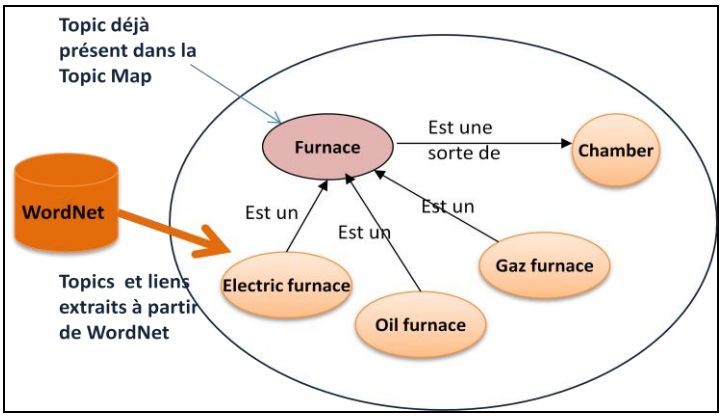


Figure 4.22 Exemple d'enrichissement de la Topic Map à partir de WordNet

Dans le cas où le résultat de la recherche sur un Topic donne deux ensembles de généralisant correspondent à deux sens différents. Seul est conservé le sens qui contient le concept racine de l'application étudiée (pour notre cas, la construction durable) par exemple le Topic « *heat* » a sept sens parmi lesquels :

heat, heat energy - a form of energy that is transferred by a difference in temperature
is a kind of energy
has particulars: geothermal energy; heat of dissociation; heat of formation; heat of solution; latent heat, heat of transformation; specific heat

heat, warmth, passion - intense passion or emotion
is a kind of emotionality, emotionalism

estrus, oestrus, **heat**, rut - applies to nonhuman mammals: a state or period of heightened sexual arousal and activity
is a kind of physiological state, physiological condition

Dans ce cas, nous considérons le sens 1 de l'exemple car il contient le terme « *energy* » qui représente le concept racine de notre application.

A ce stade, la Topic Map est enrichie et elle couvre les trois types de terminologies :

- La terminologie normalisée du domaine représentée par le thésaurus ;
- La terminologie du langage commun représentée par les synsets, les termes génériques et les termes spécifiques des Topics existants extraits de WordNet et de WOLF ;
- Et la terminologie des producteurs des documents du corpus qui peut couvrir partiellement les deux précédentes terminologies et utiliser une terminologie moins normalisée. Ceci permet de concilier les trois types de terminologies.

Au final, la Topic Map sera représentée comme **un réseau de Topics** dans lequel ces Topics sont organisés par **des liens ontologiques** qui regroupent les liens de type « est un », « partie de », les liens associatifs par exemple les liens « traite de », « réalisé par », les liens de synonymie et **les liens d'occurrences** (figure 4.23). De cette manière, l'utilisateur pourra naviguer dans ce réseau en utilisant ces liens sémantiques et localiser son centre d'intérêt ou bien découvrir des informations tout au long de sa recherche à travers la Topic Map.

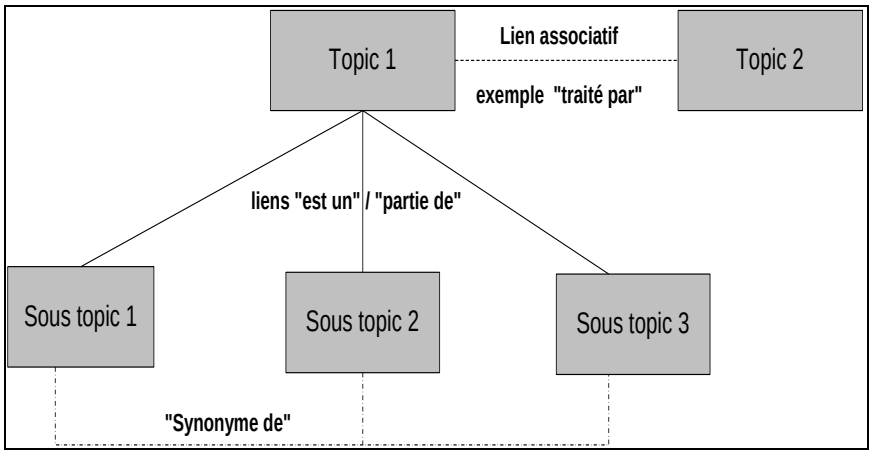


Figure 4.23 Les différents liens ontologiques dans la Topic Map

Notre idée consiste à rajouter la notion d'usage dans notre Topic Map, pour cela nous proposons d'enrichir cette Topic Map par un ensemble de **questions types et leurs réponses** et **les liens** entre eux et ce pour représenter les interrogations possibles des utilisateurs sur le contenu des documents.

4.2.4 Enrichissement de la Topic Map avec les liens d'usage

L'intérêt principal de la Topic Map globale est de guider l'utilisateur dans ses interrogations. Elle doit lui permettre de naviguer au sein de sa structure et d'accéder aux documents associés aux Topics explorés [Ellouze et al. 2009a], [Ellouze et al. 2009c].

Notre objectif à travers cette étape est d'introduire dans la Topic Map des connaissances sur les questions les plus fréquemment posées sur les documents par le biais de liens d'usage. Cela suppose bien sûr que les questions les plus fréquemment posées et leurs réponses ont été préalablement recensées. Le recensement de ces questions potentielles est fait suite à l'analyse de toutes les sources d'interrogations disponibles (les foires aux questions, les traces des discussions téléphoniques et des consultations directes avec les travailleurs du domaine, etc.). Toute question recensée est donc représentée en un Topic que l'on relie via le lien « répond à » à ses réponses c'est-à-dire les Topics référençant les documents qui permettent de répondre à cette question. Cette même question est aussi reliée à chacun des mots clés la constituant via un hyper lien de type « est composé de ». Le stockage des liens « est-composé-de » d'une question (i.e. phrase en langage naturel) vers les termes la composant permet d'une part une recherche par navigation et d'autre part une recherche automatique de « question proche » en reconstituant le vecteur de Salton [Salton, 1989] correspondant à la question, grâce à l'ensemble des termes qui la composent.

L'insertion d'un Topic obtenu par enrichissement peut se traduire soit par une action vide dans le cas où le Topic existe tel quel dans la Topic Map, soit par le rajout d'un nom à un Topic existant dans le cas où le Topic existe sous un autre nom soit par l'insertion effective de ce Topic dans le cas où il n'existe pas sous aucune des formes citées. La figure 4.24 illustre un exemple d'enrichissement de la Topic Map à partir des scénarios d'usage :

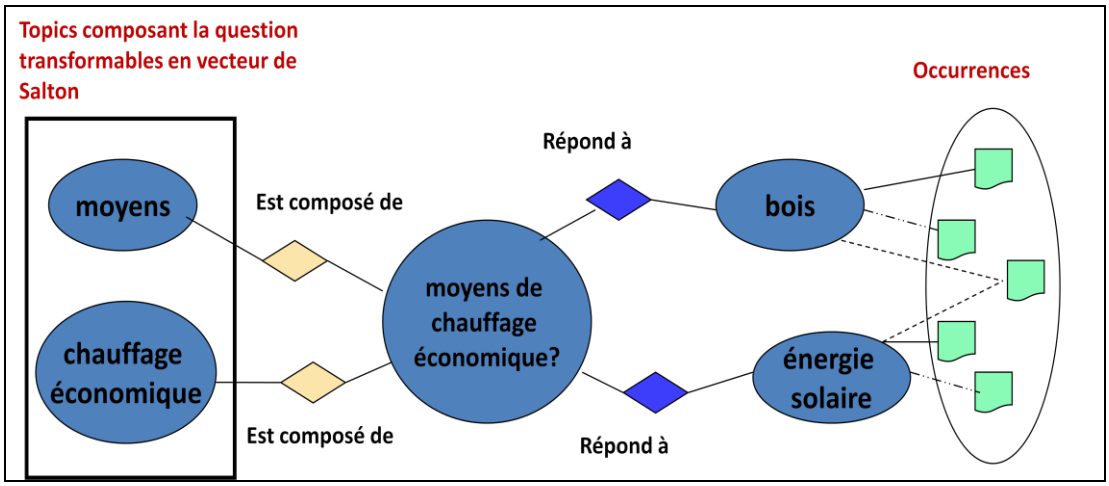


Figure 4.24 Enrichissement de la Topic Map à partir des scénarios d'usage

Puisque chaque question est représentée sous forme d'un vecteur de Salton contenant les termes qui la composent, lorsqu'un utilisateur pose sa question, cette dernière est elle aussi représentée sous forme d'un vecteur de mots, il s'agit dans ce cas par une méthode automatique (calcul de distance) retrouver le vecteur de la FAQ qui se rapproche le plus du vecteur des mots de la requête et par la suite renvoyer à l'utilisateur les réponses adéquates puisque celles-ci sont déjà prévues et recensées à partir de l'analyse des FAQ (figure 4.25).
Pour trouver le vecteur de la FAQ le plus proche du vecteur de la requête le l'utilisateur, on calcule la distance entre ces deux vecteurs en utilisant la mesure de cosinus. Cette mesure, donnée par l'équation 4.2 et initialement proposée par Salton [Salton, 1989], mesure l'angle que forme le vecteur de la requête extraite à partir de FAQ (*RF*) et le vecteur requête de l''utilisateur (*R*). Le cosinus vaut 1 si les vecteurs sont parallèles et 0 s'ils sont orthogonaux.

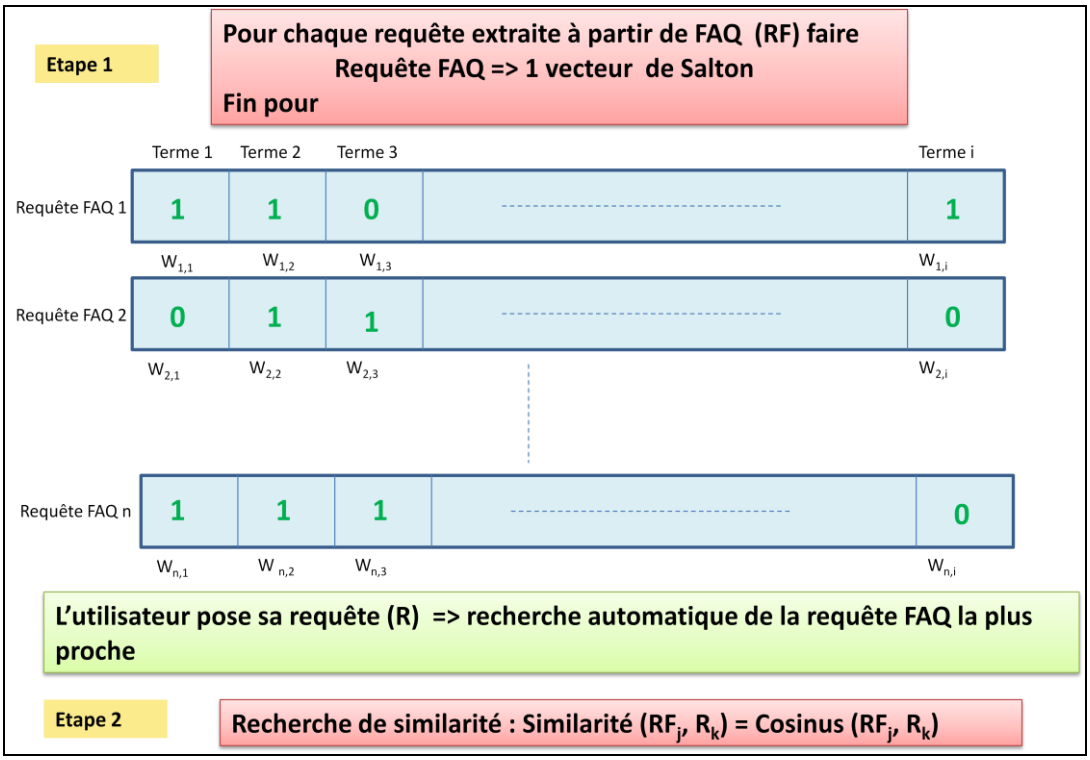


Figure 4.25 Processus de recherche de la requête FAQ la plus proche

La similarité entre deux documents RF et R serait alors mesurée par la formule suivante :

$$\cos inus(RF, R) = \frac{\sum_t W_{t,RF} W_{t,R}}{\sqrt{\sum_t W_{t,RF}{}^2 \sum_t W_{t,R}{}^2}}$$

Equation 4.10

Avec t appartenant à l'ensemble des termes dans la requête et $w_{t,RF}$ est le poids $tf \times idf$ assigné au terme t dans la requête de la FAQ RF, $w_{t,R}$ est le poids $tf \times idf$ assigné au terme t dans la requête de l'utilisateur R. $tf \times idf$ correspond au nombre de termes communs et au nombre de fois qu'ils apparaissent dans la requête.

Le tableau 4.1 présente les classes de questions que nous avons identifiées ainsi que quelques exemples de questions associés à chaque classe. Ces exemples ont été recensés à partir de ces sites Web :

http://www.ademe.fr : Site Web de l'Agence de l'Environnement et de la Maîtrise de l'Energie (France).

http://www.cstb.fr : Site Web du Centre Scientifique et Technique du Bâtiment (France).

http://www.rncan.gc.ca : Site Web sur les ressources naturelles au Canada (Canada).

http://www.ec.gc.ca : Site Web sur l'environnement au Canada (Canada).
http://www.avenir-energie.com Site Web sur les solutions de chauffage écologique et économique.

Classes de questions	Exemples de questions
Mode de fonctionnement	- Comment fonctionne une pompe à chaleur ? - Principes de fonctionnement des panneaux solaires photovoltaïques ? - Mode de fonctionnement du chauffage solaire ? - Quelle est la différence entre le mode de fonctionnement du solaire photovoltaïque et le solaire thermique ?
Guide d'installation, fiche technique	- Je recherche la documentation et la fiche technique pour chaudière murale Idéal Standard 2.24 FF - Guide d'installation d'une pompe à chaleur - A quelle hauteur doit être installé un radiateur électrique ?
Chauffage économique, écologique	- Quelle est l'économie de combustible en pourcentage avec une solution plancher chauffant ? - Economies réalisées grâce à une installation solaire Electrique / Photovoltaïque - Types de chauffage économique ? - Economies réalisées grâce à une installation solaire Thermique - Economies réalisées grâce à un récupérateur d'eau de pluie - Economies réalisées grâce à une pompe à chaleur géothermique sur nappe phréatique - D'où vient l'économie réalisée par les pompes à chaleur ? - Pourquoi la pompe à chaleur est elle considérée comme un moyen de chauffage écologique ?

Impacts sur l'environnement	- Impacts du chauffage au bois sur la santé - Impacts du chauffage au bois sur l'environnement
Normes, textes réglementaires	- Documentation sur la norme européenne EN 1264 - Texte de référence qui concerne l'ensemble des travaux à effectuer par l'installateur de chauffage central en ce qui concerne l'exécution des planchers chauffants - Le Cahier des Prescriptions Techniques du CSTB traitant des planchers chauffants et rafraîchissants - Norme d'application qui donne les épaisseurs d'isolant ou plus particulièrement les résistances thermiques minimales à respecter pour la pose du plancher chauffant. - Textes réglementaires relatif aux installations fixes destinées au chauffage et à l'alimentation en eau chaude sanitaire des bâtiments d'habitation
Liste des appareils certifiés	- Liste des récupérateurs d'eau de pluie certifiés - Liste des pompes à chaleur certifiées - Liste des chaudières à gaz certifiées
Conseils	- Conseil pour un fonctionnement optimal du chauffage au bois - Conseils sur le chauffage électrique
Devis	- Devis d'installation d'une pompe à chaleur - Devis d'installation d'un plancher chauffant électrique
Cout d'installation, Cout d'entretien	- A quel tarif puis-je revendre l'électricité produite par mon système photovoltaïque à EDF ? - Le cout d'un radiateur électrique

Liste des fournisseurs certifiés	- Comparatif des fournisseurs certifiés des appareils de chauffage solaire - Liste des fournisseurs certifiés des appareils de chauffage géothermique

Tableau 4.1 Exemples de scénarios de questions préparés à partir de FAQ

La figure 4.26 illustre l'étape d'enrichissement de la Topic Map à partir des scénarios d'usage :

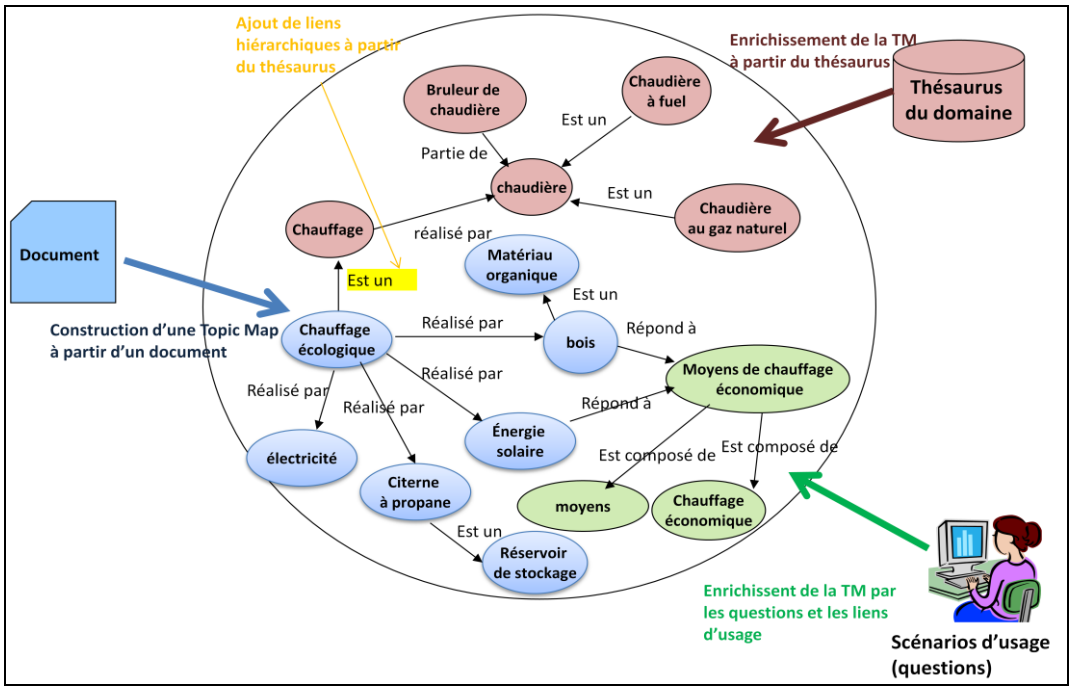


Figure 4.26 Enrichissement de la Topic Map par les questions et les liens d'usage

4.2.5 Enrichissement de la Topic Map globale par la Topic Map associée au document d_i

Cette étape a pour objectif d'intégrer et fusionner la Topic Map associée au document d_i, obtenue suite aux phases décrites précédemment, avec la Topic Map globale. Notre approche ACTOM, comme le montre la figure 4.10, est basée sur une construction incrémentale, le principe est de construire de façon incrémentale une Topic Map TM_i correspondante à un ensemble de documents D= {d_1, d_2, ..., d_i} en fusionnant la Topic Map TM_{i-1} correspondante à l'ensemble de documents D-{d_i} avec la Topic Map TMd_i associée au document d_i.

Nous avons établi un bref état de l'art, dans le deuxième chapitre, sur les techniques d'intégration de schémas conceptuels et d'ontologies. Les travaux sur l'intégration de schémas sont nombreux et variés [Do et Rahm, 2001], [Calvanese et al. 2001], [Stumme et Madeche, 2001], [Rasgado et Guzman, 2006], [Kong et al. 2006], etc. Ils se distinguent les uns des autres par les types de schémas manipulés (schéma E/A, schémas orientés objets, ontologies, documents XML, etc.), par les techniques d'appariement utilisées ou encore par les techniques de fusion adoptées.

Dans le cadre de notre travail, nous nous proposons de réutiliser nos travaux de recherche proposés par [Lammari et Métais, 2004] et [Lammari et al. 2008] pour l'intégration de deux hiérarchies dans le cadre de la maintenance d'ontologies, nous nous sommes également inspirés des travaux de [Lammari et Besbes-Essanaa, 2009] qui concernent l'intégration et la fusion de deux vues conceptuelles dans le cadre de leur projet RetroWeb.

D'après l'état de l'art sur les approches existantes d'intégration de schémas, nous remarquons que le processus général d'intégration est presque le même pour toutes les approches, ces dernières proposent un processus en quatre étapes :

1) Une étape de pré-intégration afin d'uniformiser les schémas sources et les traduire en un modèle commun ;
2) Une étape de comparaison permettant de détecter les similarités entre les entités des deux schémas à fusionner ;
3) Une étape de fusion qui rassemble les schémas sources en un seul schéma intégré en se basant sur les résultats trouvés dans la 2ème étape ;
4) Une étape de restructuration du schéma intégré pour éventuellement l'améliorer.

Dans notre approche d'intégration de Topic Maps, nous nous inspirons de cette démarche générale d'intégration de schémas et nous l'adoptons à notre cas. Notons que dans notre cas, nous n'avons pas à considérer la première étape d'uniformisation des Topic Maps à fusionner puisque nous construisons nous même ces Topic Maps selon un même processus, une même syntaxe et un même langage.

La stratégie d'intégration choisie est binaire. C'est-à-dire nous comparons et intégrons deux Topic Maps parmi celles à fusionner. Le résultat est une Topic Map intermédiaire qui, à son tour, est comparée et intégrée avec une autre Topic Map. Ce processus est répété jusqu'à intégration de toutes les Topic Maps, c'est-à-dire jusqu'à ce que nous terminions tous les documents disponibles dans notre référentiel multilingue. Ce processus est illustré par la figure 4.10.

L'algorithme d'intégration de deux Topic Maps prendra en entrée la Topic Map d_i associée au document d_i et la Topic Map globale (ou la Topic Map précédente TM_{i-1} correspondante à l'ensemble de documents D-{d_i}), la sortie sera une nouvelle Topic Map intégrée TM_i.

L'intégration de deux Topic Maps se fait en deux temps (figure 4.27). Dans un premier temps, nous procédons à un appariement des deux Topic Maps (matching). Cet appariement permet de générer deux Topic Maps en prenant en compte les résultats obtenus. Dans un second temps, nous fusionnons les deux Topic Maps résultantes de la première étape afin d'obtenir une Topic Map intégrée. A la fin, les experts du domaine seront sollicités pour valider la Topic Map résultante.

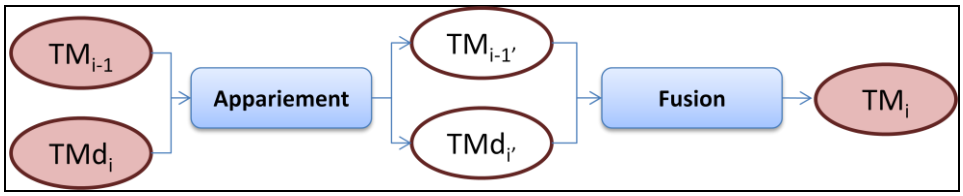


Figure 4.27 Processus d'intégration de deux Topic Maps

L'algorithme suivant résume les étapes d'intégration de deux Topic Maps dans notre approche ACTOM :

Action. *Enrichir la Topic Map globale TM_{i-1} avec la Topic Map d_i*

Action 1. *Aligner la Topic Map globale avec la Topic Map d_i*

Action 2. *Fusionner les Topics Maps résultantes de l'action 1*

Action 3. *Valider la Topic Map intégrée*

Nous présentons dans chacun des paragraphes suivants les deux étapes d'appariement et de fusion de Topic Maps.

4.2.5.1 Appariement de deux Topic Maps

L'appariement de deux Topic Maps dans ACTOM consiste en l'appariement des Topics suivi de celui des associations. **Nous limitons notre recherche de similarité à une recherche d'équivalence** prenant en compte les synonymes et le multilinguisme.

Notre approche d'appariement prend en considération les noms des Topics et des associations. Pour ce faire, nous proposons d'utiliser les deux ontologies générales (WordNet pour l'anglais et WOLF pour le français) et le thésaurus afin de minimiser les interactions avec l'utilisateur et les experts du domaine.

Appariement des Topics

Nous vérifions pour chaque Topic de la Topic Map d_i s'il existe déjà dans la Topic Map précédente (TM_{i-1} correspondante à l'ensemble de documents D-{d_i}), quatre cas sont possibles :

- Dans le cas où le Topic de la TMd_i est déjà présent dans la Topic Map précédente TM_{i-1} tel qu'il est, c'est-à-dire avec le même nom de base (*base name*) et ce pour les deux langues anglais et français, les deux Topics sont déclarés équivalents. Notons que le choix de la langue du nom de base d'un Topic est paramétrable par l'utilisateur ;
- Dans le cas où le Topic de la TMd_i existe dans la Topic Map précédente TM_{i-1} mais dans une langue différente de celle de la Topic Map d_i alors le traduire (ou rechercher son équivalent à partir du thésaurus ou de WordNet et WOLF) dans l'autre langue, mettre comme nom de base, le nom en français et l'autre nom (*variant name*), celui en anglais. Les deux Topics deviennent équivalents. Notons que la traduction d'un Topic d'une langue à une autre n'est pas toujours possible. En effet, un Topic peut ne pas avoir d'équivalent d'une langue à une autre puisque une langue est le résultat d'une culture et qu'il y a des Topics qui sont liés à la culture d'un pays ou d'un continent par exemple ceux liés aux traditions culinaires. Nous détaillerons les problèmes liés au multilinguisme dans la section suivante ;
- Dans le cas où le Topic de la TMd_i existe dans la Topic Map précédente mais avec un nom synonyme (dans une même langue) de celui qui existe dans la Topic Map associée au document d_i, alors le Topic aura **comme nom de base celui de la Topic Map précédente** et comme autre nom celui de la nouvelle Topic Map TMd_i. Les deux Topics sont considérés équivalents. Par exemple, si dans la TMd_i, nous avons identifié le Topic « habitation » et que ce Topic est déjà présent dans la Topic Map précédente mais sous un autre nom « maison », en interrogeant WOLF, nous remarquons que ces deux noms sont synonymes alors, dans ce cas, le Topic aura comme nom de base « maison » et parmi ses variantes de noms « habitation » ;
- Dans le cas où le Topic de la TMd_i n'existe pas dans la Topic Map précédente alors il n'a pas d'équivalent et aucun traitement ne lui est appliqué.

L'algorithme permettant l'appariement de deux Topics est le suivant :

Pour** chaque Topic T_i présent dans la Topic Map TMd_i **faire

***Si** le Topic T_i existe dans la Topic Map précédente TM_{i-1} (dans les deux langues français et anglais)*

Alors

Les deux Topics T_i et le Topic déjà présent dans TM_{i-1} sont déclarés équivalents.

Sinon

***Si** le Topic T_i existe dans la Topic Map précédente TM_{i-1} dans une des langues*

Alors

le traduire (ou rechercher son équivalent à partir du thésaurus ou de WordNet et WOLF) dans l'autre langue, mettre comme nom de base le nom en français et l'autre nom (variant name), celui en anglais. Les deux Topics T_i et le Topic déjà présent dans TM_{i-1} deviennent équivalents.

Sinon

***Si** le Topic T_i existe dans TM_{i-1} avec un nom synonyme (dans une même langue) de celui qui existe dans la TMd_i*

Alors

le Topic T_i aura comme nom de base celui de la Topic Map précédente et comme autre nom celui de la nouvelle Topic Map TMd_i. Les deux Topics sont considérés équivalents.

Sinon

***Si** le Topic T_i n'existe pas dans TM_{i-1}*

Alors

T_i n'a pas d'équivalent et aucun traitement ne lui est appliqué.

Fin pour

Appariement des associations

L'appariement des associations repose à la fois sur les Topics qu'elles relient et sur leurs noms identifiés durant une des différentes phases du processus de construction de la Topic Map ou attribués par l'expert du domaine lors de la phase de validation. L'appariement basé sur les noms des associations est effectué de la même façon que celui effectué pour les Topics. Cependant, pour les associations, nous ajoutons la condition suivante : deux

associations A_i et A_j appartenant respectivement à TMd_i et TM_{i-1} sont équivalentes si et seulement si :

a) elles sont équivalentes (en suivant le même algorithme que celui des Topics) ;

b) et l'ensemble des Topics participants dans A_i est égal à l'ensemble des Topics participants dans A_j.

4.2.5.2 Fusion de deux Topic Maps

L'objectif de cette étape est de fournir une Topic Map intégrée représentant le contenu sémantique des deux Topic Maps à fusionner. En s'inspirant des travaux de [Lammari et Besbes-Essanaa, 2009], la fusion des deux Topic Maps TMd_i et TM_{i-1} revient à fusionner des ensemble de hiérarchies de généralisation-spécialisation en tenant compte des associations existantes. Dans notre approche, en plus des algorithmes de fusion proposés par [Lammari et Besbes-Essanaa, 2009], [Lammari et Métais, 2004], nous nous appuyons sur le thésaurus du domaine et les deux ontologies générales pour avoir les chainons manquants lors de la fusion. Dans notre cas, chaque ensemble de hiérarchies à fusionner contient quatre types de hiérarchies : des hiérarchies issues du document d_i, des hiérarchies extraites à partir des documents $D-\{d_i\}$, des hiérarchies de WordNet et de WOLF et des hiérarchies issues du thésaurus du domaine.

Deux hiérarchies H_i et H_j appartenant respectivement à TMd_i et TM_{i-1} seront fusionnables si et seulement si H_i et H_j partagent des Topics types.

Pour les hiérarchies de WordNet (ou de WOLF) et du thésaurus, nous considérons le sous graphe minimum qui contient le plus possible de Topics présents dans les deux hiérarchies appartenant respectivement à de TMd_i et TM_{i-1}. Ces deux sous-graphes fournissent les chainons manquants.

Notre but est de fusionner deux hiérarchies H_i et H_j appartenant respectivement à TMd_i et TM_{i-1}, cependant, le résultat de la fusion pourrait contenir des chainons manquants. Par exemple, si on fusionne les deux hiérarchies H_1 et H_2 de l'exemple illustré dans la figure 4.28, on ne va jamais arriver à les fusionner car il manque les liens disant qu'un « chauffage au bois » est un « chauffage écologique », et de même pour le Topic « chauffage solaire ». Et bien justement, ces liens sont fournis par le thésaurus du domaine ou par une ontologie générale telle que WordNet. En effet, dans l'exemple de la figure 4.28, la hiérarchie H_3 issue de WordNet fournit le fait qu'un « chauffage écologique » est un « chauffage », la hiérarchie H_4 extraite du thésaurus du domaine décrit le fait que le « chauffage au bois » est un « chauffage écologique » et que le « chauffage solaire » est aussi un « chauffage écologique ».

Cela explique donc notre idée de rajouter, dans notre processus de fusion, le thésaurus du domaine qui fournit des informations techniques liées au domaine et les ontologies générales (WordNet et WOLF) contenant des informations générales du langage commun.

Dans notre approche de fusion, nous mettons plus l'accent sur une bonne reconstruction de la hiérarchie, en se basant non seulement sur **l'algorithme de [Lammari et Besbes-Essanaa, 2009]** mais aussi sur **l'ajout des chainons manquants à partir du thésaurus ou de WordNet** (ou WOLF) come l'illustre la figure 4.28.

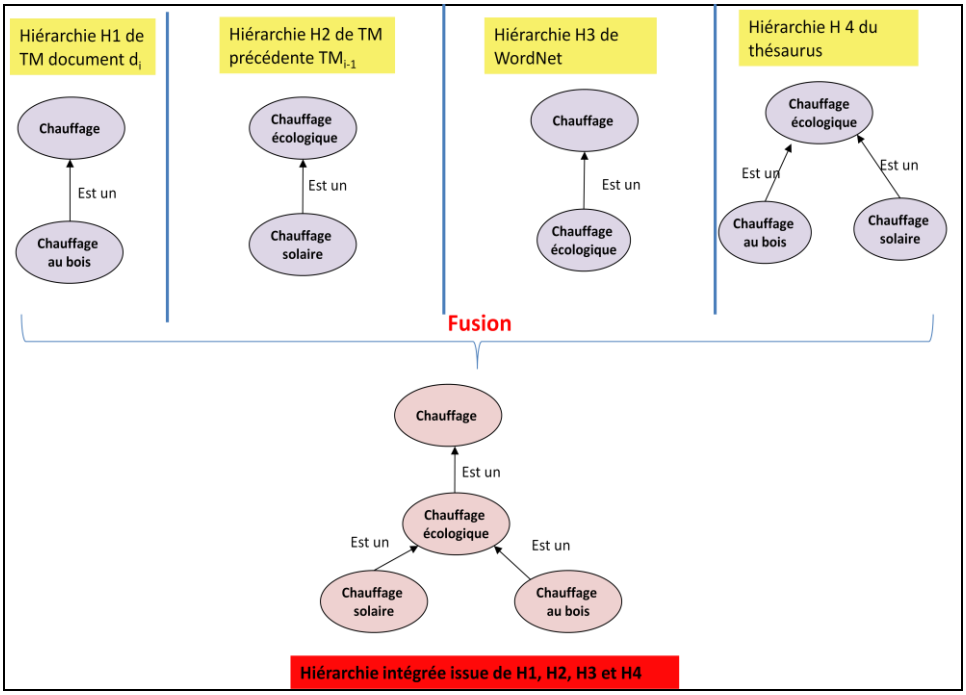


Figure 4.28 Exemple de hiérarchies intégrées

Nous proposons le processus de fusion de Topic Maps décrit dans la Figure 4.29 :

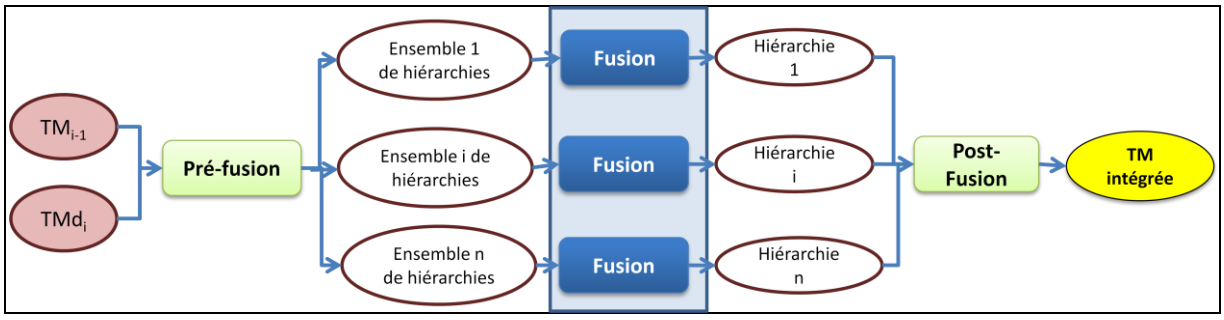


Figure 4.29 Processus de fusion de Topic Maps

Notre processus de fusion se décompose en trois grandes étapes. **La première étape est une étape de pré-fusion** permettant de sélectionner dans les deux Topic Maps TMd_i et TM_{i-1}, les ensembles de hiérarchies à fusionner. Chaque ensemble contient une hiérarchie issue du

document d_i (et qui appartient bien évidement à TMd_i), une hiérarchie extraite de l'ensemble des documents D-{d_i} (et qui appartient à TM_{i-1}), une hiérarchie extraite de WordNet (ou de WOLF) et une hiérarchie issue du thésaurus du domaine.

De plus, pour permettre un dispatching convenable des associations lors de la reconstitution de la Topic Map intégrée (étape 3 du processus d'intégration), nous sommes amenés à marquer les pattes des associations par un rôle. Une association A, reliant deux Topics T_1 et T_2 et se trouvant à la fois dans TMd_i et TM_{i-1}, verra ses deux pattes marquées de la même façon. A titre d'exemple, soit les Topic Maps TM_1 et TM_2 de la figure 4.30, dans chacune des deux Topic Maps, nous avons procédé à un marquage des pattes. L'association « permet » se trouvant aussi bien dans TM_1 et TM_2 possède les mêmes marques a et b.

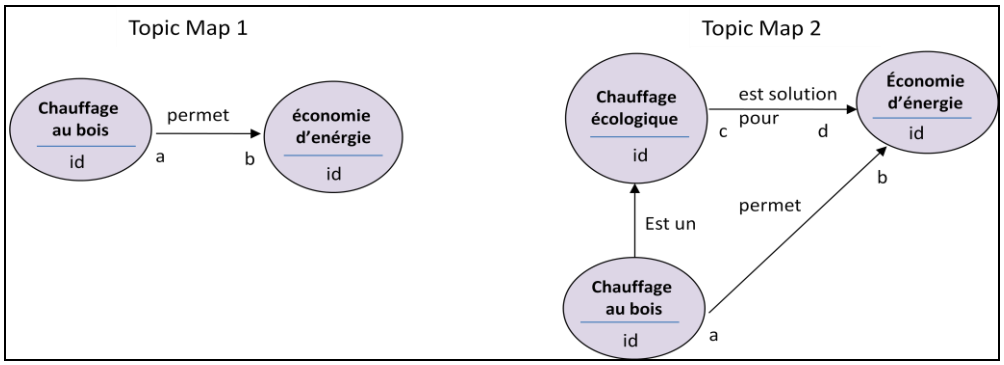


Figure 4.30 Exemple de deux Topic Maps à fusionner

La seconde étape est l'étape de fusion proprement dite. Elle s'applique à chaque groupe de hiérarchies déterminé dans la pré-fusion. Dans le cas où les deux hiérarchies sont équivalentes (même ensemble de Topics, même ensemble de liens), il n'y a pas lieu d'exécuter cette étape et le résultat sera la hiérarchie en question. Pour construire à partir de deux hiérarchies non équivalentes, une hiérarchie intégrée, le processus utilise deux techniques de transformation de modèles décrites dans [Lammari et Métais, 2004]. La première technique consiste en une traduction d'une hiérarchie en fonction booléenne (H_vers_FB). La seconde technique est l'inverse de la première. Elle permet de transformer une fonction booléenne en une hiérarchie (FB_vers_H). Ces deux techniques sont brièvement présentées dans les sections suivantes. Pour pouvoir appliquer ces deux techniques dans le contexte de notre travail, nous proposons d'ajouter à chaque Topic un attribut « id » qui représente l'identificateur du Topic.

Comme le montre la figure 4.31, pour fusionner les quatre hiérarchies (du document, de la Topic Map précédente, le sous graphe de WordNet, le sous graphe du thésaurus), nous commençons par transformer chacune d'elle en une fonction booléenne, puis les quatre

fonctions booléennes sont unies en une seule fonction qui, à son tour, est transformée en une hiérarchie.

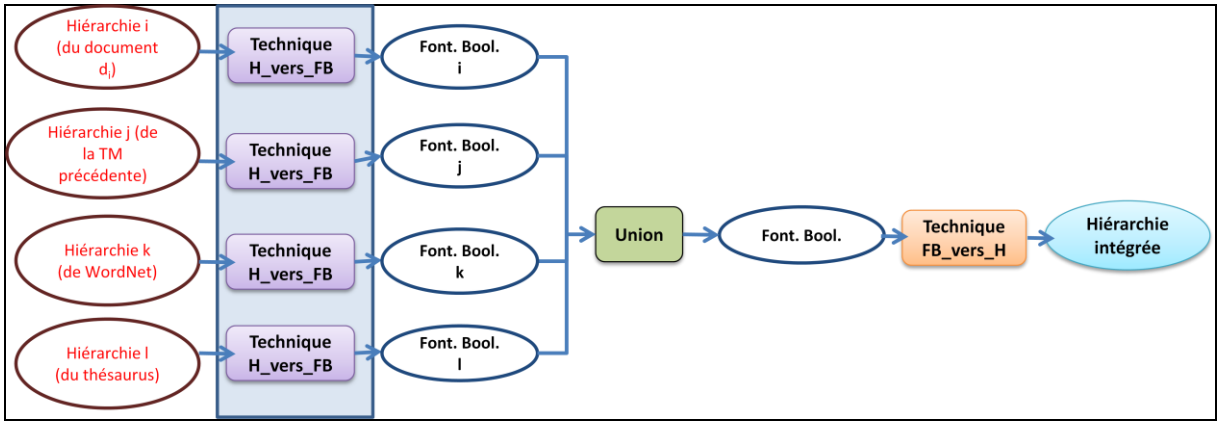


Figure 4.31 Processus d'intégration de hiérarchies

Pour unir deux fonctions booléennes φ1 et φ2 en une seule fonction booléenne φ3, on utilise l'algorithme suivant présenté dans [Lammari et Métais, 2004]

Soit B l'ensemble des variables de φ1 et φ2

Pour** chaque minterm T de φ1 et φ2 **faire

  Pour** chaque variable x dans B **faire

    Si** ni x ni $\bar{x}$ n'apparaissent dans T, **alors** $T = T.\bar{x}$ **finIf

  FinPour

FinPour

Si** T n'apparaît pas dans φ3 alors φ3 = φ3 + T **finIf

La troisième étape de la fusion est une étape de post-fusion. Elle sert à reconstituer la Topic Map intégrée à partir des hiérarchies intégrées. La reconstitution s'opère moyennant les rôles attribués aux associations dans l'étape de pré-fusion. Deux rôles d'une même association A, se trouvant chacun dans une hiérarchie intégrée, permettront de relier ces deux hiérarchies via l'association A.

Description de la technique H_vers_FB

Le but de cette technique est de représenter, sous forme de fonction booléenne, tous les types d'occurrences pouvant être représentée par une hiérarchie d'héritage. Pour ce faire, on associe à chaque attribut de la hiérarchie et à chaque patte d'association reliée à un Topic de la hiérarchie, une variable booléenne. La transformation d'une hiérarchie consiste en la construction d'une fonction booléenne $\varphi(x_1,...,x_n)$ où chaque maxterme T représente un type

d'occurrence représenté par chaque Topic de la hiérarchie. Ce maxterme est une conjonction de variables $x_i^{'}$ où chaque $x_i^{'}$ est égale :

- soit à x_i si l'attribut ou la patte d'association correspondante à x_i participe dans la description du type d'occurrence représenté par T ;
- soit à $\overline{x}_i$ dans le cas contraire.

A titre d'exemple, nous considérons les deux Topic Maps (TM_1 et TM_2) de la figure 4.30 où chaque Topic possède un attribut représenté par son identificateur « id ».

La fonction booléenne correspondante à la hiérarchie (issue de TM_2 de la figure 4.30) constituée des Topics « chauffage au bois » et « chauffage écologique » est :

$$x1.x2.x3.x4 + \overline{x1}.x2.\overline{x3}.x4$$

Les variables x1, x2, x3 et x4 correspondent respectivement aux attributs « id » du Topic « chauffage au bois » et celui du Topic « chauffage écologique » et aux rôles « a » et « c ». Le premier et le second maxterm décrivent respectivement les types d'occurrences du Topic « chauffage au bois » et du Topic « chauffage écologique ».

La fonction booléenne associée au Topic « chauffage au bois » de la Topic Map TM_1 (Figure 4.30) est : $x1.x2.x3.\overline{x4}$. L'union de ces deux fonctions booléennes donnera la fonction booléenne suivante :

$$\varphi_1 = x1.x2.x3.\overline{x4} + x1.x2.x3.x4 + \overline{x1}.x2.\overline{x3}.x4$$

Description de la technique FB_vers_H

Pour déduire, à partir d'une fonction booléenne φ($x_1,...,x_n$), la hiérarchie d'héritage qu'elle représente, on la rend, dans un premier temps, sous la forme conjonctive (en calculant $\overline{\overline{\varphi}}$) . Dans un second temps, en appliquant les règles d'extraction citées dans [Lammari et al., 2008] et [Lammari et Métais, 2004], on déduit l'ensemble des contraintes d'existence. Cet ensemble utilisé par l'algorithme de normalisation décrit dans [Lammari et Métais, 2004]. permettra la génération de la hiérarchie d'héritage correspondante à φ.

A titre d'exemple, à partir de φ_1 décrite ci-dessus, on obtient la fonction $\overline{\overline{\varphi}}_1$ de laquelle on extrait les contraintes d'existence suivantes :

id(chauffage au bois)↦id(chauffage écologique), a↦id(chauffage écologique), c↦ id(chauffage écologique), id(chauffage au bois)↔a.

De cet ensemble de contraintes, on peut déduire la hiérarchie de la figure 4.32b. La figure 4.32a représente le résultat de la fusion de la hiérarchie composée du Topic

« chauffage au bois » de la Topic Map TM_1 (Figure 4.30) avec celle de la Topic Map TM_2 composée elle aussi de l'entité « chauffage au bois ».

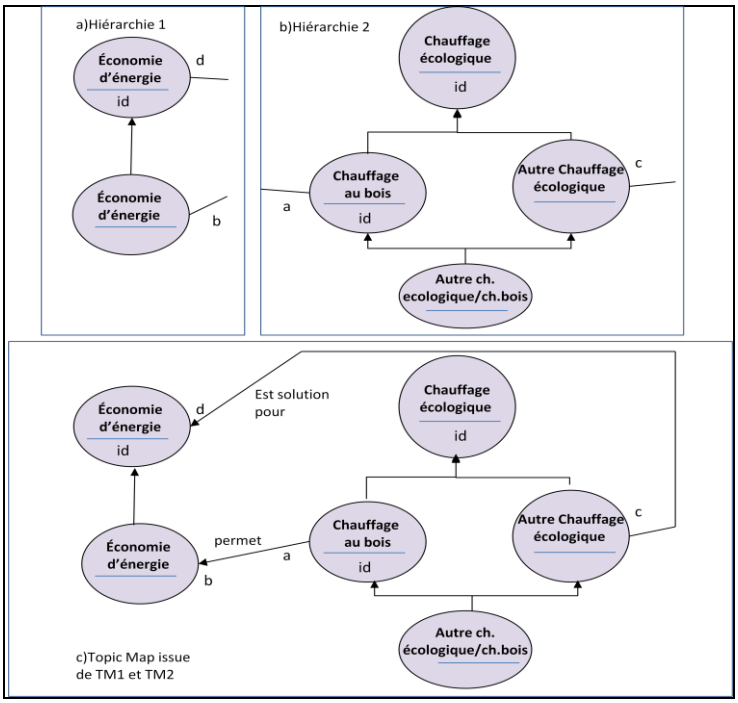


Figure 4.32 Hiérarchies et Topic Maps intégrées

La Topic Map intégrée de la figure 4.32c est produite par l'étape de post-fusion qui est chargée de reconstituer des Topic Maps intégrées à partir de hiérarchies intégrées. Une fois la Topic Map globale générée, on pourra, si nécessaire procéder à une restructuration de cette Topic Map.

4.2.5.3 Regroupement de Topics redondants par recherche de similarité

L'objectif de cette phase est de supprimer les Topics redondants dans la Topic Map globale, il s'agit d'une opération de nettoyage afin de diminuer le nombre de Topics inutilement présents dans la Topic Map.

Comme il serait trop complexe de comparer tous les Topics deux à deux vu leur nombre très élevé, notre idée consiste à définir un ensemble d'heuristiques pour le choix des Topics à comparer, par exemple, nous choisissons de comparer tous les Topics qui ont un certain nombre de documents en commun.

Une heuristique pour calculer ce nombre de documents est définie comme suit :

$$\frac{|D_{T1} \cap D_{T2}|}{|D_{T1} \cup D_{T2}|} > \frac{1}{4}$$

Avec T1 et T2 deux topics

D_{T1} l'ensemble des documents reliés à T1

D_{T2} l'ensemble des documents reliés à T2

Dans le cas où deux Topics pointent vers un nombre suffisant de documents en commun, nous vérifions en plus s'ils sont proches en se basant sur le calcul de distance et les mesures de similarité par exemple avec WordNet, et enfin nous envisageons de les fusionner (après avoir demandé la confirmation à l'utilisateur).

A titre d'exemple, nous considérons le Topic T1 « peintre » qui pointe sur les documents D_1, D_2, D_3, D_4 et un deuxième Topic T2 « entreprise de peinture » lié aux documents D_1, D_2, D_3, D_7, D_8. Nous comparons donc ces deux Topics, nous constatons que : $\frac{|D_{T1} \cap D_{T2}|}{|D_{T1} \cup D_{T2}|} = \frac{1}{2} > \frac{1}{4}$, la distance sémantique donnée par WordNet entre ces deux Topics est proche, donc nous proposons un regroupement, qui dans ce cas là est pertinent.

Un autre exemple, soit le Topic T3 « chauffage solaire » qui pointe sur les documents D_1, D_2, D_3, D_4, D_8 et le Topic T4 « réduction d'impôts » relié aux documents D_1, D_2, D_3, D_7, D_8. Nous comparons dans ce cas ces deux Topics, nous constatons que : $\frac{|D_{T3} \cap D_{T4}|}{|D_{T3} \cup D_{T4}|} = \frac{2}{3} > \frac{1}{4}$,

mais la distance sémantique donnée par WordNet est grande, donc nous ne proposons pas le regroupement.

Notons que le résultat de la fusion de deux Topics est un seul Topic qui sera relié aux documents associés aux deux Topics sources, par exemple, soit T1 et T2 (appartenant respectivement aux Topic Maps TMd_i et TM_{i-1}), T1 est relié au document D_1 et T2 est relié au document D_2, le résultat de la fusion de T1 et T2 est un Topic qui sera relié aux documents D_1 et D_2.

Nous comptons, comme perspectives de notre travail, développer plus d'heuristiques pour le choix des Topics à comparer et envisager d'utiliser d'autres mesures de distance. Un travail d'approfondissement sera également mené sur l'étude de la meilleure distance pour la comparaison des Topics choisis.

4.2.6 Annotation de la Topic Map globale par les documents et leurs segments thématiques

Suites aux étapes de création et d'enrichissement détaillés précédemment, nous obtenons une Topic Map multilingue, enrichie à partir des documents du référentiel, d'un thésaurus du domaine, de WordNet et de WOLF et des requêtes. Cette Topic Map est **multi-niveaux**, elle organise les Topics selon deux niveaux en plus du niveau ressources qui comprend les documents et leurs fragments. Le niveau le plus haut de la Topic Map contient les Topics représentant **les thèmes** traités dans les documents et reliés entre eux par **des liens ontologiques**, ainsi que des **Topics questions** et réponses reliés entre eux par **des liens d'usage** de type « composé de » entre Topics questions et les termes qui les composent et des liens de type « répond à » entre Topics questions et Topics réponses. Le deuxième niveau de la Topic Map englobe **les Topics concepts du domaine** reliés entre eux et aux Topics du premier niveau par des liens ontologiques de type « est un », « partie de », « synonyme de », « relié à » ou des liens associatifs relatifs au domaine.

Le troisième niveau contient **les ressources** multilingues (pour notre cas l'anglais et le français) c'est-à-dire les documents et leurs segments thématiques, chaque ressource est reliée au Topic qui en parle par **un lien occurrence**.

L'annotation de la Topic Map consiste à annoter chaque Topic par la liste des documents et segments qui traitent de ce Topic. Cette phase est réalisée lors des processus de création et d'enrichissement de la Topic Map à partir du référentiel puisque à partir de la matrice LSI qui nous donne l'importance de chaque terme et de chaque concept dans les documents et les segments, l'ajout d'un terme ou d'un concept comme un Topic à la Topic Map inclut l'ajout du lien occurrence entre ce Topic et la ressource (document et segment) à partir desquels il est extrait. Ce lien occurrence est étiqueté par **la mesure *tof×idf*** (si le Topic est un Topic thème ou Topic question) ou ***tf×id*** (si le Topic est un concept du domaine, un sous type de Topic ou une instance de Topic) du Topic par rapport au document ou au segment, ce qui nous permet de trier les documents et leurs segments par ordre de pertinence.

C'est l'une des extensions que nous avons proposé dans le modèle des Topic Maps pour mieux organiser et trier les ressources et faciliter la recherche à travers les Topic Maps.

Tri des ressources selon l'attribut « tof×idf » et leur filtrage selon l'attribut « Langue »

Nous proposons d'utiliser la notion de **facette** pour caractériser les ressources, en effet, comme on l'a déjà présenté, la facette est un ensemble d'attributs-valeurs liés au lien

occurrence pour décrire la ressource correspondante (document ou segment). Nous avons choisi, en plus des attributs type, taille, date, format, public cible, organisation…, deux attributs qui sont **la mesure *tof×idf*** indiquant l'importance du Topic dans le document (ou *tf×idf* dans le cas d'un terme) et un autre attribut qui est **la langue du** document (français ou anglais). La notion de facette telle qu'elle présentée dans le standard des Topic Maps, permet de filtrer les ressources selon ces attributs, justement c'est l'une des raisons qui nous ont poussés à choisir les Topic Maps puisque notre objectif est de proposer un SRI intelligent fondée sur une Topic Map multilingue, donc nous utilisons l'attribut langue pour filtrer les documents selon la langue dans le cas où l'utilisateur ne désire afficher que les documents en anglais ou en français. De plus, nous nous basons sur l'attribut *tof×idf* (ou *tf×idf)* pour trier les ressources par ordre décroissant selon leur degré de pertinence. La figure 4.33 illustre un exemple d'intégration des facettes pour la description d'un document :

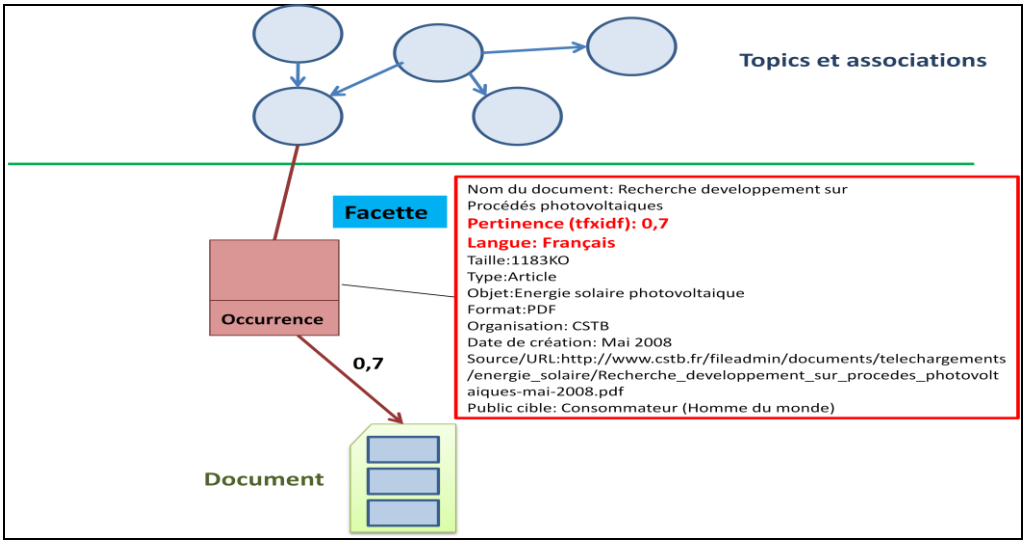


Figure 4.33 Intégration des facettes pour la description des documents

Au final, l'objectif est de permettre à l'utilisateur d'accéder directement aux ressources disponibles dans les différentes langues, à partir de la structure de Topic Map en naviguant à travers ses liens, ce ci constitue l'un des avantages du modèle des Topic Maps puisqu'il permet de représenter à la fois la partie **connaissances du domaine** et la partie **ressources** en les reliant par les liens occurrences.

Notons que toutes les étapes d'enrichissement et d'annotation de la Topic Map citées précédemment sont appliquées pour toutes les facettes linguistiques. Nous détaillons l'aspect multilingue dans la dernière section de ce chapitre.

La validation de la Topic Map résultante de chaque étape consiste à préciser la sémantique de certains liens ou encore à supprimer ou à ajouter des liens et/ou des Topics. Elle nécessite la collaboration d'un ou plusieurs experts du domaine.

En conclusion, grâce au réseau de liens sémantiques englobant les liens ontologiques et les liens d'usage entre les sujets qu'elles représentent, la Topic Map globale, enrichie et annotée vise principalement à permettre à des utilisateurs différents, ayant des cultures, des connaissances, des profils, des backgrounds différents et s'exprimant avec des vocabulaires différents de trouver facilement ce qu'ils cherchent.

La Topic Map enrichie et annotée représente une **méta-carte sémantique** médiant la terminologie normalisée du domaine (à partir du thésaurus), la sémantique des producteurs des documents (à partir des termes et des associations extraits des documents), les sources d'interrogations relatives au domaine et la sémantique du langage courant représentée par WordNet et WOLF.

4.3 Gestion du multilinguisme dans la construction de la Topic Map

Dans notre approche, nous visons à produire une Topic Map comme une carte sémantique de navigation permettant de représenter, structurer et organiser sémantiquement un ensemble de documents disponibles dans plusieurs langues. La construction d'une telle Topic Map offrira à l'utilisateur la possibilité de s'enrichir de connaissances se trouvant dans des documents écrits dans une langue autre que la sienne.

La prise en compte de multilinguisme permet à l'utilisateur, lors de sa navigation, d'avoir accès à des documents qui ne sont pas dans sa langue d'origine. Le grand intérêt de cette approche par rapport à de simples traductions de réponses est de proposer à l'utilisateur des documents correspondant à des Topics n'existant pas forcement dans sa langue ou dans sa culture. Ceci constitue à notre avis un enrichissement culturel.

Les problèmes liés à la recherche multilingue sont très nombreux, ils sont reliés principalement aux ambiguïtés sémantiques. Nous distinguons par exemple le problème de la polysémie où un même terme peut avoir différents sens et l'homographie dans le cas où deux mots différents s'écrivent de la même façon ou bien le problème du sens large : Un terme qui a un sens très large, exemple « air » peut prendre un sens particulier dans certains domaines par exemple « air bag ».

Par ailleurs, une des particularités du multilinguisme est l'absence éventuelle de termes sémantiquement équivalents d'une langue à une autre. Les solutions actuelles pour la résolution des problèmes du multilinguisme utilisent des systèmes de traduction automatique qui ne prennent pas en compte la sémantique des termes, les ambigüités sémantiques et l'absence de termes équivalents, elles proposent de traduire les documents dans la langue des requêtes ou inversement ou bien représenter les documents et les requêtes selon un langage intermédiaire ou pivot ce qui est très couteux en temps surtout pour des documents de grande taille.

Dans notre approche, nous proposons d'explorer le modèle des Topic Maps pour essayer de résoudre les problèmes liés au multilinguisme et surtout dans le cas ou un terme n'a pas d'équivalent d'une langue à une autre par exemple le terme « bsisa » en arabe désignant un aliment traditionnel n'a pas d'équivalent en français mais si un utilisateur qui parle le français navigue dans la Topic Map, il va découvrir, grâce au lien hiérarchique entre « bsisa » et « aliment », que « bsisa » est un aliment qui fait partie de la tradition alimentaire arabe.

4.5.1 Le modèle des Topic Maps pour la gestion du multilinguisme

Le standard des Topic Maps dispose du concept de **scope** (contexte) ou domaine de validité et du concept de **facette**. Le scope indique dans quel contexte tel Topic aura tel nom, telles occurrences et tels rôles. La facette permet de compléter les informations à propos d'une occurrence en ajoutant des informations de type attributs-valeurs dans le composant occurrence qui référence le document concerné. Nous avons exploité ces deux concepts pour la prise en charge du multilinguisme dans l'élaboration de la Topic Map. En effet, nous proposons de définir un scope pour chaque langue traitée dans la Topic Map avec la possibilité bien sûr d'attribuer à un Topic une liste de noms dans différentes langues.

Comme nous l'avons détaillé dans la section précédente, nous exploitons aussi le concept de facette dans un objectif de filtrage des documents selon leurs langues. Pour ce faire, nous avons défini un attribut « langue » dans la facette du composant occurrence reliant un Topic aux documents qui en parlent. La valeur prise par cet attribut dans une occurrence donnée correspondra à la langue du document référencé par cette occurrence.

4.5.2 Les liens de synonymie et les liens hiérarchiques pour la gestion du multilinguisme

Pour la résolution de problèmes d'équivalence entre les termes dans différentes langues, nous proposons d'utiliser les liens de synonymies et les liens hiérarchiques entre les Topics. En effet, le standard des Topic Maps dispose de la notion de noms multiples d'un Topic qui sont en fait les synonymes de ce Topic. Dans un contexte monolingue un Topic peut avoir plusieurs noms ou synonymes dans la même langue, comme nous l'avons mentionné précédemment, ces synonymes sont identifiés à partir du thésaurus, de WordNet (ou de WOLF) et des experts du domaine.

Dans un contexte multilingue, un Topic aura un « nom » dans chaque langue **s'il existe**, avec des possibilités de « valeurs nulles » si le Topic n'a pas d'équivalent dans une langue. Ces valeurs nulles seront prises en compte dans la navigation, par exemple en utilisant **les liens hiérarchiques** entre les Topics.

Par exemple si un utilisateur, lors de sa navigation, découvre l'existence d'un nouveau Topic qui n'est pas dans sa langue d'origine alors pour comprendre sa signification, il pourra passer à un niveau plus haut dans la Topic Map à travers le lien « est un », par exemple, et découvrir à quel type appartient ce Topic ou bien naviguer au même niveau et découvrir les synonymes de ce Topic dans les autres langues grâce aux liens de synonymie.

La figure 4.34 présente la structure d'un Topic dans un contexte multilingue, chaque Topic aura un nom dans chaque langue et des synonymes dans les différentes langues s'ils existent.

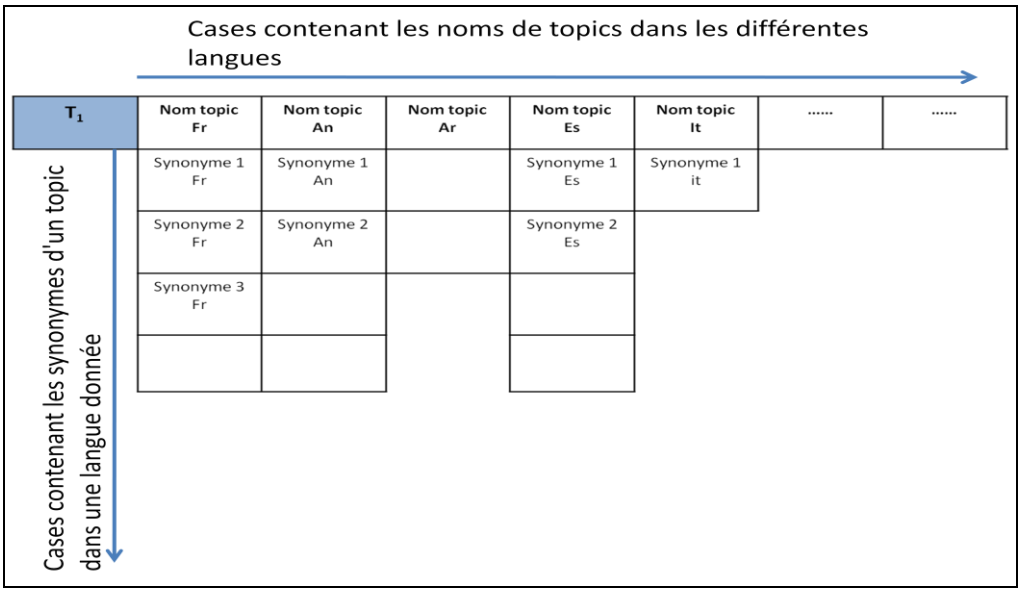


Figure 4.34 Format d'un Topic multilingue

4.4 Conclusion

Dans ce chapitre, nous avons proposé, ACTOM, une **A**pproche **incrémentale et évolutive** de **C**onstruction d'une **TO**pic Map **M**ultilingue. Cette dernière sert à organiser un contenu multilingue composé de documents textuels. Elle a pour avantage de faciliter la recherche d'information dans le contenu. ACTOM conjugue l'utilisation de quatre sources d'information : (a) un référentiel sémantique de documents disponibles dans différentes langues indexés thématiquement et sémantiquement, (b) un thésaurus du domaine, (c) deux ontologies générales WordNet pour l'anglais et WOLF pour le français ainsi que (d) l'ensemble de toutes les sources d'interrogations possibles telles que les questions relatives aux documents sources qu'un expert du domaine ou un utilisateur quelconque peut poser, les foires aux questions (FAQ), etc.

Notre approche a pour première originalité de **prendre en compte l'usage de la Topic Map** à travers la mise en œuvre de liens d'usage entre les questions potentielles extraites des sources d'interrogations disponibles et les réponses associées. Toute question potentielle (i.e. phrase en langage naturel) représentée sous forme de Topic est aussi reliée à chacun des mots clés la constituant via un hyper lien de type « est composé de ». Le stockage des liens « est-composé-de » d'une question vers les termes qui la composent permet d'une part une recherche par navigation et d'autre part une recherche automatique de « question proche ».

De part son processus **incrémental**, notre approche nous permet de réutiliser le même processus de construction pour la gestion de l'évolution de la Topic Map c'est à dire à chaque enrichissement du contenu qu'elle organise.

De plus, ACTOM présente l'avantage d'utiliser **un référentiel de documents segmentés thématiquement et indexés sémantiquement** pour la construction de la Topic Map, ce référentiel nous donne la possibilité d'indexer un Topic par un segment de document ce qui n'était pas possible avec le modèle des Topic Maps tel qu'il est défini par la norme ISO. Par ailleurs, le référentiel donne, pour chaque terme et chaque concept indexant un document et un segment, leurs $tf \times idf$ (pour les termes) *et* $tof \times idf$ *(*pour les concepts). Ces mesures sont utilisées pour pondérer le lien occurrence entre le Topic et la ressource et filtrer ainsi ces ressources selon leur degré de pertinence afin de faciliter la navigation et la recherche dans la Topic Map.

Dans notre approche, nous avons proposé **une méthode d'intégration de Topic Maps**, en effet, nous nous basons sur une construction incrémentale, qui consiste à créer une Topic Map à partir de chaque document du référentiel en utilisant comme sources le thésaurus, les

ontologies générales et les questions et ensuite intégrer cette Topic Map avec la Topic Map globale. Notre méthode d'intégration est inspirée des travaux de notre équipe sur la fusion d'ontologies [Lammari et Métais, 2004] et [Lammari et Besbes-Essanaa, 2009], elle est composée de trois étapes : une étape d'appariement, une étape de fusion et une étape de validation. Dans notre approche d'appariement, nous limitons notre recherche de similarité à une recherche d'**équivalence** prenant en compte les synonymes et le multilinguisme. L'accent sera mis sur l'étape de fusion, dans laquelle nous nous intéressons à la reconstruction de hiérarchies en se basant, d'une part, sur les techniques de transformation de hiérarchies en fonctions booléennes (et vice versa) et les algorithmes de [Lammari et Métais, 2004] et d'autre part sur le thésaurus du domaine et les deux ontologies générales (WordNet et WOLF) qui fournissent les chainons manquants dans les hiérarchies à fusionner.

Après l'étape d'intégration, notre idée est de supprimer les Topics redondants afin de diminuer la taille de la Topic Map et préparer l'étape d'élagage qui fera l'objet du chapitre suivant. Pour cela, nous procédons à la recherche de similarités entre les Topics pour éventuellement les fusionner. Cependant, comme le nombre des Topics est très grand, nous commençons par définir un ensemble d'heuristiques pour le choix des Topics à comparer, ensuite vérifier s'ils sont proches et enfin les fusionner après avoir demandé la confirmation de l'utilisateur. Comme perspectives de nos travaux de recherche, un travail d'approfondissement sera mené sur la définition des heuristiques pour le choix des Topics à comparer et l'étude de la meilleure distance pour la fusion de ces Topics.

ACTOM a aussi l'avantage de prendre en compte le **multilinguisme des ressources** qu'elle représente. Ainsi, un utilisateur pourra, lors de sa navigation, avoir accès à des documents qui ne sont pas dans sa langue d'origine. Le grand intérêt de cette approche par rapport à de simples traductions de réponses est de proposer à l'utilisateur des documents correspondant à des concepts n'existant pas forcément dans sa langue ou dans sa culture. La Topic Map constituera ainsi un moyen d'enrichissement culturel pour les utilisateurs.

Enfin, nous suggérons, dans le chapitre suivant, de travailler sur l'élagage de la Topic Map générée, nous présentons les raisons qui nous ont poussées à s'intéresser à ce problème et les différentes techniques que nous proposons pour le résoudre.

CHAPITRE 5

Prise en compte de la qualité : méthode d'élagage de la Topic Map

5.1 Introduction

Le traitement de la qualité des Topic Maps recouvre différents volets, certains sont communs avec le domaine des schémas conceptuels, d'autres sont communs avec le domaine de la recherche d'information. D'autres volets seront propres à la problématique des Topic Maps. La formalisation des critères de qualité d'une Topic Map, de métriques associées et de proposition d'algorithme d'amélioration constitue un sujet en lui-même et constitue la prolongation logique des travaux de cette thèse. Nous le citerons comme une de nos principales perspectives.

Dans cette thèse nous nous sommes bornés à traiter l'aspect de la qualité lié au volume de la Topic Map, en proposant une méthode d'élagage.

Nous commençons dans ce chapitre par exposer la notion de qualité dans les systèmes d'information, puis nous décrivons les travaux sur la qualité des ontologies et des schémas conceptuels. Nous présentons, par la suite, un bref état de l'art sur les méthodes classiques de mesure de performance des systèmes de recherche d'information (SRI). Enfin, après avoir décrit quelques travaux concernant la qualité des Topic Maps, nous présentons notre solution d'élagage dynamique pour la gestion du volume dans la Topic Map dans le but d'améliorer la navigation et faciliter la recherche à travers cette Topic Map.

5.2 La qualité dans les systèmes d'information

La notion de qualité est considérée comme une partie intégrale de tout système d'information, en particulier avec l'accroissement continu du volume de données et la diversité des applications. Il existe une variété de travaux portant sur la qualité dans les différentes phases du processus de développement, ces travaux concernent différents aspects de la qualité : la qualité des données (par exemple la qualité des données d'un système d'aide à la décision a fait l'objet du projet européen DWQ, *Data Warehouse Quality*, [Jarke et al. 2000]), la qualité des modèles conceptuels (par exemple le projet QUADRIS, *Quality of Data and Multi-Source Information Systems*, proposé par [Akoka et al. 2007b] dont l'objectif est d'offrir un cadre d'évaluation de la qualité dans les systèmes d'information multisources), la qualité des processus de développement, la qualité des processus de traitement de données, la qualité des processus métier, etc [Akoka et al. 2008].

L'étude de la qualité de la Topic Map doit à priori s'intéresser aux travaux sur la qualité des ontologies et des schémas conceptuels.

5.2.1 Travaux sur la qualité des ontologies

Les premiers travaux sur la qualité des ontologies ont été menés par Nicolas Guarino qui s'interrogeait sur la mesure de la qualité de l'ontologie (du point de vue de son adéquation avec le réel) indépendamment de sa correction. Depuis la notion de qualité d'une ontologie a fait l'objet de plusieurs recherches [Hartmann et al. 2005], [Brank et al. 2005], [Supekar, 2006], [Yang et al. 2006] et [Djedidi et Aufaure, 2008]. La plupart des approches d'évaluation proposées sont multicritères, elles proposent des modèles pour l'évaluation de la qualité des ontologies, ces modèles sont fondés sur la définition d'un ensemble de critères (attributs). Pour chaque critère, l'ontologie est évaluée et un score est attribué. De plus, un poids est assigné à chaque critère.

La norme ISO/CEI 9126 décrit les exigences en termes de qualité des produits logiciels (figure 5.1). Cette norme permet de mesurer la qualité d'un produit logiciel en prenant en considération plusieurs caractéristiques. Certaines approches de qualité se sont inspirées de cette norme pour définir leurs propres critères d'évaluation. [Bansiya et David, 2002] ont adapté cette norme pour établir un modèle hiérarchique pour l'évaluation de la qualité d'une conception orientée-objet. Ce modèle comprend une arborescence d'évaluation contenant des caractéristiques, des sous-caractéristiques, des critères et des métriques d'évaluation.

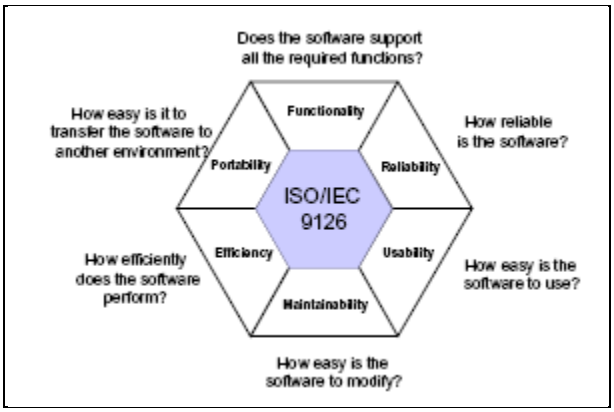


Figure 5.1 Les 6 caractéristiques de la qualité logicielle ISO/IEC 9126

Parmi les approches d'évaluation multicritères, nous citons par exemple les travaux de [Djedidi et Aufaure, 2008] qui proposent un **modèle de qualité hiérarchique** pour l'évaluation d'une ontologie décrit par la figure 5.2. Ce modèle prend en considération deux aspects de l'ontologie : l'aspect **structurel et l'usage.**

Pour chaque aspect, les auteurs ont défini un ensemble de critères. Pour la structure, les critères identifiés sont : la complexité, la cohésion, la modularité, la taxonomie et l'abstraction. Pour l'usage, trois critères ont été évalués : la complétude, la modularité et la

compréhension. A partir de ces critères, [Djedidi et Aufaure, 2008] ont proposé un modèle hiérarchique d'évaluation de l'ontologie, la racine correspond au résultat final d'évaluation, les nœuds au-dessous de la racine correspondent aux aspects pris en considération lors de l'évaluation (structure, usage), les nœuds du niveau suivant correspondent aux critères d'évaluation et les feuilles aux mesures d'évaluation ou métriques (par exemple le nombre moyen de relations par concept, la profondeur d'une hiérarchie, etc.). Cette évaluation est appliquée au fur et à mesure du processus de construction et d'enrichissement de l'ontologie.

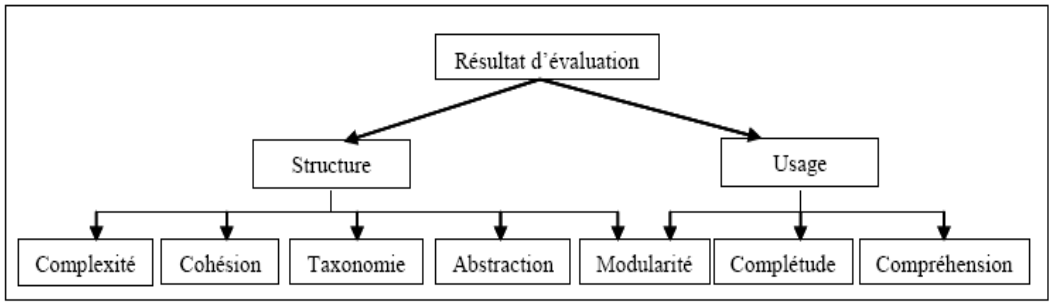


Figure 5.2 Modèle hiérarchique d'évaluation d'une ontologie [Djedidi et Aufaure, 2008]

[Burton-Jones et al. 2004] ont proposé une approche avec 10 critères : l'éligibilité (fréquence des erreurs syntaxiques), la richesse, l'interprétabilité (la présence des termes utilisés dans WordNet), la consistance (nombre de concepts impliqués dans des contradictions), la clarté (les termes utilisés dans l'ontologie ont-ils plusieurs sens dans WordNet?), la compréhensivité, l'exactitude (pourcentage de fausses relations), la pertinence, l'autorité (combien d'autres ontologies utilisent les concepts de l'ontologie à évaluer?), l'historique (nombre d'accès à l'ontologie).

[Lozano-Tello et Gomez-Perez, 2004] ont proposé 117 critères. Ces critères couvrent plusieurs aspects : le contenu de l'ontologie (concepts, relations, taxonomie, axiomes), la méthodologie utilisée lors de la construction de l'ontologie, le coût d'utilisation de l'ontologie et les outils disponibles.

[Gomez-Perez et Rojas-Amaya, 1999] proposent trois critères pour mesurer la qualité d'une ontologie : (1) La précision de la modélisation elle-même (clarté, standardisation du vocabulaire, suppression des concepts quasi-homonymes) ; (2) La fiabilité de l'ontologie (complétude, cohérence, extensibilité) et (3) La qualité de la structuration (disjonction des classes, utilisation de l'héritage multiple, modularité).

[Fox et al. 1998] ont proposé un autre ensemble de critères : la complétude fonctionnelle (l'ontologie contient-elle assez d'information?), la généralité (l'ontologie est-elle assez générale pour qu'elle soit partagée par plusieurs utilisateurs?), l'efficacité du

raisonnement supporté par l'ontologie, la compréhension, la précision/granularité (l'ontologie supporte-t-elle plusieurs niveaux d'abstraction/détail?), la minimalité (l'ontologie contient-elle tous les concepts nécessaires?).

5.2.2 Travaux sur la qualité des schémas conceptuels

Il existe des travaux visant à définir la qualité d'un modèle conceptuel [Sisaïd-Cherfi et al., 2002], [Sisaïd-Cherfi et al., 2006], [Akoka et al. 2007a], [Akoka et al. 2007b], [Akoka et al. 2008]. Ces approches travaillent sur différentes dimensions de la qualité des modèles conceptuels. Nous pouvons citer par exemple la clarté (mesurant la facilité à lire le modèle, selon une considération visuelle), la simplicité (selon la nature des concepts), l'expressivité (richesse du modèle), la justesse (correction du modèle), la complétude (niveau de couverture des besoins), la compréhension, etc.

[Akoka et al. 2007b] ont proposé un méta-modèle de la qualité, dans le cadre de leur projet QUADRIS [Akoka et al. 2007b]. Ce méta-modèle est centré sur la description des différentes dimensions de qualité. Chaque dimension peut être déclinée en plusieurs facteurs. A chaque facteur peut être associé un ensemble de métriques différentes, et à une métrique donnée peuvent correspondre différentes méthodes de mesure. Les auteurs proposent également une démarche de prise en compte de la qualité qui comprend trois volets :

1) La définition d'un cadre pour l'évaluation de la qualité qui couvre tous les aspects du développement (figure 5.3): **Usage, Spécification et Implémentation** ;
2) Le développement d'un environnement dédié à la qualité qui met en œuvre l'approche qualité proposée et permet l'automatisation de la mesure de la qualité ;
3) La conduite d'une évaluation de l'approche impliquant des professionnels de divers domaines de l'informatique (conception, développement, tests, utilisateurs, etc.).

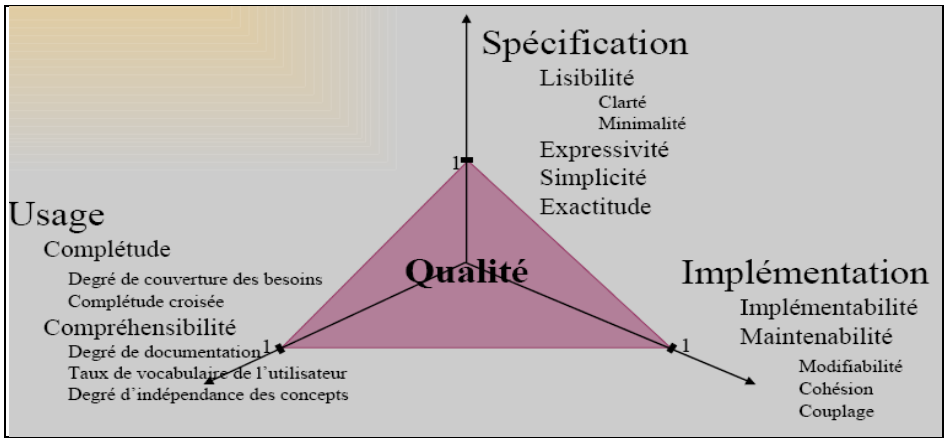


Figure 5.3 Framework proposé par [Akoka et al. 2007a] pour l'évaluation des modèles conceptuels.

Dans leurs travaux récents, [Akoka et al. 2008] proposent un cadre d'évaluation de la qualité dans les systèmes d'information multi-sources. Ce cadre permet de définir un méta-modèle pour étudier en particulier les interdépendances entre les dimensions de la qualité d'un modèle conceptuel de données et celles de la qualité des données instanciant ce modèle. Dans le cadre du projet QUADRIS [Akoka et al. 2007b], ces propositions ont été validées dans trois domaines d'application : le domaine biomédical, le domaine commercial et le domaine géographique.

5.3 Travaux sur la qualité dans les systèmes de recherche d'information

Le but de la RI est de trouver des documents pertinents à une requête, et donc utiles pour l'utilisateur. L'évaluation d'un SRI consiste à estimer son efficacité à retrouver des documents pertinents. La pertinence est une notion très complexe à évaluer. En effet, elle dépend fortement de l'utilisateur, qui est véritablement le seul à savoir si le document retourné par le système correspond à son besoin d'information initial. Il est cependant indispensable de disposer de techniques d'évaluation qui, en définissant des mesures précises, permettent de juger de la performance des SRI, quels que soient les méthodes d'indexation, de recherche ou les modèles qu'ils implémentent.

5.3.1 Critères de qualité

Les techniques d'évaluation des SRI s'appuient essentiellement sur l'estimation de la qualité des informations retrouvées par les systèmes, c'est-à-dire les documents retrouvés

sont-ils pertinents ou non pertinents ? D'autres critères peuvent toutefois être pris en considération, comme par exemple :

- Le temps mis par le système pour fournir des réponses à l'utilisateur ;
- L'effort effectué par l'utilisateur pour obtenir l'information recherchée (par exemple, le nombre de requêtes qu'il a dû formuler avant d'avoir le résultat recherché) ;
- La qualité de la présentation des résultats par le système (par exemple, à partir de la liste de résultats fournis par le système, combien de documents l'utilisateur a-t-il dû parcourir avant de trouver le document recherché ?).

Nous évoquons dans cette section uniquement les techniques liées à l'évaluation de la qualité des informations retrouvées par le SRI. La qualité d'un système de recherche d'information doit être mesurée en comparant les réponses du système avec les réponses idéales que l'utilisateur espère recevoir. Plus les réponses du système correspondent à celles que l'utilisateur espère, meilleur est le système. Diverses techniques et mesures ont été proposées pour répondre à ce problème.

5.3.2 Campagnes d'évaluation

Pour pouvoir être évalués et comparés les uns aux autres, les SRI doivent être appliqués sur un même jeu de données et utiliser des méthodologies d'évaluation identiques. De nombreux projets d'évaluation ont vu le jour depuis le début les années 70 (exemple le projet Cranfield) afin de construire des collections de test complètes et définir des protocoles d'expérimentations précis. Aujourd'hui, l'initiative la plus importante est la campagne d'évaluation TREC[38] (*Text REtrieval Conference*) qui, en plus de fournir des collections de test volumineuses, propose une infrastructure bien définie pour l'évaluation des méthodologies de recherche. TREC a permis d'encourager le développement d'autres campagnes d'évaluation, telles que CLEF[39] (*Cross Language Evaluation Forum*) pour l'évaluation de SRI multilingues, Amaryllis (pour la langue française) ou INEX (*Initiative for the Evaluation of XML Retrieval*) pour la RI structurée.

[38] http://trec.nist.gov/
[39] http ://www.clef-campaign.org/

5.3.3 Les mesures du Rappel, de la Précision et de F-mesure

Les deux mesures communément utilisées depuis plus de 30 ans pour évaluer un système de recherche d'information sont le taux de précision et celui de rappel.

Le Rappel

Le rappel mesure la proportion de documents pertinents retrouvés parmi tous les documents pertinents dans la base. La proportion complémentaire est le ***Silence*** qui correspond à la proportion de documents pertinents non retrouvés (Equation 5.1).

$$Rappel = \frac{|P \cap R|}{|P|} \in [0, 1] \text{ et } Silence = 1 - Rappel$$

Equation 5.1

Avec : *P* représente le nombre de documents pertinents dans tout le corpus.

R représente le nombre de documents retrouvés.

La Précision

La précision mesure la proportion de document pertinent retrouvé parmi tous les documents retrouvés par le système. La proportion complémentaire est le ***Bruit*** qui correspond à la proportion de documents retrouvés qui ne sont pas pertinents (Equation 5.2).

$$précision = \frac{|P \cap R|}{|R|} \in [0, 1] \text{ et } Bruit = 1 - précision$$

Equation 5.2

Plusieurs indicateurs de synthèse ont été créés à partir de deux mesures de Rappel et de la Précision, mais le plus célèbre est la F-mesure.

La F-mesure

Cette mesure correspond à une moyenne harmonique de la précision et du rappel. Cette moyenne diminue lorsque l'un de ses paramètres est petit et augmente lorsque les deux paramètres sont proches tout en étant élevés (Equation 5.3).

$$F - mesure = \frac{(1 + \beta^2)\, précision \times rappel}{(\beta^2 \times précision) + rappel}$$

Equation 5.3

Le paramètre β permet de pondérer la précision ou le rappel, il est égal généralement à la valeur 1. Pour effectuer ces mesures, il faut disposer des réponses idéales aux requêtes en question. La figure 5.4 illustre ces formules.

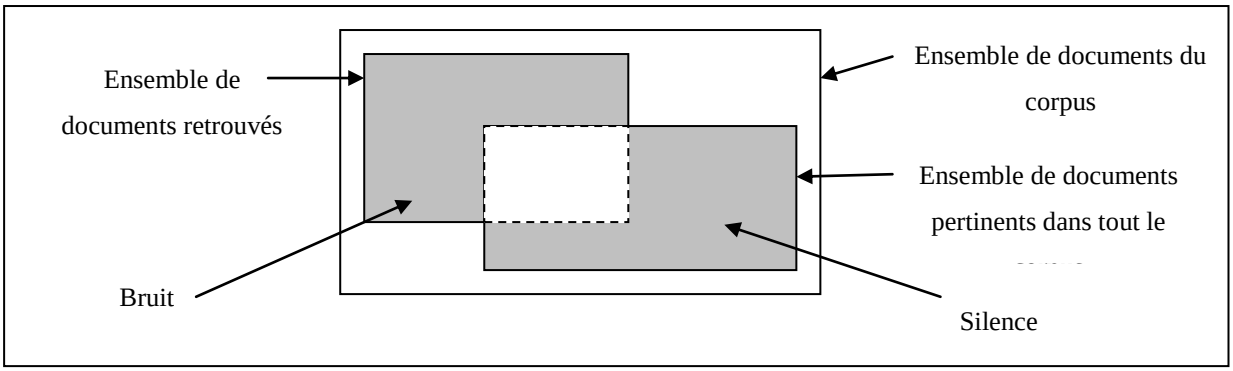


Figure 5.4 Rapprochement de pertinences système et utilisateur

Dans la prochaine section, nous présentons quelques travaux proposés pour l'étude de la qualité des Topic Maps.

5.4 Travaux sur la qualité d'une Topic Map

Nous remarquons, d'après la littérature, que très peu de travaux [Dicks et al. 2004], [Legrand et Soto, 2006], [Gyun et Park, 2007], [Godehardt et Bhatti, 2008] se sont intéressés à l'évaluation d'une Topic Map et à l'étude de sa qualité. [Gyun et Park, 2007], [Dicks et al. 2004] ont choisi d'étudier la qualité de la recherche à travers une Topic Map en la comparant avec des moteurs des recherche classiques selon des critères définis par les auteurs alors que [Legrand et Soto, 2006] et [Godehardt et Bhatti, 2008] travaillent sur la qualité de la Topic Map, par exemple Legrand dans ses travaux propose d'utiliser des techniques de représentation et de visualisation pour améliorer l'exploration de la Topic Map, ces techniques sont basées sur le filtrage et le regroupement des données de la Topic Map en utilisant les treillis de Galois.

Nous proposons de classer les approches d'évaluation de la qualité des Topic Maps en deux classes d'approches : Les approches qui s'intéressent à la qualité de la visualisation de la Topic Map et celles qui proposent d'étudier la qualité de la recherche à travers la Topic Map.

5.4.1 Les approches qui s'intéressent à la qualité de la visualisation de la Topic Map

Dans cette classe d'approches, nous citons par exemple la méthode proposée par **[Legrand et Soto, 2006]**, qui a pour objectif d'améliorer la visualisation de la Topic Map en se basant sur des techniques de filtrage et de regroupement conceptuel des données dans la Topic Map. Tous les regroupements possibles sont effectués, qu'ils réunissent des données très ressemblantes ou n'ayant que peu de facteurs en commun. De plus, les données peuvent appartenir à plusieurs classes, et non nécessairement à une seule. La technique choisie par les auteurs est l'analyse conceptuelle au moyen de treillis de Galois, ils s'intéressent tout particulièrement à un algorithme de classification conceptuelle basé sur l'Analyse Formelle de Concepts (*Formal Concept Analysis* ou FCA) et les correspondances de Galois.

L'algorithme de Galois consiste à regrouper les objets étudiés dans des classes qui matérialisent les concepts du domaine étudié. Les objets individuels sont discriminés en fonction des propriétés qu'ils ont en commun.

Dans leurs travaux, [Legrand et Soto, 2006] proposent également des techniques de représentation et de navigation pour faciliter l'exploration de Topic Map. L'idée consiste à représenter les Topic Maps par des villes virtuelles, à l'intérieur desquelles les utilisateurs peuvent se déplacer pour créer leur propre carte cognitive.

[Godehardt et Bhatti, 2008] utilisent les Topic Maps pour la visualisation de sources de données hétérogènes. Leur approche vise à améliorer l'affichage de la Topic Map à cause de la diversité et la grande quantité d'information qu'elle représente. L'idée consiste à utiliser la notion de cluster et de secteur avec l'outil TM Viewer[40] et afficher la Topic Map dans sa globalité selon différents niveaux pour que les utilisateurs puissent gérer le grand nombre de Topics. Les critères choisis pour rassembler les Topics dans un même niveau ne sont pas spécifiés dans les articles publiés par les auteurs, ils précisent juste que les Topics ayant les mêmes objectifs sont regroupés au même niveau. Ce projet à été inspiré des travaux de [Weerdt et al. 2006] qui consistent à implémenter un outil, appelé TopicMaker, pour la visualisation de Topic Maps en trois dimensions et selon plusieurs niveaux. La figure 5.5 illustre un exemple d'affichage de la Topic Map multi-niveaux avec l'outil TM Viewer.

[40] Ontologies for Education Group. Simple topic map viewer: very general and java applet based o4e (ontologies for education) portal http://o4e.iiscs.wssu.edu/drupal/

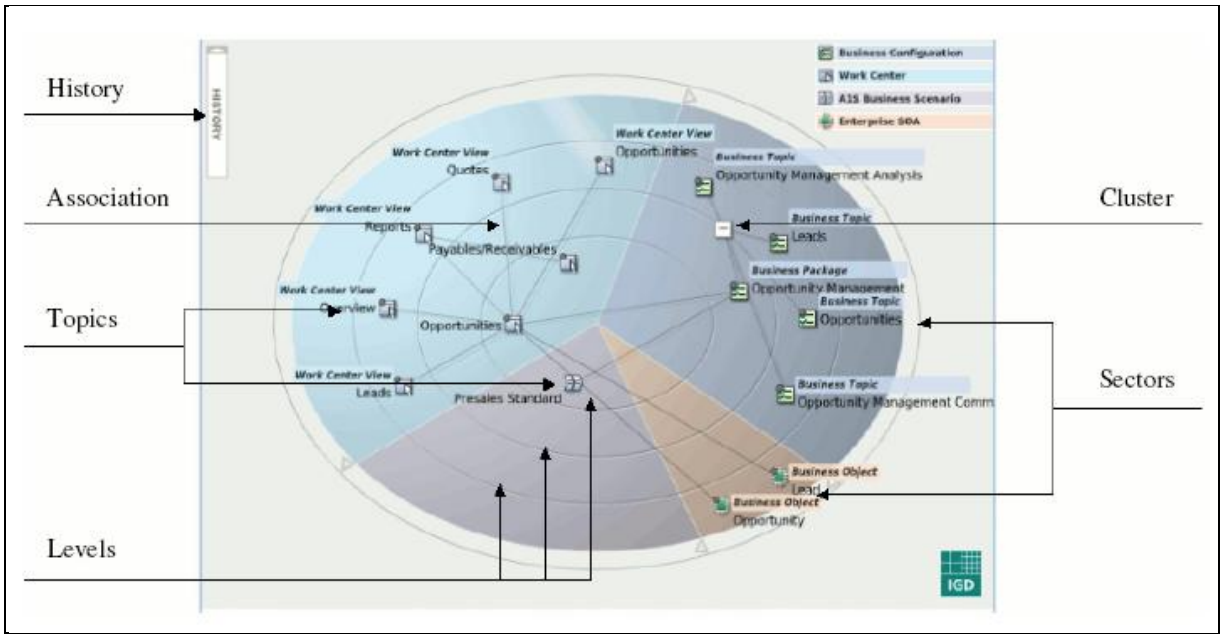


Figure 5.5 Exemple d'affichage de la Topic Map multi-niveaux avec l'outil TM Viewer [Godehardt et Bhatti, 2008]

5.4.2 Les approches qui s'intéressent à la qualité de la recherche à base de Topic Map

Dans cette classe d'approches, nous citons par exemple la méthode présentée dans **[Dicks et al. 2004]** qui travaillent sur les performances de la recherche à travers les Topic Maps. Cette méthode consiste à évaluer une application Web fondée sur le modèle des Topic Maps et développée pour la recherche d'information, cette application est implémentée et testée dans le domaine de l'éducation. [Dicks et al. 2004] proposent de la comparer avec un moteur de recherche en utilisant les mesures du rappel et de précision calculés pour les deux outils. De plus les auteurs, dans leur démarche d'évaluation, prennent en compte les avis d'un échantillon de personnes composé d'étudiants et d'enseignants qui ont testé les deux systèmes. Les résultats obtenus ont montré que la recherche à base de Topic Map donne un meilleur rappel que le moteur de recherche.

La même idée est adoptée par **[Gyun et Park, 2007]** pour l'évaluation de leur système de recherche de musique folklorique coréenne (appelée Pansori) basée sur les Topic Maps, les auteurs proposent de comparer leur outil avec un système de recherche existant et déjà utilisé par les étudiants dans les universités coréennes. Un ensemble de mesures a été défini pour cette évaluation, des mesures subjectives (complétude, facilité d'utilisation, efficacité, satisfaction, convenance) et des mesures objectives telles que le temps mis par le système pour retrouver l'information recherchée.

5.5 Problématiques particulières à la qualité des Topic Maps

D'après cet état de l'art sur la qualité des SRI, des ontologies, des schémas conceptuels et des Topic Maps, nous remarquons que la qualité des Topic Maps n'a pas été suffisamment étudiée par rapport aux travaux réalisés sur la qualité des ontologies et des schémas conceptuels. De plus, la notion de qualité de Topic Maps n'est pas la même que celle relative aux ontologies et aux schémas conceptuels vu les différences entre ces modèles. En effet, comme nous l'avons déjà expliqué dans le chapitre 2 (section 2.1.2.6), les Topic Maps sont **orientées utilisation directe par l'utilisateur**, elles cartographient le contenu de documents, alors qu'une ontologie est une spécification formelle et explicite d'un domaine afin de permettre l'échange d'informations entre applications.

Par conséquent, les travaux sur la qualité des ontologies se basent essentiellement sur la définition de critères objectifs et de métriques pour l'évaluation de l'ontologie par rapport à ces critères et à ces métriques alors que la plupart des travaux sur la qualité des Topic Maps s'intéressent surtout à la qualité de la visualisation et à l'affichage de la Topic Map vu que cette dernière est très souvent volumineuse et contient des milliers de Topics et de liens de différents types.

Comme nous l'avons déjà mentionné dans la section précédente sur les travaux concernant la qualité des Topic Maps, selon nous, la notion de qualité de Topic Map peut être vue selon **deux facettes** :

1) **La qualité de la Topic Map** c'est-à-dire la qualité de l'affichage, la qualité la visualisation de la Topic Map ;
2) **La qualité de la recherche** à travers la Topic Map c'est-à-dire la pertinence des résultats de recherche par rapport aux besoins des utilisateurs.

Ces deux facettes sont très liées puisque la qualité de la Topic Map impacte les performances de la recherche. En d'autres termes, il est difficile pour un utilisateur de retrouver l'information qu'il cherche dans une Topic Map mal structurée et mal affichée, il peut facilement se perdre en naviguant à travers ses liens et perdre ainsi beaucoup de temps.

Nous nous intéressons dans le cadre de notre travail à la première facette qui concerne la qualité de l'affichage de la Topic Map. Un des problèmes majeurs relié à la qualité des Topic Maps est que la Topic Map générée est très souvent **volumineuse** et contient trop d'information. Les Topic Maps sont généralement de très grande taille, puisqu'elles peuvent contenir des milliers de Topics et d'associations. Ce grand volume d'information et cette complexité peut entrainer, **une mauvaise organisation de la Topic Map, une difficulté au**

niveau de la recherche d'information et un encombrement au niveau de l'affichage de la Topic Map. Cela peut s'expliquer du fait qu'une Topic Map, telle qu'elle a été conçue, est orientée usage donc elle doit représenter divers points de vue et différentes visions à propos des sujets du domaine d'étude selon les différentes catégories d'utilisateurs susceptibles de s'intéresser au contenu de la Topic Map.

Face à cette énorme quantité d'information dans une Topic Map, il serait difficile pour les utilisateurs surtout ceux qui ne sont pas des experts du domaine, de trouver facilement ce qu'ils cherchent dans des temps raisonnables.

Une particularité des Topic Maps par rapport aux ontologies et aux schémas conceptuels est donc la prépondérance du **problème de volume** car elles sont destinées à être vues et utilisées directement par l'utilisateur.

5.6 Notre approche de gestion du volume de la Topic Map

Dans le cadre de notre approche ACTOM, pour gérer le volume de la Topic Map, nous proposons une méthode d'élagage dynamique de la Topic Map au niveau de l'affichage et ce grâce à la définition de métadonnées.

Une Topic Map étant utilisée essentiellement pour l'organisation d'un contenu et pour la recherche d'information dans ce contenu, il est souhaitable de réviser sa structure de façon périodique afin qu'elle puisse répondre de façon efficace aux besoins de recherche d'information et qu'elle puisse évoluer en accord avec les évolutions effectuées sur le contenu. Notre méthode d'élagage a pour objectif d'une part la gestion des évolutions de la Topic Map pour prendre en compte les éventuels changements relatifs au contenu et à l'usage de la Topic Map et d'autre part l'amélioration de l'affichage de la Topic Map afin de permettre aux utilisateurs d'accéder facilement à l'information recherchée et pouvoir naviguer dans toute la Topic Map d'une manière intelligente et intuitive.

Notre idée consiste à sélectionner ce qui est important dans la Topic Map et ce qui est moins important pour les utilisateurs. Le processus d'élagage de la Topic Map consiste à supprimer tous les Topics et leurs liens respectifs considérés non pertinents et ne garder que ceux qui sont important pour les utilisateurs, cependant dans notre approche, nous avons choisi de tout garder et de ne supprimer aucun Topic mais en revanche, nous proposons d'attribuer une note d'importance (ou un score) à chaque Topic et en fonction du temps et d'autres conditions, nous montrons à l'utilisateur que les Topics ayant un score élevé (ou supérieur à un certain seuil paramétrable par l'utilisateur).

Les scores affectés à chaque Topic sont également utilisés comme critère de choix de visualisation. En effet, un Topic ayant un très bon score pourrait être considéré comme Topic principal et dans ce cas apparaître dans une visualisation par défaut de la Topic Map.

Notre idée de choisir de garder tout dans la Topic Map et d'associer une note à chaque Topic s'explique du fait qu'on ne peut pas confirmer de manière définitive qu'un Topic est pertinent ou non. En effet, le degré d'utilisation d'un Topic peut varier au cours du temps pour différentes raisons par exemple les variations saisonnières. Nous remarquons qu'en été, beaucoup d'utilisateurs demandent de consulter des documents sur le Topic « climatisation » alors qu'en hiver, ce même Topic est très peu consulté vu qu'en hiver les gens s'intéressent surtout au moyen et aux systèmes de chauffage. Dans ce cas, la note du Topic « climatisation » va augmenter en été et diminuer en hiver et inversement pour le Topic « chauffage ».

Un autre exemple de variation de la note d'un Topic, les sujets d'actualité tels que le crash de l'avion d'Air France en juin 2009, nous remarquons que, dans cette période, les gens s'interrogent beaucoup sur ce sujet et veulent consulter tous les documents, les articles et les témoignages qui en parlent. Dans cette période, la note du Topic « crash de l'avion d'Air France » a beaucoup augmenté par contre, actuellement, sa note a diminué puisque très peu de gens veulent consulter ce Topic et s'intéressent à ce sujet.

Actuellement, il y a des sujets auxquels les utilisateurs s'intéressent de plus en plus, cet intérêt ne cesse d'augmenter au cours du temps, nous citons par exemple le domaine de l'environnement, des solutions pour l'économie d'énergie et les moyens pour préserver l'environnement. Par conséquent, la note des Topics représentant ces sujets augmente de façon continue au cours du temps.

5.6.1 Notation de Topics

Méta-propriété 1 : Importance du Topic dans la Topic Map

Dans un premier temps, nous proposons comme méta-propriété, la note (ou score) d'un Topic qui nous renseigne sur son importance dans la Topic Map et sur l'usage qu'on en fait lors de l'exploitation de la Topic Map. Ces méta-propriétés sont utilisées pour la gestion des évolutions de la Topic Map, plus précisément pour l'élagage dynamique de Topics considérés non pertinents. La note d'un Topic est initialisée dès la création de la Topic Map (figure 5.6). Elle peut être (a) **très bonne** dans le cas où le Topic concerné est obtenu à partir des trois

sources (les documents, le thésaurus et les requêtes) utilisées pour élaborer la Topic Map globale, (b) **moyenne** dans le cas où le Topic est issu de deux sources ou encore (c) **moins bonne** s'il a été extrait d'une seule source.

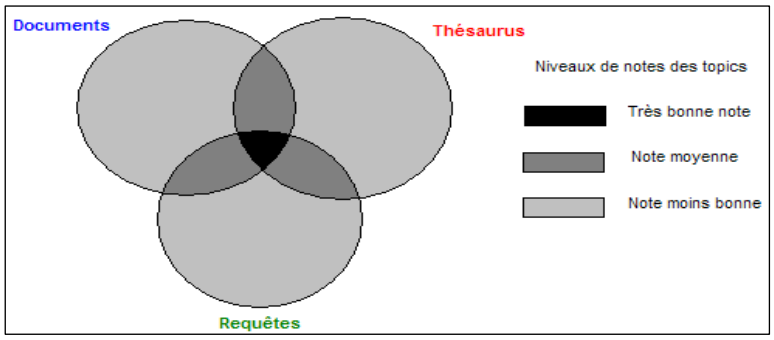


Figure 5.6 Initialisation des notes des Topics

Ces mesures de qualité doivent être transformées en une note comprise entre 0 et 1 pour permettre la réalisation du processus d'élagage dynamique au niveau de la visualisation de la Topic Map : Seuls les Topics ayant un score supérieur à un seuil fixé seront affichés. Au cours du temps et pendant la vie de la Topic Map, le score permettra de refléter le niveau de popularité de chaque Topic. Pour cela, la note est définie comme une moyenne pondérée de plusieurs critères qui sont : le nombre de documents indexés par le Topic (**ND**), le nombre de FAQs qui font référence à ce Topic (**NF**) et le nombre de consultations qui concernent ce Topic (**NC**). Nous définissions alors la formule de calcul de la note comme suit (Equation 5.4) :

$$Notetopic = \frac{\alpha \times ND + \beta \times NF + \gamma \times NC}{\alpha + \beta + \gamma}$$

Equation 5.4

Les poids sont paramétrables, en revanche, nous suggérons de paramétrer γ plus grand que α et β et ce dans le but de mieux refléter l'usage dans la Topic Map.

Méta-propriété 2 : Niveau auquel appartient le Topic dans la Topic Map

Comme nous l'avons déjà évoqué dans notre méta-modèle de Topic Map (Chapitre 3), en plus de la note du Topic, nous avons défini une deuxième méta-propriété associée à chaque Topic. Cette méta-propriété nous renseigne sur le niveau auquel appartient le Topic dans la Topic Map. Nous avons choisi d'organiser notre Topic Map en trois niveaux : le plus haut niveau contient les Topics thèmes et les Topics questions ; le niveau intermédiaire englobe les Topics concepts du domaine, les instances de Topics, les synonymes de Topics ; et le niveau

le plus bas contient les documents sources et leurs segments ainsi que les sources d'interrogation relatives à ces documents (FAQ). Nous utilisons cette métadonnée pour organiser les Topics et améliorer l'affichage de la Topic Map.

Dans nos futurs travaux, nous comptons définir d'autres critères de qualité pour l'évaluation de la Topic Map et proposer une liste exhaustive de méta propriétés qui nous serviront pour la gestion de l'évolution de la Topic Map.

5.6.2 Analyse des notes

Nous nous intéressons à la première méta-propriété qui est la note du Topic, nous proposons non seulement d'affecter cette note au Topic mais aussi d'analyser sa variation qui peut changer de différentes manières et pour plusieurs raisons. Par exemple, dans le cas des Topics « chauffage » et « climatisation », la note varie de façon **cyclique** alors que dans le cas du Topic « crash de l'avion d'Air France » ou par exemple « la mort de la princesse Diana », la note varie de façon **maximale puis minimale** puisque, en ce moment, les gens ne s'intéressent presque pas à ces sujets. Dans le cas de Topics tels que « environnement », « économie d'énergie » ou « construction durable », nous remarquons que leur note **augmente** de façon **continue** au cours du temps puisque les utilisateurs s'intéressent de plus en plus à ces sujets.

Vu le problème de variations des notes d'un Topic, nous proposons de stocker un historique de ces notes c'est-à-dire la date et la note pour pouvoir par la suite analyser la périodicité de la notation. Pour cela, nous proposons de définir une liste de **méta-méta-propriétés** pour analyser les variations des notes. En effet, comme nous l'avons déjà présenté dans l'exemple précédent, il y a des variations saisonnières (comme par exemple « chauffage » et « climatisation »), il y a aussi des variations reliées aux sujets d'actualités tels que « le crash de l'avion d'air France » ou « le mariage de Sarkozy avec Carla Bruni » mais nous trouvons également des variations liées aux sujets tendance du moment tels que : économie d'énergie, environnement et construction durable.

Une **méta-méta-propriété** est définie comme le type de la variation de la note d'un Topic (selon la saison, selon le temps, augmentation de la note, diminution de la note). Cette méta-méta-propriété peut par exemple nous permettre d'anticiper le score d'un Topic, de **changer automatiquement** la note d'un Topic à partir du moment où on connait son type et gérer dynamiquement l'élagage au moment de l'affichage de la Topic Map.

5.6.3 Utilisation des méta-propriétés pour améliorer l'affichage de la Topic Map

L'objectif de la Topic Map est de permettre aux utilisateurs de rechercher des informations et d'accéder au contenu des documents sources. Elle doit tenir compte, au niveau de la visualisation, de la complexité et de la grande quantité des informations à représenter. L'objectif n'est pas simplement de montrer ces données, mais de permettre à l'utilisateur de les comprendre (même partiellement), ou tout au moins d'en extraire quelques informations.

Dans notre approche, l'idée réside dans le fait que nous utilisons les deux méta-propriétés définies précédemment (la note et le niveau auquel appartient le Topic) pour l'élagage dynamique de la Topic Map au moment de l'affichage et ce dans le but de faciliter l'accès aux documents à travers la Topic Map.

Nous utilisons **la méta-propriété du niveau d'un Topic pour améliorer l'affichage en organisant la Topic Map en trois niveaux** : le premier contient les Topics thèmes et les questions, le deuxième contient les Topics concepts du domaine, les instances de Topics, éventuellement, les Topics réponses qui peuvent aussi appartenir au premier niveau et enfin le niveau ressource, contenant les documents et leurs segments.

Cette organisation fournit à l'utilisateur différents niveaux de détail sur le contenu de la Topic Map et lui permet de passer d'un niveau à un autre en fonction de ses choix de navigation. En effet, au départ, nous choisissons de n'afficher à l'utilisateur que les Topics questions et les Topics thèmes de premier niveau, ensuite, lors de sa navigation l'utilisateur pourrait s'intéresser à un thème particulier, il aura alors la possibilité de poursuivre sa recherche et parcourir le champ de proximité sémantique correspondant à ce Topic thème c'est-à-dire le sous arbre de la Topic Map qui contient tous les Topics concepts du domaine reliés au thème choisi par des liens ontologiques et des liens d'usage. Au niveau des feuilles de ce sous arbre, l'utilisateur pourra accéder aux documents et aux segments qui traitent de ce thème. De cette manière, l'utilisateur aura la possibilité de construire sa propre carte cognitive contenant les données qui l'intéressent (en fonction des parties qu'il a visitées).

Pour pouvoir réaliser cet affichage multi-niveaux de la Topic Map, il est nécessaire de bien choisir les techniques et les outils de visualisation adéquats qui nous permettent d'afficher tous les Topics et leurs associations, malgré leur nombre très elevé. Cet affichage permettra de faciliter l'exploration et la recherche d'information à travers la Topic Map.

En plus de l'affichage multi-niveaux, **l'attribution d'une note à un Topic nous a également servi comme critère de choix de visualisation.** En effet, un Topic ayant une note

très bonne pourrait être considéré comme Topic principal et de part ce fait apparaître dans une visualisation par défaut de la Topic Map. Nous définissons une règle pour l'affichage des Topics, cette règle est la suivante : seuls les Topics ayant une note supérieur à un seuil seront affichés par défaut, ce seuil est paramétrable, pour notre cas, nous avons choisi de le fixer à 0.5, les autres Topics seront affichés **en gris** mais l'utilisateur pourrait quand même les consulter s'il veut. Par exemple, selon la saison, il y a des Topics en gris tel que « climatisation » en hiver et les autres sont affichés par défaut, (par exemple « climatisation » en été).

La partie qui concerne la visualisation de la Topic Map sera présentée en détail dans le chapitre suivant « Plateforme de mise en œuvre de l'approche proposée ».

5.7 Conclusion

Dans ce chapitre, nous avons exposé les problèmes liés à la qualité des Topic Maps, à savoir la complexité et le volume qui est le principal problème rencontré par les créateurs des Topic Maps vu que cette dernière doit tout représenter sur un domaine donné selon plusieurs points vue et est destinée à être vue directement par les utilisateurs. Nous avons également réalisé un état de l'art sur la qualité des SRI, en particulier les mesures utilisées pour mesurer les performances de la recherche dans les SRI, nous avons aussi présenté quelques travaux de recherche sur la qualité des ontologies, des schémas conceptuels ainsi que ceux qui s'intéressent la qualité des Topic Maps. Nous avons par la suite proposé de classifier les approches existantes concernés par la qualité des Topic Maps, bien qu'elles ne sont pas très nombreuses, celles qui travaillent sur la qualité de la visualisation de la Topic Map et celles qui concernent la qualité de la recherche à travers les Topic Maps. Nous avons remarqué que une des particularités des Topic Maps par rapport aux ontologies et aux schémas conceptuels est **la prépondérance du problème de volume,** par conséquent l'objectif principal à travers l'étude de la qualité des Topic Maps est de gérer le grand nombre de Topics et d'associations dans une Topic Map.

Dans notre approche ACTOM, nous avons essayé de résoudre ce problème en proposant une méthode d'élagage dynamique à l'affichage de la Topic Map grâce à des métadonnées associés aux Topics. Nous avons défini deux types de métadonnées : **(1) la première est la note du Topic**, elle est initialisée à la création de la Topic Map, cette méta-propriété nous renseigne sur l'importance des Topics et sur l'usage qu'on en fait lors de l'exploitation de la Topic Map. Elle est utilisée pour la gestion des évolutions de la Topic Map, plus précisément

pour l'élagage de Topics considérés non pertinents. La note d'un Topic dépend de sa présence dans les trois sources d'informations utilisés comme entrée pour la construction de la Topic Map ; le thésaurus, les documents et les scénarios d'usage ; (2) **la deuxième méta-propriété** indique **le niveau** auquel appartient le Topic dans la Topic Map organisée en trois niveaux selon le méta-modèle que nous avons défini précédemment.

Nous utilisons ces méta-propriétés pour améliorer l'affichage de la Topic Map, en effet, cette dernière est affichée en trois niveaux le premier contient les Topics thèmes et les questions, le deuxième contient les Topics concepts du domaine, les instances de Topics, éventuellement, les Topics réponses qui peuvent aussi appartenir au premier niveau et enfin le niveau ressource, composé de nœuds faisant référence aux documents et à leurs segments. La note du Topic est utilisé pour afficher la Topic Map selon l'usage, en fixant un seuil pour la note, les Topics ayant une note inférieur au seuil seront affichés en gris clair et les autres, considérés des Topics pertinents sont affichés par défaut. Cette visualisation permet de faciliter la navigation dans la Topic Map et améliorer ainsi les performances et la qualité de la recherche à base de Topic Map.

Nous avons également proposé de sauvegarder un historique des notes qui comprend la note et sa date, ce ci nous a permis d'analyser les types de variations que peut avoir la note d'un Topic par exemple les variations saisonnières où la note varie de façon cyclique ou les variations liés au sujet d'actualité où la note varie de façon maximale puis minimale. Nous avons défini le type de variation comme **une méta-méta-propriété** qui nous avons utilisé pour changer automatiquement la note d'un Topic.

Nous projetons, dans nos prochains travaux de faire une étude permettant de réunir les critères de qualité d'une Topic Map afin de déterminer de façon plus ou moins exhaustive la liste de méta-propriétés utiles à la gestion des évolutions d'une Topic Map.

En plus de la solution de l'élagage dynamique de la Topic Map pour ne garder que les Topics qui intéressent les utilisateurs et prendre en compte les changements qui peuvent survenir au niveau de l'utilisation (consultation) de Topics, nous comptons étudier d'autres solutions pour résoudre le problème de volume dans une Topic Map par exemple :

- Création de vues synthétiques dans la Topic Map au niveau de l'affichage ;
- Clustering des données pour une meilleure visualisation de la Topic Map.

Dans le chapitre suivant, nous proposons d'implémenter et de mettre en œuvre les modèles et les techniques que nous avons présentés tout au long de ce mémoire. Une étape d'expérimentation et de validation est ensuite réalisée pour tester notre plateforme.

CHAPITRE 6

Plateforme de mise en œuvre de l'approche proposée

Ce chapitre est consacré à la description de la plateforme de recherche intelligente que nous avons développée afin de mettre en œuvre l'approche, les idées et les modèles que nous avons proposés dans ce travail de thèse. Notre plateforme permet la création semi automatique d'une Topic Map à partir de documents textuels multilingues, cette Topic Map permet l'annotation sémantique de ces documents pour en faciliter la recherche. L'utilisateur aura alors la possibilité de naviguer dans la Topic Map et accéder à l'information qu'il cherche à travers les liens de la Topic Map et il pourra également l'interroger par des requêtes classiques.

6.1 Domaine d'application : La construction durable

Nous avons choisi le domaine de la construction durable pour l'opérationnalisation de notre plateforme en particulier le sous domaine relatif aux solutions pour l'économie d'énergie. Le choix de ce domaine est justifié par le fait que c'est un sujet d'actualité, tout le monde s'intéresse actuellement aux moyens pour préserver l'environnement, aux solutions pour l'économie d'énergie, etc. Par ailleurs, à notre connaissance aucune application utilisant les Topic Maps n'a été testée dans ce domaine, la plupart des travaux utilisant les Topic Maps pour la recherche travaillent dans les domaines de l'éducation, du e-learning, de la musique, etc.

6.1.1 Présentation du thésaurus CTCS

Dans notre travail, nous utilisons un thésaurus du domaine de la science et de la construction *Canadian Thesaurus of Construction Science and Technology CTCS*[41] développé par l'université de Montréal. Le thésaurus CTCS est très générique et couvre tout le domaine de la construction et du bâtiment. Ce thésaurus comporte actuellement 15331 termes répartis sur 10 niveaux de hiérarchies. Chacun de ces termes est décrit dans un fichier html. Les relations entre termes sont des hyperliens dans ce fichier html. Nous présentons ci-dessous les liens entre les termes utilisés dans ce thésaurus suivi d'un exemple de hiérarchie entre ces termes du domaine de la construction durable :

[41] http://irc.web-p.cisti.nrc.ca/thesaurus/welcome.html

BT Broader term **GT General term** **WT Whole term** (generalization) (specialization)
NT Narrower term **PT Part term**
AT Associated term (these are at the same level) **RT Related term**
US Use (this) term (synonyms) **UF** Use (this) term instead
FT **French term**
SN Scope Notes (This is extra information and description)

6.1.2 Présentation du corpus de test

Notre corpus de test est constitué de documents textuels au format pdf, HTML et word appartenant du domaine de la construction et développement durable, en particulier le sous domaine relatif aux solutions pour l'économie d'énergie. Le contenu de ce corpus est récapitulé dans le tableau 6.1. Cette base contient 120 documents, répartis sur 9 thèmes et de taille globale 14 Méga Octets. Les ressources ont été téléchargées à partir des sites Web suivants déjà présentés dans la section 4.2.4 du quatrième chapitre : http://www.ademe.fr, http://www.cstb.fr, www.rncan.gc.ca, http://www.ec.gc.ca, http://www.avenir-energie.com.

Thème	Les Topics les plus fréquents associés aux thèmes	Nombre de documents	Types de documents	Taille
Chauffage	Equipements de chauffage Types de chauffage Organisme de certification Chauffage résidentiel Comparatif modes de chauffage Devis de chauffage Sécurité Coût d'entretien	19	HTML pdf	4,45 Mo
Chauffage écologique	Moyens de chauffage écologique Mode de fonctionnement Consommation d'énergie Energie renouvelable Rapport et publications Coût d'installation	25	Doc Pdf html	6,31 Mo
Chauffage au bois	Impact sur l'environnement Pollution Effets sur la santé Appareils certifiés Normes – réglementation Chaudière bois Cheminée Label qualité Poêle	20	pdf	6,17 Mo

	Citerne à propane			
Chauffage à l'électricité	Installation de chauffage électrique Radiateur Plancher chauffant Pompe à chaleur Types de chaudières électriques Systèmes de chauffage mixte bois- electricité	8	Html pdf	3,57 Mo
Chauffage solaire	Systèmes de production d'énergie solaire Fonctionnement du chauffage solaire Capteur solaire Performances et certification Budget pour installation du chauffage solaire	13	Pdf Html	9,29 Mo
Chauffage par géothermie	Guide d'installation Principe de fonctionnement d'une pompe à chaleur géothermique Coût d'installation Rendement de la pompe à chaleur géothermique	15	Html pdf	4,99 Mo
Chauffage au gaz naturel	Pompe à chaleur Conseils d'installation Types de chaudière à gaz Liste des modèles de chaudières au gaz Utilisation conseillée	11	Doc Pdf html	3,19 Mo
Chauffage au fuel	Caractéristiques du fuel Utilisation du fuel Entretien de chauffage au fuel Chaudière à condensation fuel	9	pdf	4,29 Mo

Tableau 6.1 Description du corpus de test

6.2 Présentation de la plateforme

Nous avons développé une plateforme de recherche intelligente fondée sur le modèle des Topic Maps. Cette plateforme permet de réaliser les fonctionnalités suivantes :

- **La création semi-automatique de la Topic Map** à partir des documents textuels multilingues, du thésaurus du domaine, des deux ontologies générales WordNet et WOLF et des scénarios d'usage. L'utilisateur commence par définir les paramètres de la Topic Map, à savoir le corpus, qui doit être un ensemble de

documents à priori textes, word, pdf ou html, et le thésaurus qui, dans notre cas doit être un fichier csv ;

- **Le stockage et la sauvegarde de la Topic Map générée** sous le format XTM par défaut ou bien sous un autre format par exemple RDF ou OWL ;
- **Le paramétrage du module de création de la Topic Map** : l'utilisateur a la possibilité de choisir le format des sources en entrée, il peut choisir un thésaurus, une ontologie du domaine ou un dictionnaire. Il peut également choisir différents types de documents grâce à une fonction qui permet de convertir tous les types de documents au format texte. Le paramétrage de la Topic Map inclut aussi le choix du format de sauvegarde de la Topic Map (XTM, RDF, OWL) et le choix du système de visualisation de la Topic Map (par exemple affichage selon la langue choisie par l'utilisateur, affichage selon le profil de l'utilisateur) ;
- **La visualisation de la Topic Map** pour la recherche et la navigation. L'interrogation de la Topic Map peut se faire soit par requête soit par navigation.

La figure 6.1 présente le diagramme de cas d'utilisation illustrant les fonctionnalités de notre plateforme.

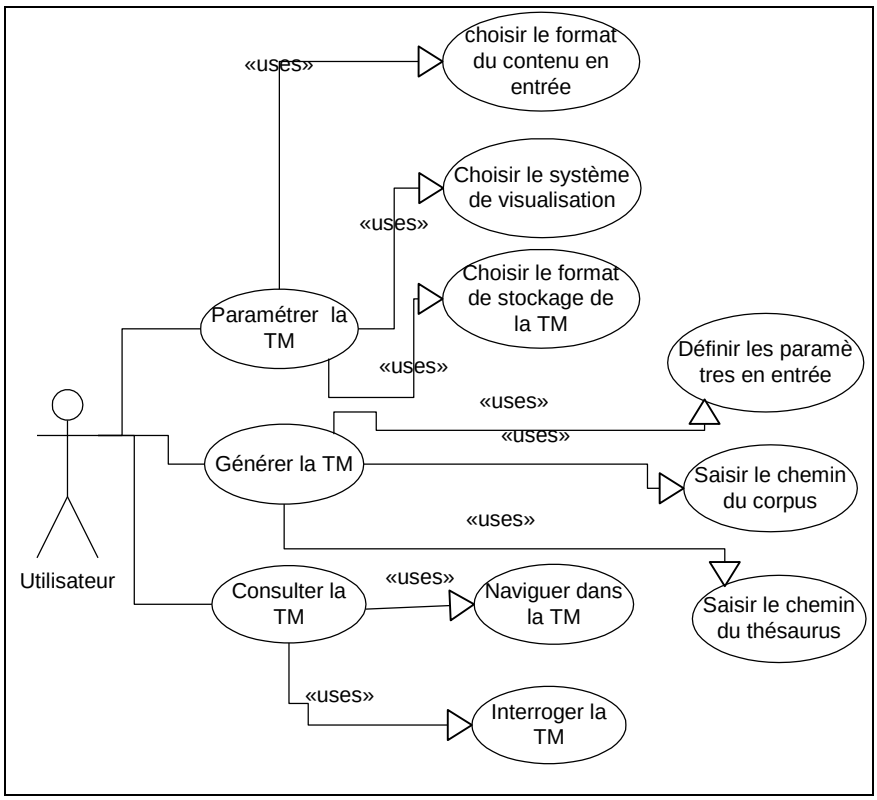


Figure 6.1 Diagramme des cas d'utilisation

6.3 Architecture générale

L'architecture générale de notre plateforme est basée sur trois modules déployés sur une architecture n-tier. Ces modules sont indépendants mais utilisent les mêmes données (thésaurus, documents, ontologies générales, scénarios d'usage …).

Le module Paramétrage de la Topic Map : TM_PARAMETRING

Ce module prend comme entrée les choix de l'utilisateur en ce qui concerne le corpus et le thésaurus ainsi que le choix du format de sauvegarde de la Topic Map et le système de visualisation. Ce module transmet les chemins vers le corpus et le thésaurus au module construction et la Topic Map au format choisi et le système de visualisation au module consultation.

Le module Construction de la Topic Map : TM_GENERATOR

Après avoir préparé le corpus déjà choisi par l'utilisateur en le transformant en format texte, ce module assure la connexion avec le serveur R afin de réaliser l'indexation de ses documents. Une segmentation éventuelle peut s'avérer nécessaire dans le cas de documents de grande taille, cette tâche est aussi assurée par ce module. En se basant sur le résultat de l'indexation, avec le package LSA de R, sous forme de matrice termes/ documents ou segments, ce module assure la construction automatique de la Topic Map à l'aide du thésaurus donné en paramètre.

Le module consultation de la Topic Map : TM_CONSULT

Deux axes constituent ce module : requête et navigation. Selon le choix de l'utilisateur, le fichier XTM généré est soit transmis à ce module afin de répondre aux requêtes des utilisateurs écrites en langage Tolog (langage de requêtes des Topic Maps proposé par Ontopia), soit transmis au module paramétrage. En effet, l'utilisateur, après avoir choisi le système de visualisation, il pourra naviguer dans la Topic Map et accéder aux documents qui l'intéressent.

La figure 6.2 présente l'architecture générale de notre plateforme de recherche intelligente fondée sur le modèle des Topic Maps.

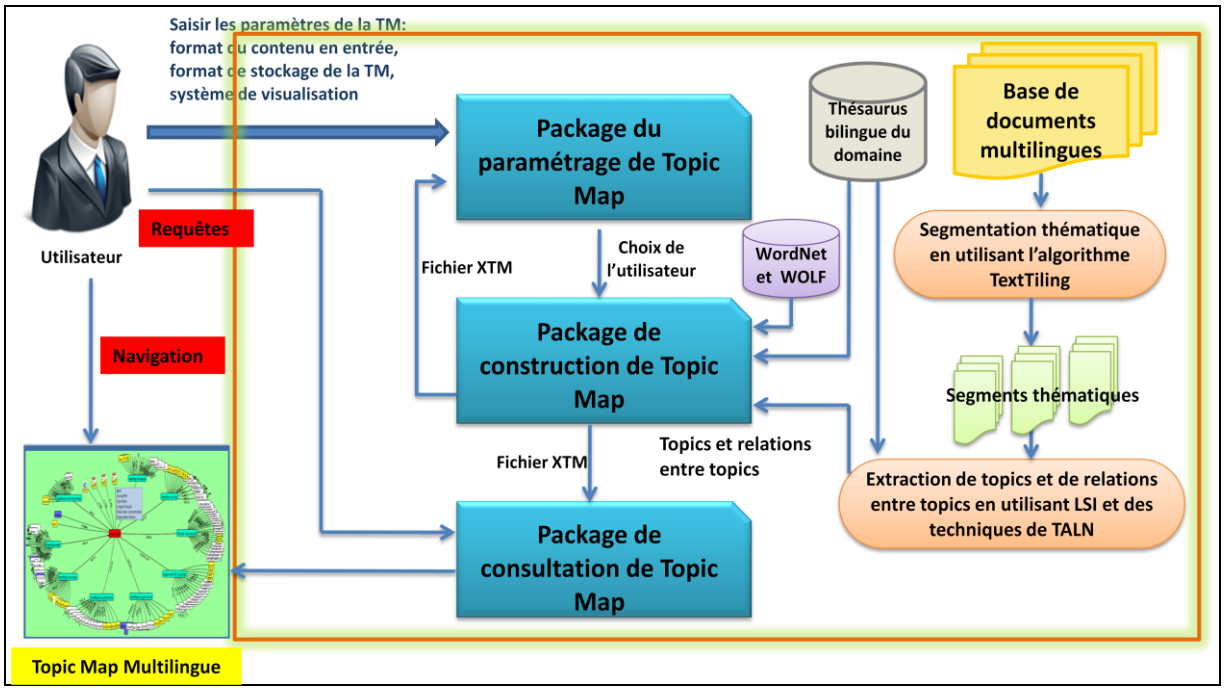


Figure 6.2 Architecture globale de la plateforme proposée

6.4 Réalisation et expérimentations

Nous avons implémenté notre plateforme en prenant en compte quatre principes : (i) la modularité - la plateforme est composée de quatre modules un module pour chaque fonctionnalité - (ii) l'extensibilité, (iii) le guidage pour supporter l'utilisateur qu'il soit expert ou non et (iv) la portabilité. Ce dernier principe est essentiellement lié au paramétrage de la Topic Map et plus précisément de celui des formats de sauvegarde. L'utilisateur a la possibilité de sauvegarder la Topic Map sous d'autres formats autre que XTM tels que RDF ou OWL. Il aura alors la possibilité d'utiliser et de visualiser la Topic Map sur d'autres plateformes et outils de visualisation.

Notre plateforme est implémentée en Java JDK 1.6 en utilisant l'environnement NetBeans IDE Dev 6.9 et, les bibliothèques Rengine, RserveEngine de LSI et l'API TM4J dédiée aux Topic Maps.

6.4.1 Environnement matériel et logiciel

L'application a été réalisée sous windows XP et en utilisant :

- Le langage JAVA Java JDk 1.6 pour la construction de la Topic Map, avec l'environnement de développement NetBeans IDE Dev 6.9 ;
- Le langage R pour l'indexation sémantique des documents, version 2.10.1 et les bibliothèques Rengine, RserveEngine du package LSI ;

- Les bibliothèques TM4J qui contient les classes de bases des Topic Maps ;
- Le standard XTM pour la représentation des Topic Maps, version 1.0.4.1.2.1 ;
- WordNet Online Search - 3.0 (http://wordnet.princeton.edu/).

Présentation du langage R

R[42] est un système d'analyse statistique et graphique créé par Ross Ihaka et Robert Gentleman. R est à la fois un logiciel et un langage qualifié de dialecte du langage S créé par *AT&T Bell Laboratories*. S est disponible sous la forme du logiciel S-PLUS commercialisé par la compagnie Insightful. Il y a des différences importantes dans la conception de R et celle de S. R est distribuée librement sous les termes de la GNU (*General Public Licence).* Son développement et sa distribution sont assurés par plusieurs statisticiens rassemblés dans le R Development Core Team.

De R vers JAVA

Pour la connexion R-Java, deux solutions sont possibles : la première consiste à utiliser l'API RClient et implémenter les classes nécessaires pour exécuter les scripts R à partir de l'application Java, la deuxième se base sur une architecture client/serveur grace à la bibliothèque REngine qui permet d'exécuter les scripts R sans avoir besoin d'implémenter les classes nécessaires. Nous avons adopté la deuxième solution, cette dernière nécessite la réalisation des étapes suivantes :

- Installer le serveur Rserve dans RGUI ;
- Installer l'API REngine et l'importer dans l'application Java ;
- Ouvrir une connexion au serveur et lancer le serveur ;
- Ecrire le script R directement dans le programme Java et l'exécuter, cette exécution permet de manipuler les objets R directement dans les variables et objets Java ;
- A la fin fermer la connexion.

6.4.2 Implémentation des modules

Nous présentons respectivement dans les figures 6.3 et 6.4, le diagramme de classe global et le diagramme de packages de notre application.

[42] http://www.r-project.org/

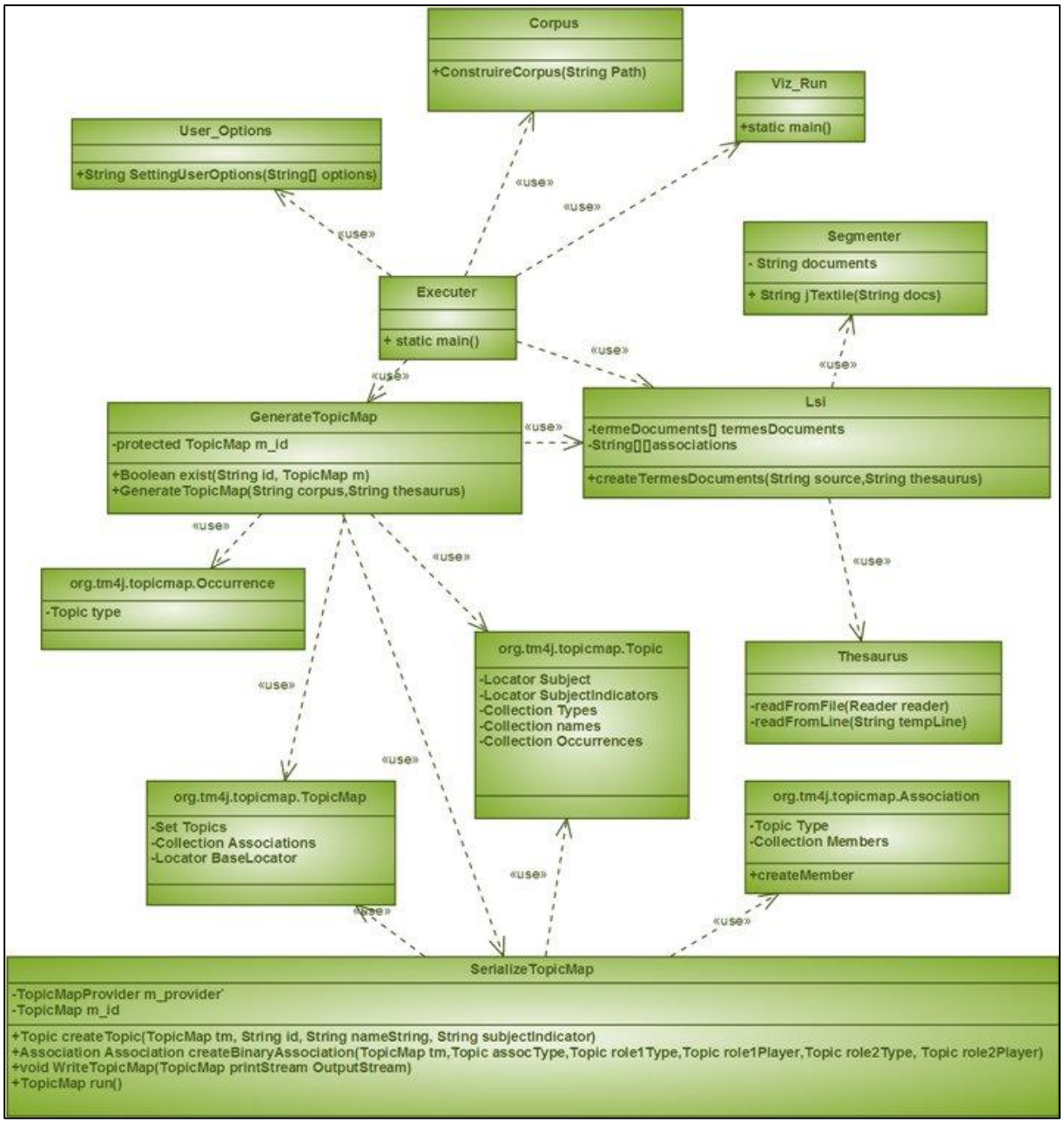


Figure 6.3 Diagramme de classes global

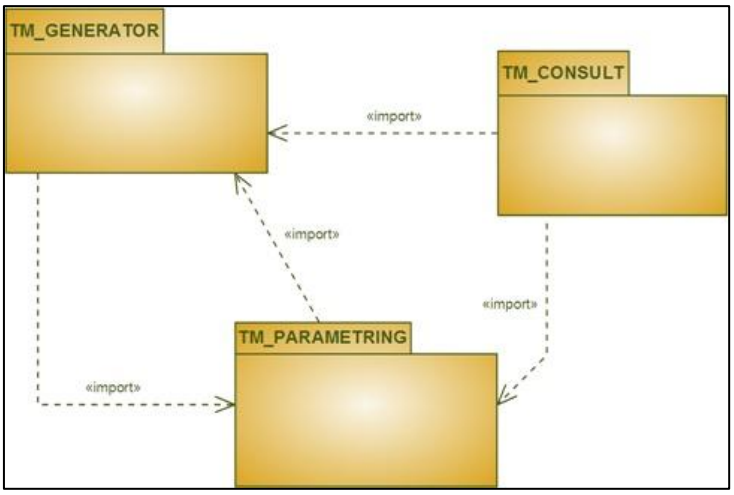


Figure 6.4 Diagramme de packages

Nous décrivons dans ce qui suit les différents packages implémentés dans notre plateforme ainsi que les classes de l'API TM4J que nous avons utilisées dans le cadre de notre travail.

6.4.2.1 Les classes de l'API TM4J

TM4J est une API Java open source pour la manipulation de Topic Maps. Parmi les interfaces de cette API auxquelles nous avons eu recours pour l'implémentation de notre application, les classes du package *org.tm4j.topicmap* :

- La classe *org.tm4j.topicmap.TopicMap* implémente les principales caractéristiques permettant de définir les concepts de modèle des Topic Maps et les différentes façons de les construire et les manipuler ;
- La classe *org.tm4j.topicmap.Topic* décrit un Topic dans une Topic Map ;
- La classe *org.tm4j.topicmap.Association* implémente une association entre deux Topics de la Topic Map ;
- La classe *org.tm4j.topicmap.Occurrence* définit une occurrence d'un Topic.

6.4.2.2 Le package TM_GENERATOR

Ce package - illustré par la figure 6.5 - renferme la majorité des classes de l'application puisqu'il permet d'implémenter la fonctionnalité principale du système, à savoir la génération automatique de la Topic Map à partir des paramètres d'entrée fournis par le package de paramétrage. Ses classes sont les suivantes :

- **Corpus** : A l'aide des fonctions *PdfTextConverter*, *WordTextConverter* et *HtmlTextConverter* importées du package TM_PARAMETRING, cette classe

assure la transformation du corpus fourni par l'utilisateur, sous un format très souvent hétérogène, vers le format texte brut afin de le préparer pour la segmentation et l'indexation ;

- **Thesaurus** : Le package de construction de la Topic Map prend en entrée le thésaurus CTCS du domaine pour enrichir la Topic Map avec les associations entre les Topics ;
- **Segmenter** : Cette classe permet de segmenter les documents du corpus afin d'extraire les segments thématiquement homogènes à partir de chaque document, ces segments représentent le thématiques abordées dans un même document. Nous avons utilisé l'algorithme TextTiling pour la segmentation thématique ;
- **LSI** : A ce stade, les documents du corpus sont prétraités et segmentés, cette classe permet l'indexation de l'ensemble des contextes ainsi crées (documents et segments). L'API REngine est implémentée selon une architecture client/serveur, nous avons donc implémenté la classe LSI comme un client au serveur R, elle fournit un script R au serveur permettant de réaliser l'indexation et par la suite elle reçoit le résultat sous forme de matrice termes/contextes ;
- **GenerateTopicMap** : Cette classe constitue le cœur de l'application, c'est elle qui assure la fonctionnalité la plus importante, à savoir la génération automatique de la Topic Map, en d'autres termes, la création des Topics, des associations et des occurrences ;
- **SerializeTopicMap** : Servant de base pour la classe *GenerateTopicMap*, cette classe sert à initialiser la Topic Map c'est-à-dire la construction de la première Topic Map 0 qui contient un seul Topic représenté par le concept clé du domaine « construction durable » ;
- **Execute** : cette classe permet le déclenchement des différents algorithmes et l'exécution du main principal.

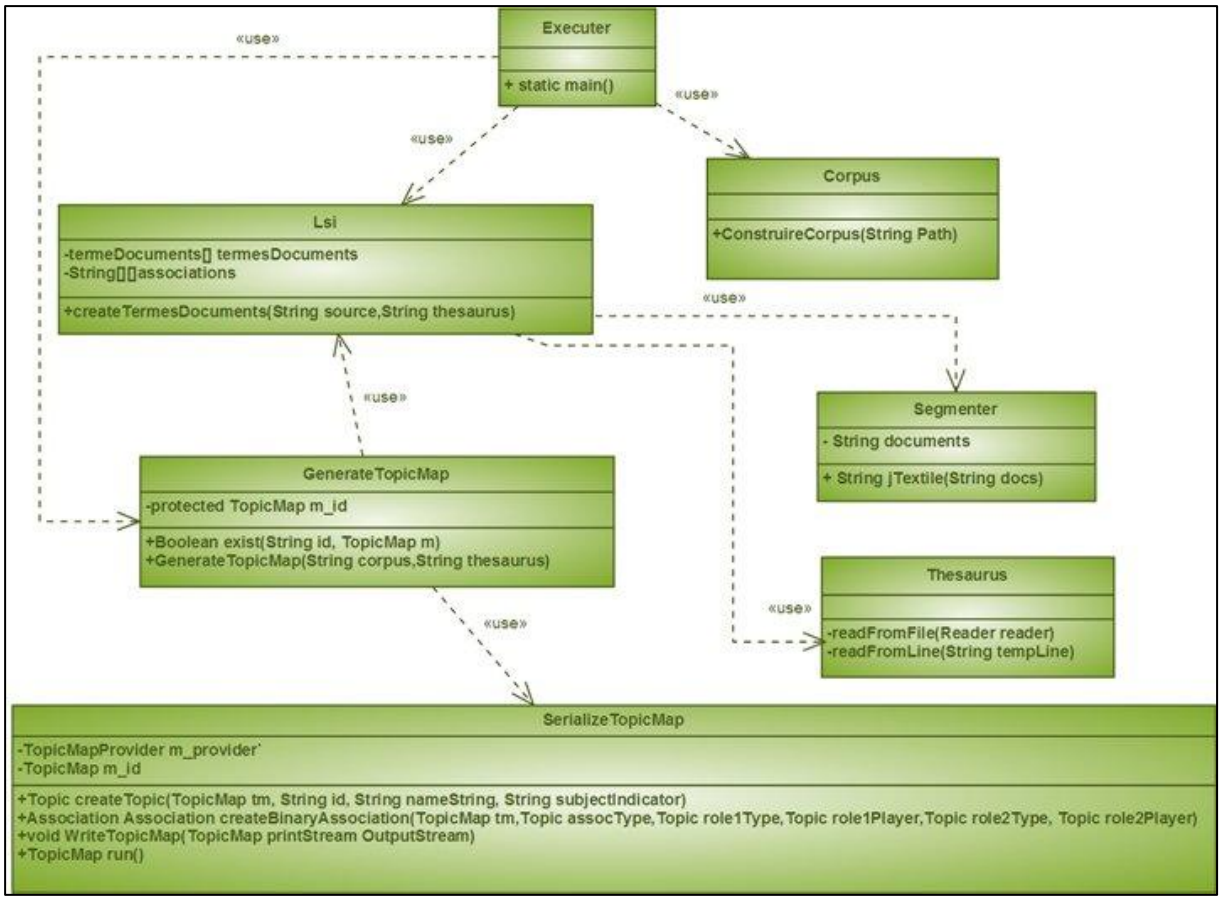


Figure 6.5 Diagramme de classes du package TM_GENERATOR

6.4.2.3 Le package TM_ PARAMETRING

Comme le montre la figure 6.6, ce package renferme la classe *User_Options* qui est une interface permettant à l'utilisateur de saisir les différents paramètres d'entrées et ses choix concernant le format de sauvegarde de la Topic Map et le système de visualisation. Cette classe fait appel aux autres classes du package pour garantir cette fonctionnalité. Par exemple, la classe *SaveFormat* et ses deux sous classes (*SaveFormat_RDF* et *SaveFormat_OWL) permettent* la prise en charge des formats RDF et OWL pour la transformation de la Topic Map selon ces deux formats et éventuellement d'autres modèles de représentation de connaissances. Concernant le système de visualisation, l'utilisateur pourrait par exemple choisir d'afficher la Topic Map en une seule langue (telle que l'anglais).

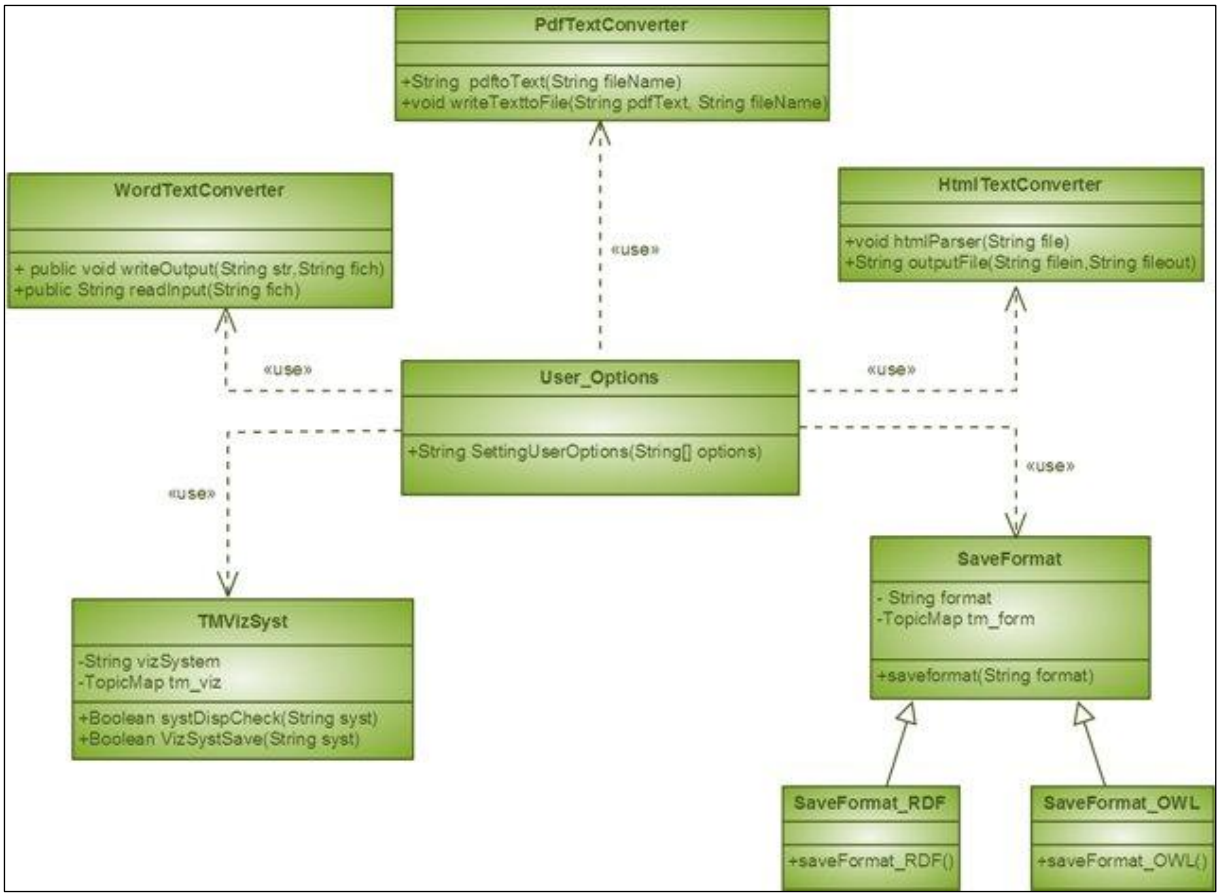


Figure 6.6 Diagramme du package TM_PARAMETRING

6.4.2.4 Le package TM_CONSULT

Selon les choix préalablement définis dans le module de paramétrage, une des deux classes *TMQuery* et *Visualizer* est appelée par la classe *Viz_Run* (figure 6.7). La première classe *TMQuery* permet par l'intermédiaire du langage de requête Tolog de répondre aux requêtes de l'utilisateur. La deuxième classe, *Visualizer*, permet d'afficher la Topic Map sous forme de carte sémantique contenant les Topics, les associations et les liens occurrences qui permettent d'accéder aux ressources reliées aux Topics. L'utilisateur pourra, à travers ce mode d'interrogation, naviguer à travers les liens de la Topic Map et consulter les documents indexés par les Topics.

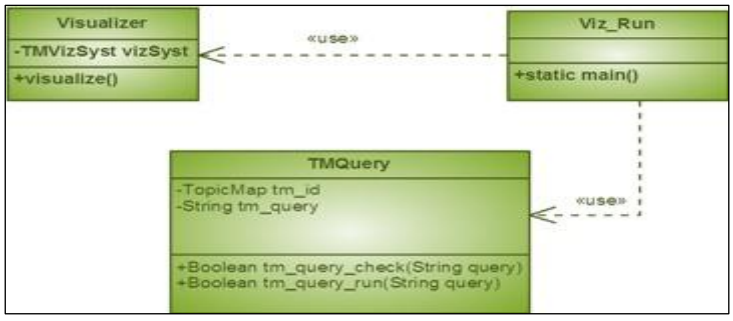


Figure 6.7 Diagramme du package TM_CONSULT

6.4.2.5 Diagrammes de séquences

Nous présentons dans la figure 6.8 un scénario qui décrit les interactions et les messages échangés entre les objets lors de la phase de construction de la Topic Map. L'objet « Rserve » dans le diagramme représente le serveur R qui assure la connexion R-Java.

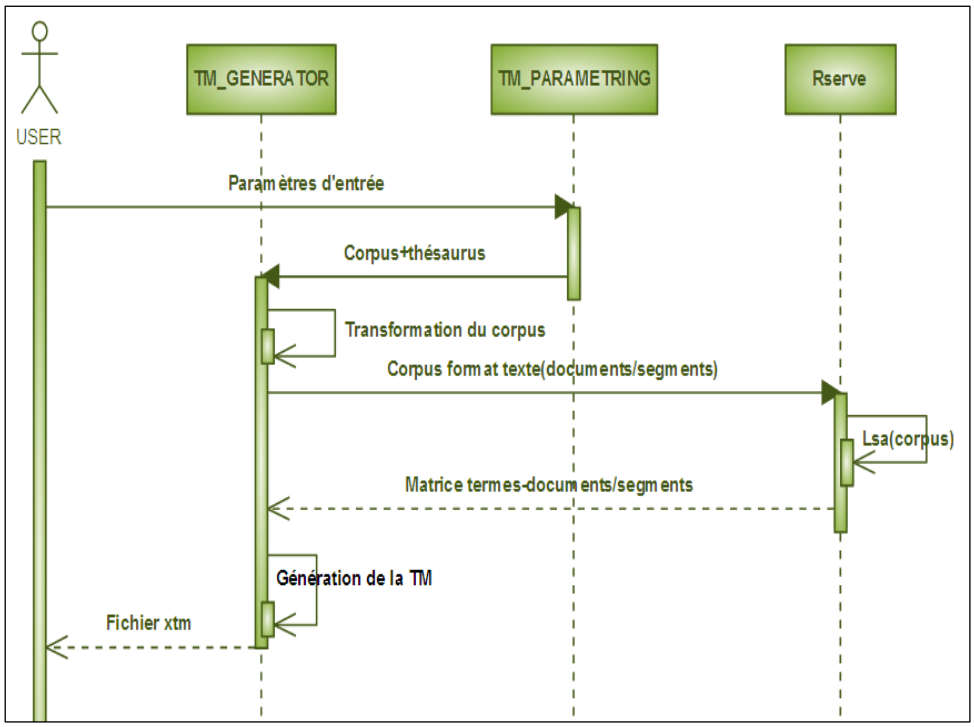


Figure 6.8 Diagramme de séquence du cas d'utilisation « construire la Topic Map »

Le diagramme de séquences de la figure 6.9 montre les interactions entre les différents composants dans un scénario pour la consultation de la Topic Map.

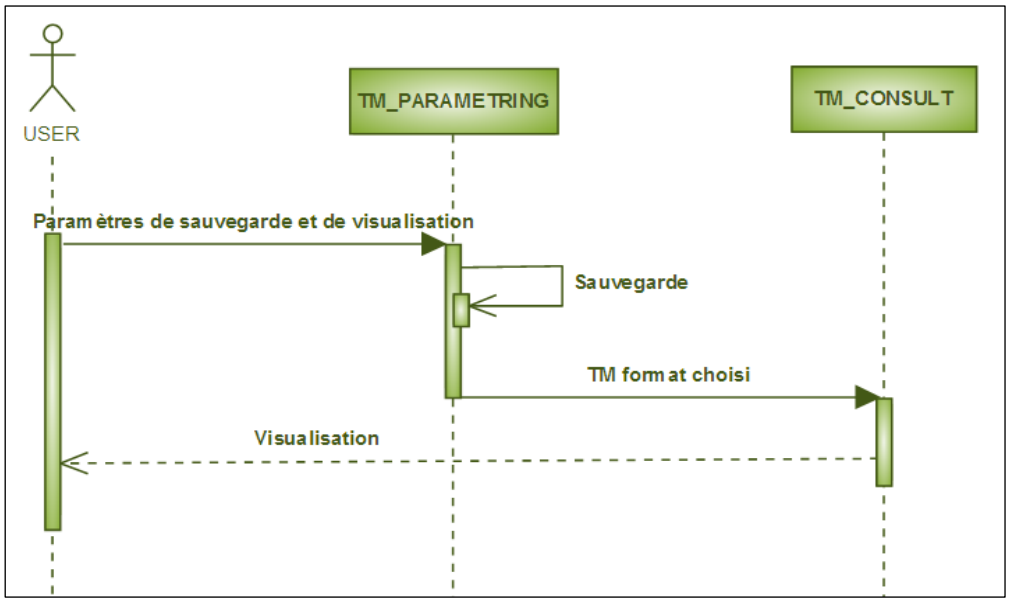
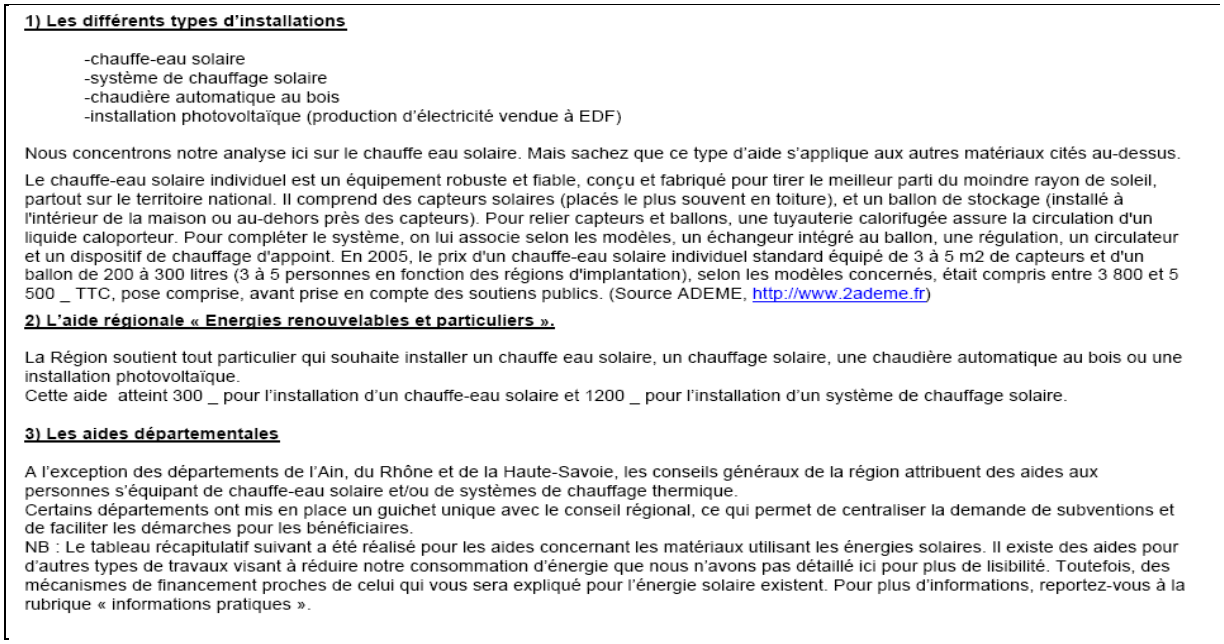


Figure 6.9 Diagramme de séquences du cas d'utilisation « consulter la Topic Map »

6.4.3 Expérimentations et résultats

Pour tester notre plateforme, une Topic Map a été crée à partir des documents, du thésaurus CTCS, des ontologies WordNet et WOLF et de l'ensemble des scénarios d'usage que nous avons élaborés. La figure 6.10 présente un extrait d'un document du corpus de test.



1) Les différents types d'installations

-chauffe-eau solaire
-système de chauffage solaire
-chaudière automatique au bois
-installation photovoltaïque (production d'électricité vendue à EDF)

Nous concentrons notre analyse ici sur le chauffe eau solaire. Mais sachez que ce type d'aide s'applique aux autres matériaux cités au-dessus.

Le chauffe-eau solaire individuel est un équipement robuste et fiable, conçu et fabriqué pour tirer le meilleur parti du moindre rayon de soleil, partout sur le territoire national. Il comprend des capteurs solaires (placés le plus souvent en toiture), et un ballon de stockage (installé à l'intérieur de la maison ou au-dehors près des capteurs). Pour relier capteurs et ballons, une tuyauterie calorifugée assure la circulation d'un liquide caloporteur. Pour compléter le système, on lui associe selon les modèles, un échangeur intégré au ballon, une régulation, un circulateur et un dispositif de chauffage d'appoint. En 2005, le prix d'un chauffe-eau solaire individuel standard équipé de 3 à 5 m2 de capteurs et d'un ballon de 200 à 300 litres (3 à 5 personnes en fonction des régions d'implantation), selon les modèles concernés, était compris entre 3 800 et 5 500 _ TTC, pose comprise, avant prise en compte des soutiens publics. (Source ADEME, http://www.2ademe.fr)

2) L'aide régionale « Energies renouvelables et particuliers ».

La Région soutient tout particulier qui souhaite installer un chauffe eau solaire, un chauffage solaire, une chaudière automatique au bois ou une installation photovoltaïque.
Cette aide atteint 300 _ pour l'installation d'un chauffe-eau solaire et 1200 _ pour l'installation d'un système de chauffage solaire.

3) Les aides départementales

A l'exception des départements de l'Ain, du Rhône et de la Haute-Savoie, les conseils généraux de la région attribuent des aides aux personnes s'équipant de chauffe-eau solaire et/ou de systèmes de chauffage thermique.
Certains départements ont mis en place un guichet unique avec le conseil régional, ce qui permet de centraliser la demande de subventions et de faciliter les démarches pour les bénéficiaires.
NB : Le tableau récapitulatif suivant a été réalisé pour les aides concernant les matériaux utilisant les énergies solaires. Il existe des aides pour d'autres types de travaux visant à réduire notre consommation d'énergie que nous n'avons pas détaillé ici pour plus de lisibilité. Toutefois, des mécanismes de financement proches de celui qui vous sera expliqué pour l'énergie solaire existent. Pour plus d'informations, reportez-vous à la rubrique « informations pratiques ».

Figure 6.10 Extrait d'un document écrit en français de notre corpus

6.4.3.1 Segmentation et indexation de documents

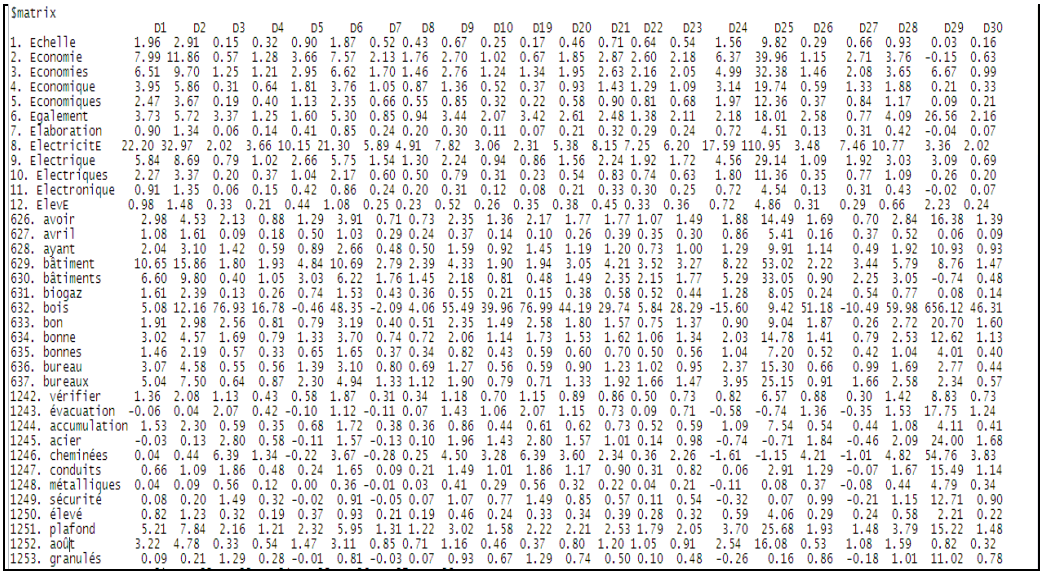
A l'aide du package LSA de R, qui prend en entrée un ensemble de contextes (documents ou segments) et donne une matrice termes/contextes pondérée avec les *tf×idf* des termes dans les contextes, nous avons écrit un script R qui permet de construire la matrice termes-contextes avec la fonction de normalisation *gwidf* qui permet d'avoir des valeurs proches des valeurs *tf×idf* normalisées. Le script est montré par la figure 6.11.

```
>mymatrix=textmatrix('corpus',minDocFreq=2,minGlobFreq=4,removeNumbers=TRUE)
>myspace=lsa(mymatrix,dims=2)

>tfidf = gw_idf(as.textmatrix(myspace))*as.textmatrix(myspace)
>space=lsa(tfidf)

>X=as.textmatrix(space)
>term=rownames(X)
```

Figure 6.11 Script permettant la génération de la matrice termes/contextes à partir des documents segmentés

Le résultat de l'indexation par LSA est donné sous forme de matrice dont les lignes sont les termes, les colonnes sont les contextes (document ou segment) et les cellules représentent

les densités (pour notre cas la mesure *tf*×*idf*) des termes dans les contextes. La figure 6.12 montre un extrait de la matrice générée et affichée sur l'interface R.



```
$matrix
                      D1    D2   D3   D4    D5    D6    D7   D8   D9  D10  D19  D20  D21 D22  D23   D24   D25  D26   D27   D28   D29  D30
1. Echelle          1.96  2.91 0.15 0.32  0.90  1.87  0.52 0.43 0.67 0.25 0.17 0.46 0.71 0.64 0.54  1.56  9.82 0.29  0.66  0.93  0.03 0.16
2. Economie         7.99 11.86 0.57 1.28  3.66  7.57  2.13 1.76 2.70 1.02 0.67 1.85 2.87 2.60 2.18  6.37 39.96 1.15  2.71  3.76 -0.15 0.63
3. Economies        6.51  9.70 1.25 1.21  2.95  6.62  1.70 1.46 2.76 1.24 1.34 1.95 2.63 2.16 2.05  4.99 32.38 1.46  2.08  3.65  6.67 0.99
4. Economique       3.95  5.86 0.31 0.64  1.81  3.76  1.05 0.87 1.36 0.52 0.37 0.93 1.43 1.29 1.09  3.14 19.74 0.59  1.33  1.88  0.21 0.33
5. Economiques      2.47  3.67 0.19 0.40  1.13  2.35  0.66 0.55 0.85 0.32 0.22 0.58 0.90 0.81 0.68  1.97 12.36 0.37  0.84  1.17  0.09 0.21
6. Egalement        3.73  5.72 3.37 1.25  1.60  5.30  0.85 0.94 3.44 2.07 3.42 2.61 2.48 1.38 2.11  2.18 18.01 2.58  0.77  4.09 26.56 2.16
7. Elaboration      0.90  1.34 0.06 0.14  0.41  0.85  0.24 0.20 0.30 0.11 0.07 0.21 0.32 0.29 0.24  0.72  4.51 0.13  0.31  0.42 -0.04 0.07
8. ElectricitE     22.20 32.97 2.02 3.66 10.15 21.30  5.89 4.91 7.82 3.06 2.31 5.38 8.15 7.25 6.20 17.59 110.95 3.48  7.46 10.77  3.36 2.02
9. Electrique       5.84  8.69 0.79 1.02  2.66  5.75  1.54 1.30 2.24 0.94 0.86 1.56 2.24 1.92 1.72  4.56 29.14 1.09  1.92  3.03  3.09 0.69
10. Electriques     2.27  3.37 0.20 0.37  1.04  2.17  0.60 0.50 0.79 0.31 0.23 0.54 0.83 0.74 0.63  1.80 11.36 0.35  0.77  1.09  0.26 0.20
11. Electronique    0.91  1.35 0.06 0.15  0.42  0.86  0.24 0.20 0.31 0.12 0.08 0.21 0.33 0.30 0.25  0.72  4.54 0.13  0.31  0.43 -0.02 0.07
12. ElevE           0.98  1.48 0.33 0.21  0.44  1.08  0.25 0.23 0.52 0.26 0.35 0.38 0.45 0.33 0.36  0.72  4.86 0.31  0.29  0.66  2.23 0.24
626. avoir          2.98  4.53 2.13 0.88  1.29  3.91  0.71 0.73 2.35 1.36 2.17 1.77 1.77 1.07 1.49  1.88 14.49 1.69  0.70  2.84 16.38 1.39
627. avril          1.08  1.61 0.09 0.18  0.50  1.03  0.29 0.24 0.37 0.14 0.10 0.26 0.39 0.35 0.30  0.86  5.41 0.16  0.37  0.52  0.06 0.09
628. ayant          2.04  3.10 1.42 0.59  0.89  2.66  0.48 0.50 1.59 0.92 1.45 1.19 1.20 0.73 1.00  1.29  9.91 1.14  0.49  1.92 10.93 0.93
629. bâtiment      10.65 15.86 1.80 1.93  4.84 10.69  2.79 2.39 4.33 1.90 1.94 3.05 4.21 3.52 3.27  8.22 53.02 2.22  3.44  5.79  8.76 1.47
630. bâtiments      6.60  9.80 0.40 1.05  3.03  6.22  1.76 1.45 2.18 0.81 0.48 1.49 2.35 2.15 1.77  5.29 33.05 0.90  2.25  3.05 -0.74 0.48
631. biogaz         1.61  2.39 0.13 0.26  0.74  1.53  0.43 0.36 0.55 0.21 0.15 0.38 0.58 0.52 0.44  1.28  8.05 0.24  0.54  0.77  0.08 0.14
632. bois           5.08 12.16 76.93 16.78 -0.46 48.35 -2.09 4.06 55.49 39.96 76.99 44.19 29.74 5.84 28.29 -15.60  9.42 51.18 -10.49 59.98 656.12 46.31
633. bon            1.91  2.98 2.56 0.81  0.79  3.19  0.40 0.51 2.35 1.49 2.58 1.80 1.57 0.75 1.37  0.90  9.04 1.87  0.26  2.72 20.70 1.60
634. bonne          3.02  4.57 1.69 0.79  1.33  3.70  0.74 0.72 2.06 1.14 1.73 1.53 1.62 1.06 1.34  2.03 14.78 1.41  0.79  2.53 12.62 1.13
635. bonnes         1.46  2.19 0.57 0.33  0.65  1.65  0.37 0.34 0.82 0.43 0.59 0.60 0.70 0.50 0.56  1.04  7.20 0.52  0.42  1.04  4.01 0.40
636. bureau         3.07  4.58 0.55 0.56  1.39  3.10  0.80 0.69 1.27 0.56 0.59 0.90 1.23 1.02 0.95  2.37 15.30 0.66  0.99  1.69  2.77 0.44
637. bureaux        5.04  7.50 0.64 0.87  2.30  4.94  1.33 1.12 1.90 0.79 0.71 1.33 1.92 1.66 1.47  3.95 25.15 0.91  1.66  2.58  2.34 0.57
1242. vérifier      1.36  2.08 1.13 0.43  0.58  1.87  0.31 0.34 1.18 0.70 1.15 0.89 0.86 0.50 0.73  0.82  6.57 0.88  0.30  1.42  8.83 0.73
1243. évacuation   -0.06  0.04 2.07 0.42 -0.10  1.12 -0.11 0.07 1.43 1.06 2.07 1.15 0.73 0.09 0.71 -0.58 -0.74 1.36 -0.35  1.53 17.75 1.24
1244. accumulation  1.53  2.30 0.59 0.35  0.68  1.72  0.38 0.36 0.86 0.44 0.61 0.62 0.73 0.52 0.59  1.09  7.54 0.54  0.44  1.08  4.11 0.41
1245. acier        -0.03  0.13 2.80 0.58 -0.11  1.57 -0.13 0.10 1.96 1.43 2.80 1.57 1.01 0.14 0.98 -0.74 -0.71 1.84 -0.46  2.09 24.00 1.68
1246. cheminées     0.04  0.44 6.39 1.34 -0.22  3.67 -0.28 0.25 4.50 3.28 6.39 3.60 2.34 0.36 2.26 -1.61 -1.15 4.21 -1.01  4.82 54.76 3.83
1247. conduits      0.66  1.09 1.86 0.48  0.24  1.65  0.09 0.21 1.49 1.01 1.86 1.17 0.90 0.31 0.82  0.06  2.91 1.29 -0.07  1.67 15.49 1.14
1248. métalliques   0.04  0.09 0.56 0.12  0.00  0.36 -0.01 0.03 0.41 0.29 0.56 0.32 0.22 0.04 0.21 -0.11  0.08 0.37 -0.08  0.44  4.79 0.34
1249. sécurité      0.08  0.20 1.49 0.32 -0.02  0.91 -0.05 0.07 1.07 0.77 1.49 0.85 0.57 0.11 0.54 -0.32  0.07 0.99 -0.21  1.15 12.71 0.90
1250. élevé         0.82  1.23 0.32 0.19  0.37  0.93  0.21 0.19 0.46 0.24 0.33 0.34 0.39 0.28 0.32  0.59  4.06 0.29  0.24  0.58  2.21 0.22
1251. plafond       5.21  7.84 2.16 1.21  2.32  5.95  1.31 1.22 3.02 1.58 2.22 2.21 2.53 1.79 2.05  3.70 25.68 1.93  1.48  3.79 15.22 1.48
1252. août          3.22  4.78 0.33 0.54  1.47  3.11  0.85 0.71 1.16 0.46 0.37 0.80 1.20 1.05 0.91  2.54 16.08 0.53  1.08  1.59  0.82 0.32
1253. granulés      0.09  0.21 1.29 0.28 -0.01  0.81 -0.03 0.07 0.93 0.67 1.29 0.74 0.50 0.10 0.48 -0.26  0.16 0.86 -0.18  1.01 11.02 0.78
```

Figure 6.12 Extrait de la matrice termes/contextes de LSI

6.4.3.2 Génération et sauvegarde de la Topic Map

La Topic Map est stockée sous le format XTM. Le document XTM généré est conforme à la norme représentée par la DTD XTM 1.0 décrite par la figure 6.13. Nous avons choisi ce format car il est proposé par la norme. Il sera ainsi possible de visualiser la Topic Map automatiquement avec un outil dédié à la visualisation de Topic Map (ces derniers supportant ce type de format conforme à la DTD XTM 1.0 de la norme).

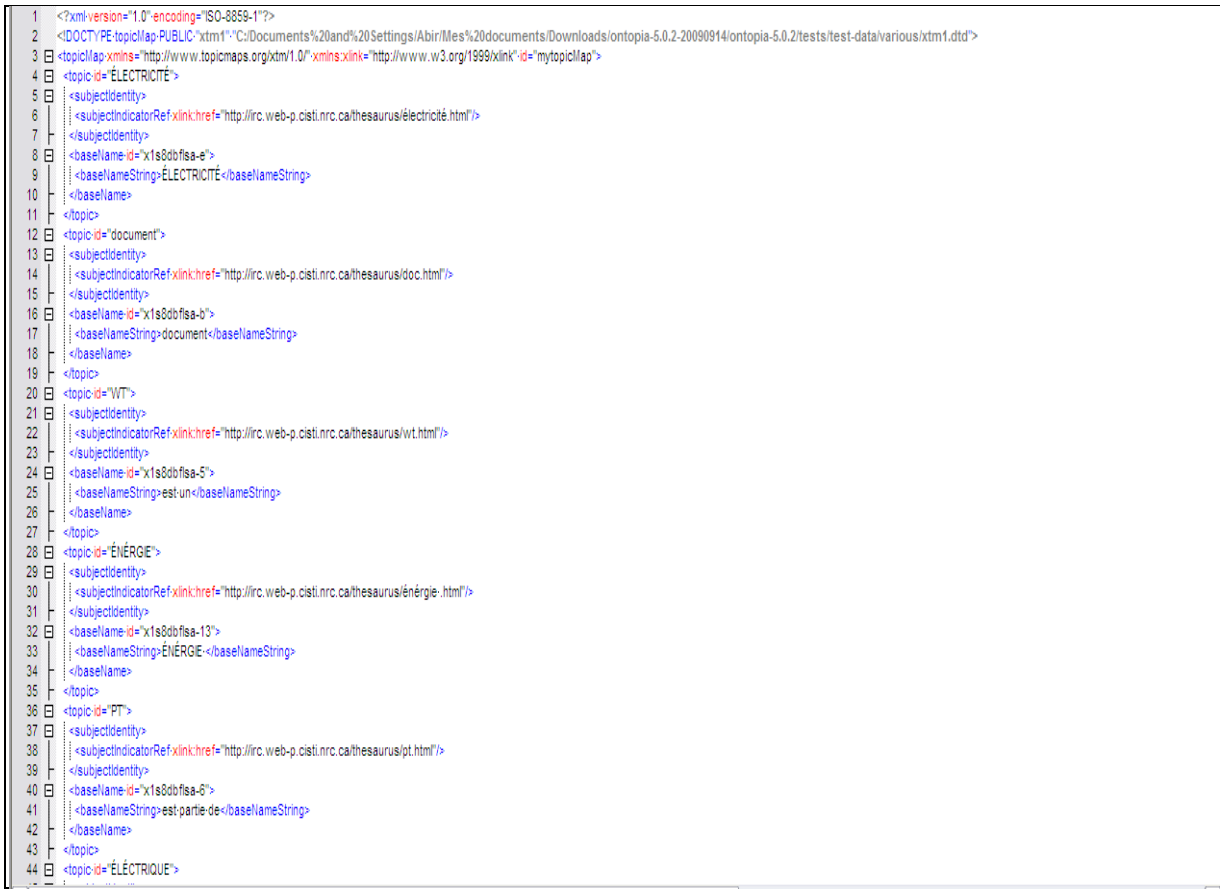
```
<!-- ..... -->
<!-- XML Topic Map DTD ..... -->
<!-- file: xtm1.dtd
-->
<!-- XML Topic Map (XTM) DTD, Version 1.0
    This is XTM, an XML interchange syntax for ISO 13250 Topic Maps.
    XML Topic Map (XTM)
     Copyright 2000-2001 TopicMaps.Org, All Rights Reserved.
      Editors:   Steve Pepper     <pepper@ontopia.net>
                 Graham Moore     <gdm@empolis.co.uk>
-->
<!-- Use this URI to identify the default XTM namespace:
      "http://www.topicMaps.org/xtm/1.0/"
  -->
<!-- topicMap: Topic Map document element ..... -->

<!ELEMENT topicMap
  ( topic | association | mergeMap )*
>
<!ATTLIST topicMap
  id           ID      #IMPLIED
  xmlns        CDATA   #FIXED 'http://www.topicmaps.org/xtm/1.0/'
  xmlns:xlink  CDATA   #FIXED 'http://www.w3.org/1999/xlink'
  xml:base     CDATA   #IMPLIED
>
<!-- topic: Topic element ..... -->
<!ELEMENT topic
  ( instanceOf*, subjectIdentity?, ( baseName | occurrence )* )
>
```

```
<!ATTLIST topic
  id          ID      #REQUIRED
>
<!-- instanceOf: Points To a Topic representing a class ..... -->
<!ELEMENT instanceOf  ( topicRef | subjectIndicatorRef ) >
<!ATTLIST instanceOf
  id          ID      #IMPLIED
>
<!-- subjectIdentity: Subject reified by Topic ..... -->
<!ELEMENT subjectIdentity
  ( resourceRef?, ( topicRef | subjectIndicatorRef )* )
>
<!ATTLIST subjectIdentity
  id          ID      #IMPLIED
>
<!-- topicRef: Reference to a Topic element ..... -->

<!ELEMENT topicRef  EMPTY >
<!ATTLIST topicRef
  id          ID      #IMPLIED
  xlink:type     NMTOKEN   #FIXED 'simple'
  xlink:href     CDATA     #REQUIRED
>
<!-- subjectIndicatorRef: Reference to a Subject Indicator ..... -->
<!ELEMENT subjectIndicatorRef  EMPTY >
<!ATTLIST subjectIndicatorRef
  id          ID      #IMPLIED
  xlink:type     NMTOKEN   #FIXED 'simple'
  xlink:href     CDATA     #REQUIRED
>
<!-- baseName: Base Name of a Topic ..... -->
<!ELEMENT baseName  ( scope?, baseNameString, variant* ) >
<!ATTLIST baseName
  id          ID      #IMPLIED
>
<!-- baseNameString: Base Name String container ..... -->
<!ELEMENT baseNameString  ( #PCDATA ) >
<!ATTLIST baseNameString
  id          ID      #IMPLIED
>
<!-- variant: Alternate forms of Base Name ..... -->
<!ELEMENT variant  ( parameters, variantName?, variant* ) >
<!ATTLIST variant
  id          ID      #IMPLIED
>
<!-- variantName: Container for Variant Name ..... -->
<!ELEMENT variantName  ( resourceRef | resourceData ) >
<!ATTLIST variantName
  id          ID      #IMPLIED
>
<!-- parameters: Processing context for Variant ..... -->
<!ELEMENT parameters  ( topicRef | subjectIndicatorRef )+ >
<!ATTLIST parameters
  id          ID      #IMPLIED
>
<!-- occurrence: Resources regarded as an Occurrence ..... -->
<!ELEMENT occurrence
  ( instanceOf?, scope?, ( resourceRef | resourceData ) )
>
<!ATTLIST occurrence
  id          ID      #IMPLIED
>
<!-- resourceRef: Reference to a Resource ..... -->
<!ELEMENT resourceRef  EMPTY >
<!ATTLIST resourceRef
  id          ID      #IMPLIED
  xlink:type     NMTOKEN   #FIXED 'simple'
  xlink:href     CDATA     #REQUIRED
>
<!-- resourceData: Container for Resource Data ..... -->
<!ELEMENT resourceData  ( #PCDATA ) >
<!ATTLIST resourceData
  id          ID      #IMPLIED
```

```
>
<!-- association: Topic Association  ..... -->

<!ELEMENT association
  ( instanceOf?, scope?, member+ )
>
<!ATTLIST association
  id          ID      #IMPLIED
>
<!-- member: Member in Topic Association ..... -->

<!ELEMENT member
  ( roleSpec?, ( topicRef | resourceRef | subjectIndicatorRef )* )
>
<!ATTLIST member
  id          ID      #IMPLIED
>
<!-- roleSpec: Points to a Topic serving as an Association Role .. -->
<!ELEMENT roleSpec  ( topicRef | subjectIndicatorRef ) >
<!ATTLIST roleSpec
  id          ID      #IMPLIED
>
<!-- scope: Reference to Topic(s) that comprise the Scope ..... -->

<!ELEMENT scope  ( topicRef  | resourceRef | subjectIndicatorRef )+ >
<!ATTLIST scope
  id          ID      #IMPLIED
>
<!-- mergeMap: Merge with another Topic Map ..... -->
<!ELEMENT mergeMap  ( topicRef | resourceRef | subjectIndicatorRef )* >
<!ATTLIST mergeMap
  id          ID      #IMPLIED
  xlink:type     NMTOKEN   #FIXED 'simple'
  xlink:href     CDATA     #REQUIRED
>
<!-- end of XML Topic Map (XTM) 1.0 DTD -->
```

Figure 6.13 DTD XTM 1.0

La figure 6.14 montre un extrait du fichier XTM généré automatiquement par la plateforme réalisée.



```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE topicMap PUBLIC "xtm1" "C:/Documents%20and%20Settings/Abir/Mes%20documents/Downloads/ontopia-5.0.2-20090914/ontopia-5.0.2/tests/test-data/various/xtm1.dtd">
<topicMap xmlns="http://www.topicmaps.org/xtm/1.0/" xmlns:xlink="http://www.w3.org/1999/xlink" id="mytopicMap">
 <topic id="ÉLECTRICITÉ">
  <subjectIdentity>
   <subjectIndicatorRef xlink:href="http://irc.web-p.cisti.nrc.ca/thesaurus/électricité.html"/>
  </subjectIdentity>
  <baseName id="x1s8dbflsa-e">
   <baseNameString>ÉLECTRICITÉ</baseNameString>
  </baseName>
 </topic>
 <topic id="document">
  <subjectIdentity>
   <subjectIndicatorRef xlink:href="http://irc.web-p.cisti.nrc.ca/thesaurus/doc.html"/>
  </subjectIdentity>
  <baseName id="x1s8dbflsa-b">
   <baseNameString>document</baseNameString>
  </baseName>
 </topic>
 <topic id="WT">
  <subjectIdentity>
   <subjectIndicatorRef xlink:href="http://irc.web-p.cisti.nrc.ca/thesaurus/wt.html"/>
  </subjectIdentity>
  <baseName id="x1s8dbflsa-5">
   <baseNameString>est un</baseNameString>
  </baseName>
 </topic>
 <topic id="ÉNÉRGIE">
  <subjectIdentity>
   <subjectIndicatorRef xlink:href="http://irc.web-p.cisti.nrc.ca/thesaurus/énérgie-.html"/>
  </subjectIdentity>
  <baseName id="x1s8dbflsa-13">
   <baseNameString>ÉNÉRGIE-</baseNameString>
  </baseName>
 </topic>
 <topic id="PT">
  <subjectIdentity>
   <subjectIndicatorRef xlink:href="http://irc.web-p.cisti.nrc.ca/thesaurus/pt.html"/>
  </subjectIdentity>
  <baseName id="x1s8dbflsa-6">
   <baseNameString>est partie de</baseNameString>
  </baseName>
 </topic>
 <topic id="ÉLÉCTRIQUE">
```

```
 <topic id="POMPE">
  <subjectIdentity>
   <subjectIndicatorRef xlink:href="http://irc.web-p.cisti.nrc.ca/thesaurus/pompe.html"/>
  </subjectIdentity>
  <baseName id="x1s8dbflsa-2a">
   <baseNameString>POMPE</baseNameString>
  </baseName>
 </topic>
 <topic id="chauffage">
  <instanceOf>
   <topicRef xlink:href="#CHAUFFAGE"/>
  </instanceOf>
  <baseName id="x1s8dbflsa-1c">
   <baseNameString>chauffage</baseNameString>
  </baseName>
  <occurrence id="occurrence1">
   <instanceOf>
    <topicRef xlink:href="#document"/>
   </instanceOf>
   <resourceData>i2141.txt</resourceData>
  </occurrence>
 </topic>
 <topic id="AÉRATION">
  <subjectIdentity>
   <subjectIndicatorRef xlink:href="http://irc.web-p.cisti.nrc.ca/thesaurus/aération.html"/>
  </subjectIdentity>
  <baseName id="x1s8dbflsa-5d">
   <baseNameString>AÉRATION</baseNameString>
  </baseName>
 </topic>
 <topic id="AT">
  <subjectIdentity>
   <subjectIndicatorRef xlink:href="http://irc.web-p.cisti.nrc.ca/thesaurus/at.html"/>
  </subjectIdentity>
  <baseName id="x1s8dbflsa-7">
   <baseNameString>est lié à</baseNameString>
  </baseName>
 </topic>
 <topic id="energy">
  <instanceOf>
   <topicRef xlink:href="#ENERGY"/>
  </instanceOf>
  <baseName id="x1s8dbflsa-16">
   <baseNameString>energy</baseNameString>
```

Figure 6.14 Extrait du fichier XTM généré avec notre plateforme

6.4.4 Recherche par requête

Concernant la recherche par requête, nous avons développé une interface en Java qui aide un utilisateur à formuler sa requête. Une requête est composée de deux dimensions : une partie spécifiant les annotations descriptives des documents cherchés, telles que type, format, ou date de création et une deuxième partie spécifiant la requête qui contient les Topics cherchés par l'utilisateur. Pour concrétiser ce service, nous avons choisi le langage Tolog pour l'interrogation de notre Topic Map. Le tableau 6.2 présente les classes et les principales méthodes des modules de traduction et d'exécution de requêtes.

Modules de traduction et d'exécution de requêtes	
Classes	***Méthodes Principales***
Traducteur	***getTraducteurTolog () :*** *permet de traduire la requête de l'utilisateur en tolog.*
	Pred () *: permet de traduire une nouvelle relation donnée par l'utilisateur en une règle d'inference en tolog.*
Exécution	***readTopicMap ()*** *: permet de charger la Topic Map.*
	getexecutionTolog () *: permet d'exécuter la requête de l'utilisateur traduite en tolog sur la Topic Map chargé.*

Tableau 6.2 Classes et méthodes principales correspondant aux modules traduction et exécution de requêtes

Ce service permet une recherche précise, en spécifiant les deux dimensions de la requête, l'utilisateur aura une liste des documents ou des segments de documents, classés selon leurs degrés de pertinence.

6.4.5 Visualisation de la Topic Map

Pour visualiser la Topic Map générée, nous avons testé plusieurs outils, parmi eux, il y a ceux qui ont été spécialement développés pour le modèle de Topic Map comme par exemple TM4J, Ontopia Vizdesktop[43] et TMNav et des outils de visualisation de graphes tels que Treebolic[44], Hypergraph[45]. Nous avons également étudié la possibilité d'intégrer des API pour l'affichage de la Topic Map telles que l'API Touchgraph[46].

6.4.5.1 Etat de l'art sur les techniques et les outils de visualisation

De nombreux travaux de recherche se sont intéressés à la représentation et la visualisation de données de grande taille. Legrand [Legrand et Soto, 2006] dans ses travaux

[43]http://ontopia.topicmapslab.de/omnigator/docs/vizigator/userguide.html
[44]http ://treebolic.sourceforge.net/
[45]http://hypergraph.sourceforge.net
[46] http://www.touchgraph.com/navigator.html

de recherche propose une étude comparative des techniques de visualisation proposées pour l'exploration d'ensembles de données multidimensionnelles de grande taille. Des interfaces visuelles intuitives peuvent réduire de manière significative la charge cognitive de l'utilisateur lorsqu'il travaille avec des systèmes complexes contenant une grande quantité de données. La visualisation est une technique prometteuse, à la fois pour améliorer la perception de la structure dans de grands espaces d'informations et pour faciliter la navigation. Elle permet également aux utilisateurs de mieux comprendre les données et leurs structures et extraire des connaissances plus efficacement.

Il existe différentes techniques pour la fouille visuelle de données : les graphes, les arbres et les cartes.

Graphes et arbres

Les graphes et les arbres sont bien adaptés à la représentation de la structure globale d'un ensemble de données multidimensionnelles. L'avantage des arbres par rapport aux autres types de graphes est que leur interprétation est plus facile, car les relations sont hiérarchiques. Cependant, les représentations arborescentes deviennent – comme les graphes de manière générale - confuses lorsque la taille du système augmente.

Diverses méthodes ont été proposées afin de pouvoir représenter visuellement un grand nombre d'informations. L'une des solutions consiste à déformer l'espace de représentation afin d'obtenir une représentation visuelle, dense, mais localement intelligible ; la vue en « œil de poisson » (Fisheye) [Sarkar et Brown, 1994], les arbres hyperboliques [Munzner, 1997] et le mur perspectif [Mackinlay et al. 1991] sont des exemples de visualisations utilisant cette technique.

Un outil de visualisation interactif, UNIVIT (*UNiversal Interactive VIsualization Tool*) [Legrand et Soto, 1999], a été développé au LIP6 pour permettre de visualiser tout type de système hiérarchique, sous la forme d'un arbre conique en trois dimensions. *UNIVIT* fournit une vue globale du système, ainsi que des vues détaillées affichées à la demande de l'utilisateur. Ces représentations mettent en œuvre des techniques apparentées à la réalité virtuelle, en particulier la tridimensionnalité, l'interactivité et l'utilisation de différents niveaux de détail. Des couleurs et des formes différentes sont utilisées pour symboliser les dimensions des nœuds et des arcs de l'arbre. Cependant le nombre de couleurs, formes, icônes et textures aisément différentiables est limité et le problème du passage à l'échelle se pose.

L'avantage des arbres et des graphes est de faire apparaître clairement la structure des données, et, en particulier, les relations entre ces données ; il est facile de distinguer les éléments isolés ou au contraire regroupés. Ceci est un point très positif pour faciliter la compréhension et la mémorisation des structures.

L'inconvénient est que ces bénéfices disparaissent lorsque le volume des données augmente, puisqu'il n'est plus possible de tout afficher. En particulier pour les graphes, lorsque le nombre d'arcs est trop important, la visualisation devient illisible. Les représentations arborescentes sont plus intuitives, mais elles manquent de généricité, puisqu'elles ne s'appliquent qu'à des structures hiérarchiques.

Les cartes

Les techniques de visualisation exploratoire ont pour objectif d'aider l'utilisateur à retrouver les informations qu'il cherche dans un grand nombre de documents. Dans le monde réel, nous utilisons des cartes pour satisfaire ces besoins. Dans cette section, nous étudions l'intérêt des visualisations exploratoires sous forme de carte.

NicheWorks [Wills, 1999] fournit une représentation schématique qui met en évidence la structure globale d'un système et permet par exemple de détecter des *outliers*, éléments marginaux qui sont isolés des autres. Par contre, une interprétation précise de cette carte est difficile tant le volume des données est grand.

Une solution pour résoudre ce problème consiste à réorganiser les données, ce que font les Tree-Maps [Johnson et Shneiderman, 1991], [Van Wijk et Van De Wetering, 1999]. Une Tree-Map permet de déceler au premier coup d'œil les éléments prépondérants du système, par l'utilisation de couleurs différentes pour chaque zone. Par contre, la compréhension détaillée devient difficile lorsque le volume des données augmente.

L'algorithme des cartes auto-organisatrices (*Self-Organizing Maps* ou *SOM*) [Kaski et al. 1998] de Kohonen permet de représenter les données sous forme de carte, en tenant compte de la similarité de ces données. Si l'on applique cet algorithme à un ensemble de documents, il les organise automatiquement sur un plan, de telle sorte que les documents ayant des points communs apparaissent proches les uns des autres. Il est ainsi possible de savoir quels éléments possèdent des similarités. Cependant, cette représentation n'indique pas de quelle manière les éléments sont reliés les uns avec les autres : il manque des informations sur la nature et donc la signification des relations entre les constituants du système pour comprendre pourquoi tel élément est situé à tel endroit, etc.

6.4.5.2 Interfaces de visualisation de la Topic Map

Nous avons testé plusieurs outils de visualisation déjà développés pour la visualisation de Topic Map comme par exemple TM4L Viewer, Ontopia Vizdesktop et TMNav qui est une partie du projet TM4J pour la visualisation de Topic Maps.

Ces outils prennent en entrée la Topic Map sous format XTM selon la DTD et permettent automatiquement d'afficher la Topic Map à condition que le fichier soit conforme à la DTD de la norme XTM ce qui est le cas du fichier qu'on a généré avec notre plateforme. Cependant, ces outils présentent quelques limites, par exemple TMNav et TM4L Viewer ne permettent pas la visualisation de la Topic Map entière dans un seul écran, dans TMNav l'utilisateur doit choisir parmi la liste des Topics, celui qu'il veut visualiser et l'outil affiche seulement le graphe ou les Topics reliés à ce Topic. Cependant il a l'avantage de l'affichage des occurrences liées aux concepts que l'utilisateur pourrait consulter directement à partir de la carte.

Contrairement à TMNav et TM4L Viewer, Ontopia Vizdesktop offre une visualisation de toute la Topic Map avec les liens sémantiques et les instances, mais les occurrences ne sont pas directement et facilement accessibles tel que pour TMNav, il faut pour les afficher, accéder à la carte d'identification de chaque concept dans le menu des propriétés.

Pour toutes ces raisons, nous avons testé d'autres outils de visualisation de grande quantité de données qui ne sont pas spécifiques aux Topic Maps, comme par exemple l'outil Hypergraph qui utilise une représentation en graphe de type noeud-lien et un paradigme de visualisation « fisheye » permettant de mettre en valeur le centre d'intérêt de navigation de l'utilisateur.

Nous avons également essayé d'intégrer l'API Touchgraph dans notre plateforme et **finalement nous avons choisi l'outil de visualisation Treebolic** qui une appliquette Java destinée à visualiser un ensemble hiérarchisé de données sous forme d'arbre hyperbolique. Cet arbre est dynamique : une animation amène un noeud visité au centre de l'affichage. Les noeuds peuvent contenir des liens hypertextes et mener le navigateur vers d'autres pages. Une appliquette Java destinée à visualiser un ensemble hiérarchisé de données sous forme d'arbre hyperbolique. Un arbre est affiché avec ses noeuds et ses arcs mais l'espace de visualisation subit une courbure en allouant davantage de place au centre, le noeud parent et les descendants immédiats apparaissent légèrement plus petits. Les grands parents et petits fils sont toujours visibles mais sont encore plus petits. Le noeud « père » sera placé au centre de l'écran par le biais d'une animation. Les noeuds fils s'afficheront ensuite autour. À mesure que

l'on s'éloigne du foyer, les noeuds se voient allouer moins d'espace et disparaissent ainsi virtuellement vers la limite du disque de visualisation, comme si toute cette hiérarchie était vue à travers un objectif fisheye, visible dans sa totalité mais sans que ce soit au détriment du contexte. L'arbre hyperbolique est dynamique : une animation amène un noeud visité au centre de l'affichage. Les noeuds peuvent contenir des liens hypertextes et mener le navigateur vers d'autres pages.

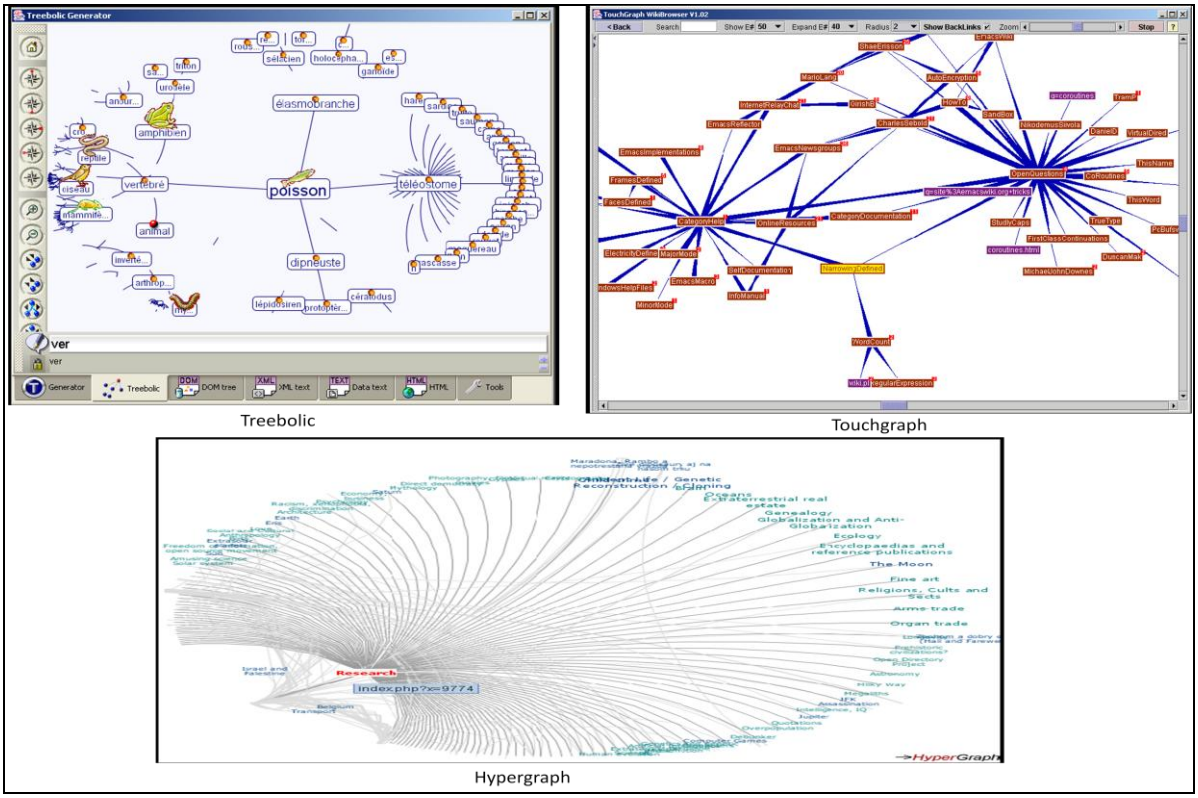


Figure 6.15 Exemples de représentation avec Treebolic, Hypergraph et Touchgraph

L'outil choisi Treebolic n'étant pas dédié aux Topic Maps, nous avons dû **adapter** cet applet dans notre plateforme. Cette adaptation a nécessité un travail d'implémentation pour pouvoir intégrer la DTD XTM dans l'applet et visualiser ainsi la Topic Map.

De plus, nous avons **étendu** Treebolic en développant le code nécessaire pour que l'utilisateur puisse accéder à un document de son choix à partir de la Topic Map tout en navigant dans sa structure. Malgré ce travail d'adaptation et d'extension, nous avons obtenu de meilleurs résultats au niveau de la visualisation avec Treebolic en comparaison avec les outils de visualisation dédiés aux Topic Maps.

La figure 6.16 présente un aperçu de la Topic Map générée et visualisée avec notre version adaptée de Treebolic. La Topic Map affiche au centre le Topic principal et les Topics thèmes (de premier niveau).

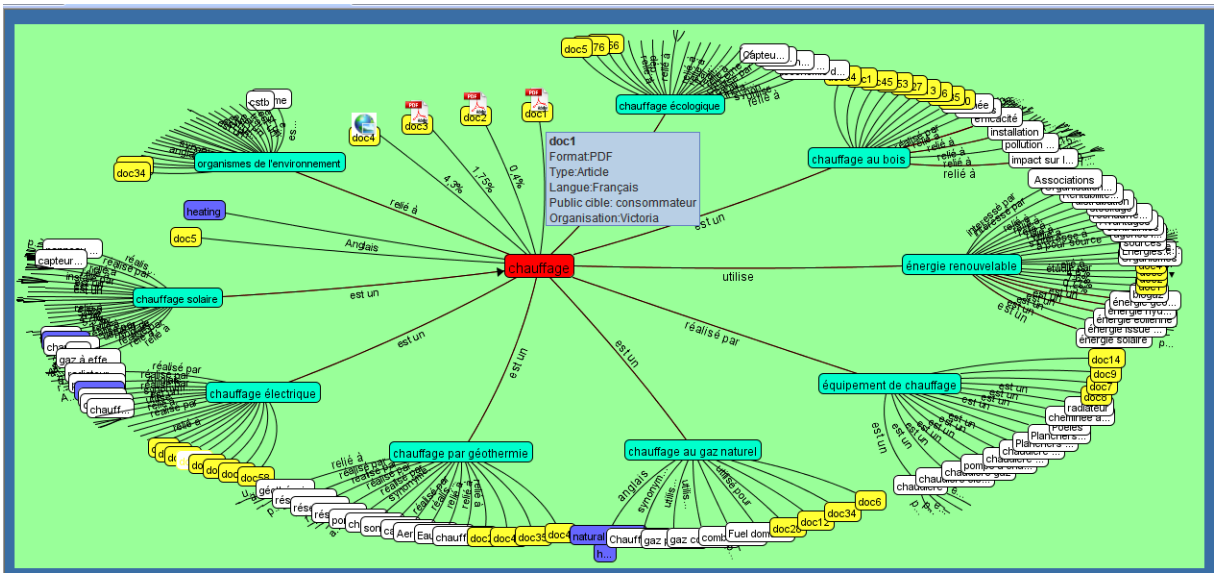


Figure 6.16 Interface de visualisation de la Topic Map

L'utilisateur peut afficher les méta-informations permettant de décrire un document comme par exemple son type, sa langue, le format, le public cible. Il a aussi la possibilité d'accéder au document en entier en cliquant sur le nœud du document sur la carte. Notre interface de visualisation offre des possibilités de mise en relief, c'est à dire quand un nœud est sélectionné, il est automatiquement agrandi en montrant la partie de la Topic Map reliée à ce nœud.

La figure 6.17 montre un exemple où l'utilisateur a sélectionné le Topic « chauffage solaire » comme Topic principal.

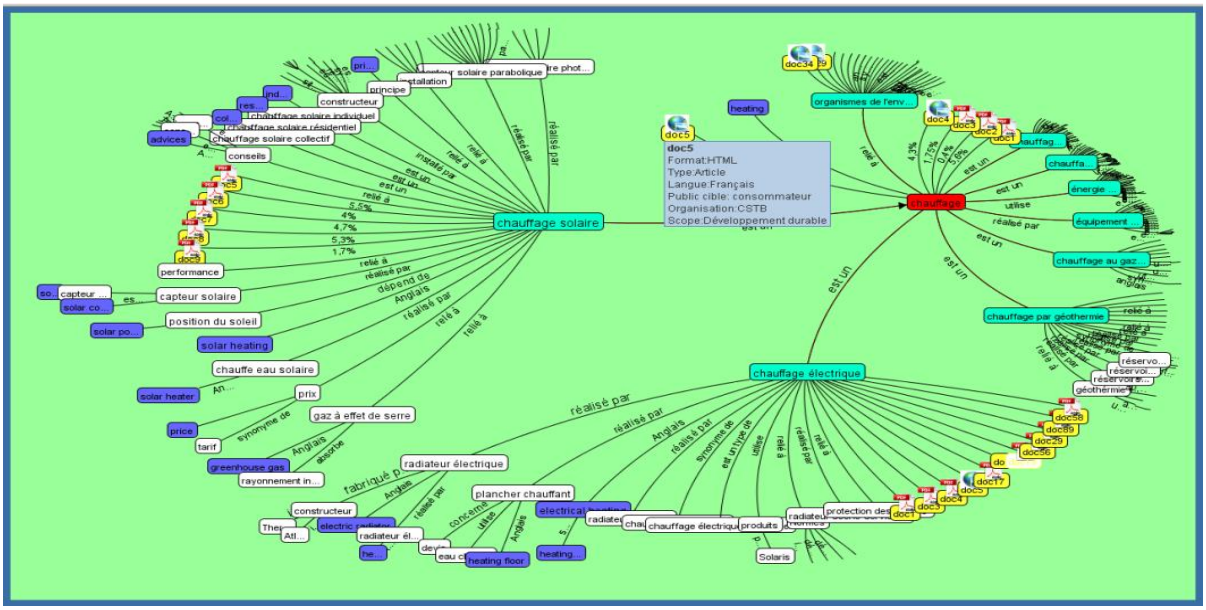


Figure 6.17 Exemple : Choix du Topic « chauffage solaire » comme Topic principal

La figure 6.18 montre un autre exemple où l'utilisateur a choisi de visualiser le Topic « chauffage au bois » comme Topic principal.

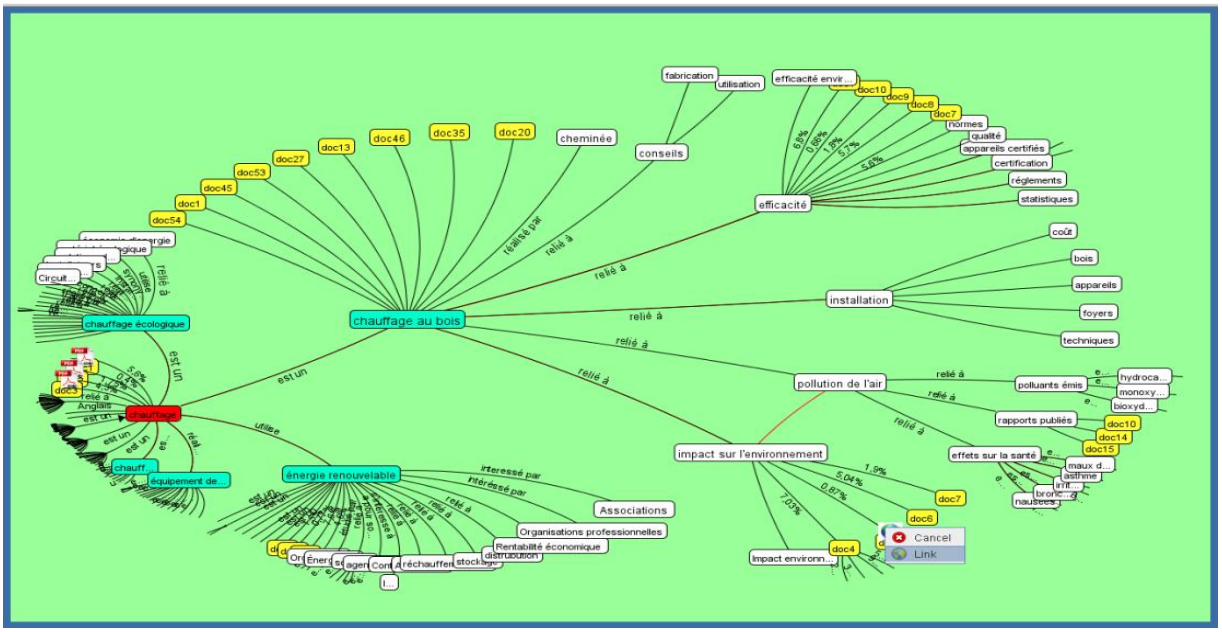


Figure 6.18 Exemple : Choix du Topic « chauffage au bois » comme Topic principal

Comme nous l'avons déjà mentionné, lors de la navigation l'utilisateur peut afficher le contenu du document en entier en cliquant sur le nœud correspondant à travers l'interface, par exemple dans la figure 6.19 l'utilisateur a choisi d'afficher un document sur les systèmes de chauffage.

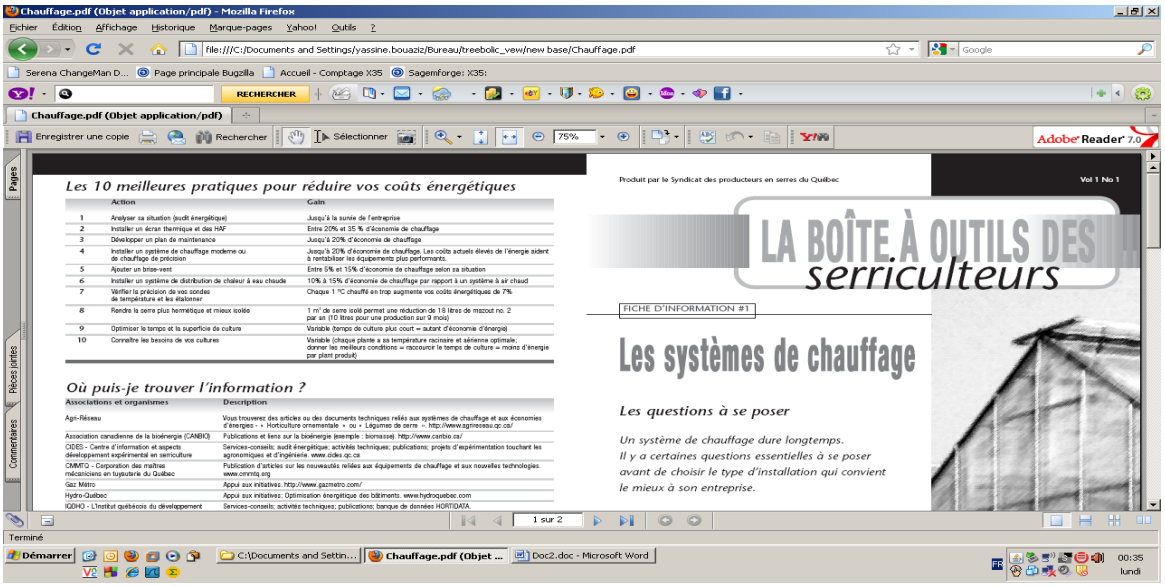


Figure 6.19 Visualisation d'un document sur les systèmes de chauffage à partir de la Topic Map

6.5 Conclusion

Dans ce chapitre, nous avons validé notre approche de construction incrémentale de Topic Map à partir d'un corpus de documents textuels multilingues, ce corpus contient des documents qui sont écrits en français et en anglais. Nous avons développé une plateforme de recherche intelligente fondée sur le modèle des Topic Maps. Nous avons montré que la Topic Map générée par notre plateforme, permet de faciliter la recherche dans les documents du corpus. En effet, en plus du mode de recherche classique par requête, nous avons représenté la Topic Map dans une interface interactive sous la forme d'une carte sémantique qui contient les Topics, les associations et les liens occurrences entre les Topics et les ressources, l'utilisateur aura alors la possibilité de naviguer dans cette carte et accéder aux documents et aux segments de documents qu'il cherche à travers les liens de la Topic Map.

Un sous-produit de ce travail est l'adaptation et l'extension du logiciel de visualisation de données Treebolic à la visualisation de Topic Maps. Cette nouvelle version du logiciel pourrait être utilisée par d'autres outils utilisant des Topic Maps. La principale extension a été la fonctionnalité de « *drill through* » d'un Topic vers les documents qu'il référence.

Dans nos travaux futurs, nous comptons agrandir le nombre de personnes qui participeront à la validation de notre plateforme. Le panel d'utilisateurs devra comprendre d'une part des « créateurs » de Topic Maps et d'autre part des « usagers » effectuant des recherches sur la plateforme. Nous devrons sélectionner des créateurs du domaine de la gestion documentaire qui sont aguerris à l'annotation des documents et des créateurs hors du domaine pour mesurer le degré d'assistance.

Parmi les usagers de la plateforme du côté interrogation, nous devrons établir un échantillon comprenant des personnes d'âge, de sexe et de niveau socioprofessionnel différents, notre plateforme étant destinée aux « hommes et aux femmes du monde ». L'étude de leur comportement face à la Topic Map pourra influencer les fonctionnalités destinées à améliorer la qualité de la Topic Map, par exemple celle proposée dans cette thèse concernant le volume de la Topic Map. Nous pourrons valider notre postulat de ne leur proposer qu'une partie de la Topic Map ou au contraire constater que l'utilisateur apprécie un affichage plus détaillé et moins ciblé.

CHAPITRE 7

Conclusion et perspectives

Dans le cadre de cette thèse, nous nous sommes intéressés aux modèles et aux techniques d'annotation sémantique de ressources Web dans le but de faciliter la recherche d'information dans ces ressources. Parmi les modèles existants, nous nous sommes intéressés aux Topic Maps, proposés par l'ISO comme un concurrent de RDF du W3C. Nous avons exposé le problème de construction de Topic Map à partir de documents textuels, nous avons également établi un état de l'art détaillé et une étude comparative des approches et des outils existants pour l'élaboration de Topic Maps. A partir de cette étude, nous avons montré que les principales limites de ces approches concernent la prise en compte de documents multilingues, l'intégration des requêtes dans le processus de construction de Topic Map et enfin l'évaluation de la qualité de la Topic Map générée.

Nous avons donc proposé une démarche pour la construction d'une Topic Map qui prend en compte un contenu textuel multilingue et l'évolution de la Topic Map selon le contenu et l'usage à travers l'exploration des questions potentielles relatives aux documents sources. Nous avons aussi proposé une méthode originale d'élagage dynamique de la Topic Map en rajoutant des méta-propriétés aux caractéristiques d'un Topic servant à mesurer son importance dans le temps. Dans la suite, nous exposons nos principales contributions et les perspectives ouvertes par nos travaux.

7.1 Contributions

7.1.1 Sur le plan théorique

1) Nous avons défini **un méta-modèle de Topic Map** comme une extension du modèle des Topic Maps proposé par l'ISO, dans ce méta-modèle, nous avons rajouté deux méta-propriétés aux caractéristiques de Topics, une méta-propriété servant à mesurer la pertinence du Topic au cours du temps et une deuxième méta-propriété pour l'organisation des Topics dans la Topic Map. En plus des liens d'occurrences, nous avons classé les liens de la Topic Map en deux classes : les liens ontologiques et les liens d'usage extraits à partir des requêtes utilisateurs. Les liens ontologiques et structurels regroupent les liens de spécialisation, le lien de composition ainsi que les liens associatifs que nous pourrions identifier suite à l'analyse des documents à organiser.
2) Pour enrichir et compléter notre méta-modèle de Topic Map, nous avons proposé un **méta-modèle de référentiel** sémantique de documents. Dans ce référentiel, après une phase de prétraitement, les documents sont segmentés thématiquement

en appliquant l'algorithme TextTiling et indexés sémantiquement en utilisant le modèle LSI. Nous avons généré pour chaque document, la liste de ses segments thématiques et la liste des termes et des concepts représentatifs de son contenu, ces termes et ces concepts sont pondérés par leurs degrés de pertinence (*tof*×*idf* pour les concepts et *tf*×*idf* pour les termes). Ces mesures ont servi pour pondérer les liens occurrences reliant un Topic aux documents (ou segments) qui lui font référence et filtrer ainsi ces ressources selon leur importance. Nous avons utilisé le référentiel pour l'annotation de la Topic Map c'est-à-dire l'ajout des liens occurrences pour annoter chaque Topic par la liste des documents et des segments qui en parlent triés selon leurs degrés de pertinence.

3) Nous avons **combiné le méta-modèle du référentiel avec celui de Topic Map,** cette combinaison nous a permis d'intégrer la notion de segments dans notre méta-modèle de Topic Map et nous a donné la possibilité d'indexer un Topic par un fragment au lieu du document en entier surtout dans le cas de documents de grande taille et qui traitent de plusieurs thématiques.
4) En se basant sur les deux méta-modèles définis, nous avons proposé une **A**pproche de **C**onstruction d'une **TO**pic Map **M**ultilingue, nommée **ACTOM**. Notre approche est **incrémentale et évolutive**, elle est basée sur un processus automatisé qui prend en entrée le référentiel de documents multilingues. Pour chaque document, nous commençons par extraire une liste de Topics et d'associations, à partir de ce document, en se basant sur le thésaurus du domaine, ensuite la Topic Map résultante est enrichie avec, d'une part les synsets et les liens des deux ontologies générales (WordNet et WOLF) et d'autre part, de nouveaux Topics et des liens d'usage extraits à partir des questions. La Topic Map obtenue est, par la suite, intégrée avec la Topic Map associée au deuxième document du référentiel. Ce processus est répété jusqu'à ce que nous terminions tous les documents disponibles. De part son processus incrémental, notre approche trouve son utilité toutes les fois où le contenu en entrée est enrichi par de nouveaux documents.
5) Notre approche **d'intégration de Topic Maps** présente l'avantage d'utiliser, non seulement les algorithmes de fusion de hiérarchies proposés par [Lammari et Métais, 2004], [Lammari et al. 2008] mais aussi le thésaurus du domaine et les ontologies générales qui fournissent les chainons manquants. En effet, le

thésaurus décrit le vocabulaire du domaine c'est-à-dire les termes techniques ainsi que les relations entre eux (généralisation/spécialisation, équivalence, associatives) alors que les ontologies telles que WordNet fournissent des informations générales du vocabulaire commun.

6) Nous avons proposé de prendre en compte **l'usage dans la Topic Map**, cette dernière est enrichie à partir d'un ensemble de scénarios d'usage, cet enrichissement consiste à ajouter les liens d'usage. Le lien d'usage est un hyper lien de type « répond à » (hyper lien questions/réponses) entre la question représentée comme un Topic et les réponses associées, c'est-à-dire les Topics référençant les documents qui permettent de répondre à la question. Nous avons proposé dans ce contexte de relier la question à chacun des mots clés la constituant via un hyper lien de type « est composé de ». Le stockage des liens « est-composé-de » d'une question vers les termes la composant nous a permis d'une part une recherche par navigation et d'autre part une recherche automatique de « question proche ».
7) Nous avons aussi travaillé sur **l'élagage de la Topic Map** générée, nous avons défini une méthode originale d'élagage dynamique à l'affichage de Topic Map dans le but de résoudre le problème de volume souvent rencontré dans les Topic Maps, notre méthode permet de faciliter la recherche et la navigation dans la Topic Map, elle est basée sur la définition de méta-propriétés que nous avons ajoutés aux caractéristiques d'un Topic, la première méta-propriété, la note du Topic, a servi pour mesurer la pertinence du Topic par rapport à son usage au cours du temps. La deuxième méta-propriété, le niveau auquel appartient le Topic, a servi pour organiser et clustériser les Topics dans la Topic Map.
8) A travers notre approche, nous avons produit une Topic Map globale multilingue comme **un espace sémantique de navigation.** Ce dernier est destiné à être la cible des procédures de recherche par navigation en plus du mode de recherche classique par requête en utilisant le référentiel de documents.

7.1.2 Sur le plan pratique

Nous avons développé une plateforme logicielle basée sur l'approche que nous avons proposée, notre plateforme est testée sur un corpus réel du domaine de la construction durable en particulier le sous domaine relatif aux solutions pour l'économie d'énergie.

7.2 Perspectives

Les travaux réalisés dans cette thèse ouvrent diverses perspectives.

Suppression de documents

Nous prévoyons d'étudier le problème de suppression d'un document. Une première solution possible est d'identifier les Topics reliés à ce document; s'ils ne sont reliés à aucun autre document et n'appartiennent pas au thésaurus, alors le document peut être enlevé. Une autre alternative pour résoudre ce problème est d'utiliser l'algorithme de [Lammari et Métais, 2004] qui propose de reconstruire une hiérarchie de concepts lorsqu'on en a enlevé des bouts. Pour notre cas, il serait alors possible d'enlever chaque Topic relié au document à supprimer en reliant son générique à ses spécifiques. Nous prévoyons dans le futur d'étudier ces solutions et éventuellement d'autres solutions pour résoudre le problème de suppression d'un document.

Extension de la norme XTM pour indexer un Topic par un segment

Une des originalités de notre approche consiste à combiner les deux méta-modèles de référentiel et de Topic Map que nous avons définis pour la conception de notre approche de construction de Topic Map. Nous prévoyons, à long terme, d'inclure le méta-modèle du référentiel dans la norme XTM pour, en particulier, intégrer la notion de segment dans la syntaxe XTM et pouvoir indexer un Topic par un segment de document et pondérer les liens occurrences entre Topics et documents par leurs degrés de pertinence.

Enrichissement du processus d'élagage de la Topic Map

Pour la gestion de la qualité de la Topic Map, nous avons choisi la solution de l'élagage dynamique. Nous souhaitons, dans nos futurs travaux, enrichir le processus d'élagage de la Topic Map, en intégrant, plus de critères (méta-propriétés) permettant de juger de la pertinence d'un Topic. Nous étudierons aussi la possibilité de leur généralisation aux autres concepts tels que le concept d'association.

Nous avons également défini le type de variation de la note d'un Topic comme une méta-méta-propriété qui nous avons utilisée pour changer automatiquement cette note. Nous projetons, dans nos prochains travaux, de faire une étude permettant de réunir les critères de qualité d'une Topic Map afin de déterminer de façon plus ou moins exhaustive la liste de méta-propriétés utiles à la gestion des évolutions d'une Topic Map.

Extension sur la qualité de la Topic Map

Une autre perspective de notre travail concerne la qualité de la Topic Map, la continuation naturelle de cette thèse serait, dans le futur, d'étudier les autres critères de qualité d'une Topic Map et de proposer des méthodes d'évaluation, des règles et des algorithmes d'amélioration de la qualité d'une Topic Map.

Bibliographie

[Abel et al. 2003] Abel M.-H., Lenne D., Moulin C., Benayache A. : Gestion des ressources pédagogiques d'une e-formation. ***In*** Documents Numériques, pp. 111-128, 2003.

[Abel, 2004] Abel M.-H.: Utilisation de normes et standards dans le projet MEMORAe, ***In*** Distances et savoirs. Vol 2/4, pp. 487- 511, 2004.

[Agirre et al. 2000] Agirre E., Ansa O., Hovy E., Martinez D. : Enriching very large ontologies using the WWW. ***In*** Proceedings of ECAI 2000 workshop on Ontology Learning, 2000.

[Agrawal et Srikant, 1997] Agrawal R., Srikant R.: Mining generalized association rules. Future Generation Computer Systems, 13(2–3), pp. 161–180, 1997.

[Ahmed, 2003] Ahmed K.: TMShare – Topic Map Fragment Exchange in a Peer-To-Peer Application, 2003. [On Line]: http://www.idealliance.org/papers/dx_xmle03/papers/02-03-03/02-03-03.pdf

[Ahmed, 2005] Ahmed K. : Topic Map Relational Query Language: TMRQL. Networked Planet White paper, 2005.[On Line]: http://www.networkedplanet.com/download/TMRQL.pdf

[Ahmed, 2009] Ahmed K. : Making Topic Maps SPARQL, 2009. [On Line]: http://www.networkedplanet.com/ontopic/2009/11/making_topic_mas_sparql.html

[Akoka et al. 2007a] Akoka J., Comyn-Wattiau I., Sisaïd-Cherfi S. : Évaluation de la qualité des modèles conceptuels, Projet QUADRIS, Marseille, juillet 2007, [On Line] : http://www.prism.uvsq.fr/apmd-quadris/exposes/Akoka.pdf

[Akoka et al. 2007b] Akoka J., Berti-Équille L., Boucelma O., Bouzeghoub M., Comyn-Wattiau I., Cosquer M., Goasdoué V., Kedad Z., Nugier S., Peralta V., et Sisaïd-Cherfi S.: A Framework for Quality Evaluation in Data Integration Systems. ***In*** Proceedings of the 9th International Conference on Enterprise Information Systems (ICEIS 2007), pp. 170-175, 2007.

[Akoka et al. 2008] Akoka, J., Berti-Équille L., Boucelma O., Bouzeghoub M., Comyn-Wattiau I., Cosquer M., Goasdoué V., Kedad Z., Nugier S., Peralta V., et Sisaïd-Cherfi S.: Évaluation de la qualité des systèmes multisources. Une approche par les patterns. Proceedings of the 2nd Workshop on Data and Knowledge Quality (QDC 2008) in conjunction with the French National Conf. on Extraction and Management of Knowledge (Extraction et Gestion des Connaissances - EGC), 2008.

[Aleksovski et al. 2006] Aleksovski Z., Klein M., Kate W. T., Harmelen F. V.: Matching Unstructured Vocabularies using a Background Ontology. ***In*** Proceedings of the 15th International Conference on Knowledge Engineering and Knowledge Management, pp. 182-197, 2006.

[Aljlayl et Frieder, 2001] Aljlayl M., Frieder O.: Effective Arabic-English Cross-Lingual Information Retrieval via Machine-Readable Dictionaries and Machine Translation. ***In*** ACM Tenth Conference on Information and Knowledge Management, Atlanta, Georgia, November, 2001.

[Aussenac-Gilles et al. 2000] Aussenac-Gilles N., Biébow B., Szulman S. : Modélisation du domaine par une méthode fondée sur l'analyse de corpus. ***In*** Actes de la conférence IC'2000, Journées Francophones d'Ingénierie des connaissances, pp. 93-103, 2000.

[Azouaou, 2005] Azouaou F., Dung Cao T., Dehors S., Desmoulins C., Dieng-Kuntz R., Faron-Zucker C. : Les Outils du Web sémantique et du E-Learning. ***In*** Actes de la Journée thématique WebLearn sur le Web sémantique pour le e-Learning, plate-forme AFIA'2005, Nice, 31 mai 2005.

[Bach et al. 2004] Bach T.-L., Dieng-Kuntz R., Gandon F. : On Ontology Matching Problems for building a Corporate Semantic Web in a Multi-Communities Organization. ***In*** the proceedings of the 6th International Conference on Enterprise Information Systems (ICEIS 2004), pp. 236-243, 2004.

[Baget et al. 2004] Baget J.F., Canaud E., Euzenat J., Kacid M.D. : Les langages du web sémantique. ***In*** le Web sémantique, Charlet J., Laublet p. & Reynaud C. (Eds.), Hors série de la revue Information - Interaction - Intelligence (I3), 4(1), Cépaduès, Toulouse, pp. 21-43, 2004.

[Ballestros et Croft, 1998] Ballesteros L., Croft W.: Resolving Ambiguity for Cross-Language Retrieval. ***In*** Proceedings of the 21st ACM SIGIR'98, pp. 64-71, 1998.

[Bansiya et David, 2002] Bansiya J., Davis C.G. : A Hierarchical Model for Object-Oriented Design Quality Assessment. ***In*** IEEE Transactions on Software Engineering, vol. 28, n° 1, pp. 4-17, January 2002.

[Baziz et al. 2003] Baziz M., Boughanem M., Nassr N. : La recherche d'information multilingue : désambiguïsation et expansion de requêtes basées sur WordNet. ***In*** International Symposium On Programming and Systems (ISPS 2003), pp. 175-186, May 2003.

[Baziz, 2005] Baziz M. : Indexation conceptuelle guidée par ontologie pour la recherche d'information. Thèse de doctorat, IRIT, Institut de recherche en informatique de Toulouse. Université Paul Sabatier, 2005.

[Benayache, 2005] Bennayache A. : Construction d'une mémoire organisationnelle de formation et évaluation dans un contexte e-learning: le projet MEMORAE. Thèse de doctorat. Université Technologique de Compiègne (UTC), décembre 2005.

[Bendaoud et al. 2007] Bendaoud R., Rouane Hacene M., Toussaint Y., Delecroix B., Napoli A.: Construction d'une ontologie à partir d'un corpus de textes avec l'ACF. ***In*** Actes des 18ème journées francophones d'ingénierie des connaissances, IC 2007, F. Trichet (Ed.), pp. 121-133, 2007.

[Bergamaschi et al. 2001] Bergamaschi, S., Castano S., Vincini M., Beneventano D. : Semantic Integration of Heterogeneous Information Sources. ***In*** Data & Knowledge Engineering 36: 3, pp. 215–249, 2001.

[Berners-Lee et al. 2001] Berners-lee T., Hendler J., Lassila O.: The Semantic Web. ***In*** Scientific American, pp. 20-88, May 2001.

[Berners-Lee, 1998] Berners-Lee T. : Semantic Web road map. Internal note, World Wide Web Consortium, September 1998. [OnLine]: http://www.w3.org/DesignIssues/semantic.html

[Böhm et al. 2002] Böhm K., Heyer G., Quasthoff U., Wolff Ch.: Topic Map Generation Using Text Mining. ***In*** Journal of Universal Computer Science, vol. 8, N° 6, pp. 623-633; 2002

[Borst, 1997] Borst, W. : Construction of Engineering Ontologies for Knowledge Sharing and Reuse: Thèse de doctorat, University of Twente, 1997.

[Bourigault et al. 2004] Bourigault D., Aussenac-Gilles, N., Charlet J. : Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas, ***In*** Techniques Informatiques et Structuration de Terminologies, Pierrel J.M. et Slodzian M. (Eds.), Numéro Spécial de la Revue d'Intelligence Artificielle (RIA), 18(1), Paris, 2004.

[Bourigault et al. 2005] Bourigault D., Fabre C., Frérot C., Jacques M.-P., Ozdowska S.: Syntex : analyseur syntaxique de corpus. ***In*** Actes des 12èmes journées sur le Traitement Automatique des Langues Naturelles, 2005.

[Bourigault, 1996] Bourigault D.: LEXTER: a Natural Language Processing tool for terminology extraction. ***In*** Proceedings of the 7th EURALEX International Congress, 1996.

[Bozsak et al. 2002 Bozsak E., Ehrig M., Handschuh S., Hotho A., Maedche A., Motik B., Oberle D., Schmitz C., Staab S., Stojanovic L., Stojanovic N., Studer R., Stumme G., Sure Y., Tane J., Volz R., Zacharias V. : KAON: Towards a large scale Semantic Web. ***In*** Proceedings of the Third International Conference on E-Commerce and Web Technologies (ECWeb), LNCS vol. 2455, pp. 304–313, 2002

[Brank et al. 2005] Brank J., Grobelnik M., Mladenic D. : A survey of ontology evaluation techniques. ***In*** Proceedings of conference on data mining and data warehouses (SIKDD'05), 2005. [On Line]: http://kt.ijs.si/dunja/sikdd2005/Papers/BrankEvaluationSiKDD2005.pdf

[Burton-Jones et al. 2004] Burton-Jones A., Storey V.C., Sugumaram V., Ahluwalia P.: A semiotic metrics suite for assessing the quality of ontologies. ***In*** Data and Knowledge Engineering, 2004.

[Caillet et al. 2004] Caillet M., Pessiot J.F., Amini M., and Gallinari P.: Unsupervised learning with term clustering for thematic segmentation of texts. ***In*** the 7th Proceedings of Recherche d'Information Assistée par Ordinateur (RIAO'04), pp. 1–11, March 2004.

[Calvanese et al. 2001] Calvanese D., Giacomo G. D., Lenzerini M.: A framework for ontology integration, ***In*** Proceedings of the First Semantic Web Working Symposium, 2001.

[Caussanel et al. 2002] Caussanel J., Cahier J.-P., Zacklad M., Charlet J. : Les Topic Maps sont-ils un bon candidat pour l'ingénierie du Web Sémantique ?, ***In*** Actes de IC' 2002, 2002.

[Chaar, 2003] Chaar S. L. : Extraction de segments thématiques pour la classification de résumé multi-document orienté par un profil utilisateur, ***In*** RECITAL 2003, juin 2003

[Charlet et al. 2004] Charlet J., Bachimont B., Troncy R. : Ontologies pour le web sémantique. ***In*** Le web sémantique, Charlet J., Laublet p. & Reynaud C. (Eds.), Hors série de la revue Information - Interaction - Intelligence (I3), 4(1): pp. 69-100, 2004.

[Choi et al. 2001] Choi F. Y., Wiemer-Hastings P., Moore J.: Latent semantic analysis for text segmentation. ***In*** Proceedings of Conference on Empirical Methods in Natural Language Processing, pp. 109–117, 2001.

[Choi, 2000] Choi F. Y.: Advances in domain independent linear text segmentation. ***In*** Proceedings of the first conference on North American chapter of the Association for Computational Linguistics, pp. 26–33, 2000.

[Cimiano et Volker, 2005] Cimiano P., Volker J.: Text2onto: a framework for ontology learning and data-driven change discovery. ***In*** A. MONTOYO, R. MUNOZ & E. METAIS (Eds.): Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB). Lecture Notes in Computer Science (LNCS) volume 3513, pp. 227–238, 2005.

[Cunningham et al. 2002] Cunningham H., Maynard D., Bontcheva K., Tablan V.: GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. ***In*** Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02*)*, pp. 168-175, July 2002.supp

[Daille, 1999] Daille B. : Identification des adjectifs relationnels en corpus. ***In*** Actes de la Conférence de Traitement Automatique du Langage Naturel (TALN'99), ATALA, pp. 105-114, 1999.

[Davies et al. 2002] Davies J., Fensel D., Van Harmelen F.: Towards the Semantic Web, ***In*** Ontology-driven Knowledge Management. Wiley, First Edition January 21, 2003.

[Deerwester et al. 1990] Deerwester S.C., Dumais S., Lander T.K., Furnas G.W., Harshan R.A.: Indexing by Latent Semantic Analysis. ***In*** Journal of the American Society of Information Science, volume 41, pp. 391-407, 1990.

[Dicheva et Dichev, 2006] Dicheva D., Dichev C.: TM4L: Creating and Browsing Educational Topic Maps, ***In*** British Journal of Educational Technology (BJET), 37(3):391-404, 2006.

[Dicks et al. 2004] Dicks, D., Venkatesh, V., Shaw, S., Lowerison, G., Zhang, D.: An Empirical Evaluation of Topic Map Search Capabilities in an Educational Context. ***In*** L. Cantoni et C. McLoughlin (Eds.): Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2004, pp. 1031-1038, 2004.

[Dietel et al. 2001] Deitel A., Faron C., Dieng R.: Learning ontologies from RDF annotations. ***In*** Proceedings of IJCAI Workshop on Ontology learning, 2001

[Djedidi et Aufaure, 2008] Djedidi R., Aufaure M.A. : Enrichissement d'ontologies : maintenance de la consistance et évaluation de la qualité, ***In*** Actes 19èmes journées francophones d'Ingénierie des Connaissances (IC'08), 2008.

[Do et Rahm, 2001] Do H., Rahm E.: COMA – A system for flexible combination of schema matching approaches. ***In*** the proceedings of the 27th International Conference on Very Large DataBases (VLDB 2001), pp. 610-621, 2001.

[Doan et al. 2001], Doan, A.H., Domingos P., Halevy A.: Reconciling Schemas of Disparate Data Sources: A Machine-Learning Approach. ***In*** SIGMOD 2001, 2001.

[Doan et al. 2004] Doan A., Madhavan J., Domingos P., Halevy A.: Ontology Matching: A Machine Learning Approach, ***In*** Handbook on Ontologies in Information Systems, S. Staab and R. Studer (eds.), Springer-Velag, pp. 397-416, 2004.

[Dumais et al. 1996] Dumais, S., Landauer, T., Littman, M.: Automatic cross-linguistic information retrieval using Latent Semantic Indexing. SIGIR - Workshop on Cross-Linguistic Information Retrieval, pp. 16-23, 1996.

[Dumas et al. 1997] Dumas L., Plante A., Plante P. : ALN: Analyseur Linguistique de ALN, vers.1.0. ATO, UQAM, 1997.

[Ehrig et Staab, 2004] Ehrig M., Staab S. : QOM - Quick Ontology Mapping. ***In*** Proceedings of the International Semantic Web Conference 2004: pp. 683-697, 2004.

[Ellouze et al. 2008a] Ellouze N., Ben Ahmed M. Métais E.: State of the Art on Topic Maps Building Approaches, ***In*** Proceedings of MBSDI 2008, Model Based Software and Integration Systems, R.-D. Kutsche and N. Milanovic (Eds.), pp. 102-112, 2008.

[Ellouze et al. 2008b] Ellouze N., Métais E., Ben Ahmed M.: State of The Art of Topic Maps Building Approaches, ***In*** International Journal of Computer Science and Software Technology, Vol. 1, No. 1, International Sciences Press, pp. 51-57, January-June 2008.

[Ellouze et al. 2009a] Ellouze N., Lammari N., Métais E., Ben Ahmed M. : CITOM: Approche de construction incrémentale d'une Topic Map multilingue, ***In*** Actes du Workshop RISE Recherche d'information sémantique (associé à INFORSID 2009), Catherine Roussey et Jean-Pierre Chevallet, pp. 65-85, Mai 2009.

[Ellouze et al. 2009b] Ellouze N., Lammari N., Métais E., Ben Ahmed M. : Usage-Oriented Topic Map Building Approach, ***In*** Proceedings of Third International Conference on Metadata and Semantics Research MTSR 2009, Fabio Sartori, Miguel Angel Sicilia et Nikos Manouselis, volume 46, pp. 13-23, 2009.

[Ellouze et al. 2009c] Ellouze N., Lammari N., Métais E., Ben Ahmed M. : CITOM: Incremental Construction of Topic Maps, ***In*** Proceedings of 14th International Conference on Applications of Natural Language to Information Systems NLDB 2009, pp. 49-61, June 2009.

[Euzenat et al. 2005] Euzenat, J., Guégan P., Valtchev P.: OLA in the OAEI 2005 Alignment Contest. In Integrating Ontologies, 2005.

[Euzenat et Valtchev, 2004] Euzenat J., Valtchev P.: Similarity-based ontology alignment in OWL-lite. ***In*** Proceedings of the European Conference on Artificial Intelligence (ECAI), pp. 333–337, 2004.

[Faatz et Steinmetz, 2002] Faatz A., Steinmetz R.: Ontology enrichment with texts from the WWW, ***In*** Proceedings of the Semantic Web Mining Conference WS'02, 2002.

[Fellah et al. 2008] Fellah A., Malki M., Zahaf A. : Alignement des ontologies : utilisation de WordNet et une nouvelle mesure structurelle, ***In*** CORIA 2008 Conférence en Recherche d'Information et Applications, pp 401-408, 2008.

[Fellbum, 1998] Fellbum, C.: WordNet: An Electronic Lexical Database. MIT press, 1998.

[Fernandez et al. 1997] Fernandez M., Gómez-Pérez A., Juristo N.: METHONTOLOGY: from ontological art towards ontological engineering. ***In*** Proceedings of AAAI Spring Symposium on Ontological Engineering, pp. 33-40, 1997. [On Line]: http:// delicias.dia.fi.upm.es/miembros/ASUN/SSS97.ps

[Fernandez et al. 1999] Fernández M, Gómez-Pérez A., Pazos J., Pazos A. : Building a Chemical Ontology using Methondology and ontology design environment. ***In*** IEEE Intelligent Systems and their applications, vol.14, pp. 37-46, 1999.

[Ferret, 2002] Ferret O. : Segmenter et structurer thématiquement des textes par l'utilisation conjointe de collocations et de la récurrence lexicale. ***In*** Actes de TALN 2002, pp. 155–165, 2002.

[Ferret, 2006] Ferret O. : Approches endogène et exogène pour améliorer la segmentation thématique de documents. ***In*** Traitement Automatique des Langues, Volume 47, N° 2, pp. 111-135, 2006.

[Ferruci et Lally, 2004] Ferruci D., Lally A.: UIMA: an architecture approach to unstructured information processing in a corporate research environment. ***In*** Natural Language Engineering, 10(3-4): 327–348, 2004.

[Folch et Habert, 2002] Folch H., Habert H.: Articulating conceptual spaces using the Topic Map standard. ***In*** Proceedings of XML'2002, pp. 8-13, December 2002.

[Fortuna et al. 2006] Fortuna B., Grobelnik M., Mladenic D.: Semi-automatic data driven ontology construction system. ***In*** Proceedings of the 9th International multiconference Information Society (IS-2006), 2006.

[Fox et al. 1998] Fox M.S., Barbuceanu M., Gruninger M., Lin J. : An organization ontology for enterprise modelling, Simulating organizations, MIT Press, 1998.

[Frath et al. 2000] Frath P., Oueslati R., Rousselot F. : Identification de relations sémantiques par repérage et analyse de cooccurrences de signes linguistiques, ***In*** Ingénierie des connaissances. Évolutions récentes et nouveaux défis. Paris : Eyrolles, Jean Charlet, Manuel Zacklad, Gilles Kassel, Didier Bourigault (Eds.), pp. 291-304, 2000.

[Galley et al. 2003] Galley M., McKeown K., Fosler-Lussier E., Jing H.: Discourse Segmentation of Multi-party Conversation. ***In*** Proceedings ACL. Sapporo, Japan,.pp. 562-569, 2003.

[Gangemi et al. 1999] Gangemi A., Pisanelli D. M., Steve G.: An Overview of the ONIONS Project: Applying Ontologies to the Integration of Medical Terminologies. ***In*** Data Knowledge Engineering. V31. 1999.

[Garshol, 2001] Garshol L. M.: Tolog: A Topic Map Query Language. ***In*** Proceedings of XML Europe 2001, May 2001. [On Line] : http://www.ontopia.net/

[Garshol, 2003] Garshol L. M.: Living with Topic Maps and RDF, ***In*** XML Europe 2003, 2003. [On Line] : http://www.ontopia.net/topicmaps/materials/tmrdf.html

[Garshol, 2005a] Garshol L. M.: Development Manager: TMQL: An introduction, 2005, [On Line]: http://www.emnekart.no/2005/forum-04-19/tmql-intro.pdf

[Garshol, 2005b] Garshol L. M.: tolog: Language tutorial. Ontopia Knowledge Suite documentation, 2005 [On Line]: http://www.ontopia.net/omnigator/docs/query/tutorial.html

[Giunchiglia et al. 2004] Giunchiglia F., Shvaiko P., Yatskevich M.: S-Match: an Algorithm and an Implementation of Semantic Matching. ***In*** proceedings of ESWS 2004, pp. 61-75, 2004.

[Godehardt et Bhatti, 2008] Godehardt E., Bhatti N. : Using Topic Maps for Visually Exploring Various Data Sources in a Web-Based Environment. ***In*** Scaling Topic Maps: Third International Conference on Topic Maps Research and Applications, TMRA 2007, pp. 51–56, 2008.

[Gollins et Sanderson, 2000] Gollins T., Sanderson M.: Sheffield University CLEF 2000 – Bilingual Track: German to English. ***In*** Lecture Notes in Computer Science, volume 2069, pp 248-255, 2000.

[Gomez-Perez et Macho, 2003] Gomez-Perez A., Manzano Macho, D.: Survey of ontology learning methods and techniques. Deliverable 1.5 OntoWeb Project Documentation, Universidad Politécnica de Madrid, 2003. [On Line]: http://www.deri.at/fileadmin/documents/deliverables/Ontoweb/D1.5.pdf

[Gomez-Perez et Rojas-Amaya, 1999] Gomez P.A., Rojas-Amaya D.: Ontological Reengineering for Reuse. ***In*** Fensel D., Studer R. (Eds.): Proceedings of the 11th European Workshop on Knowledge Acquisition, Modeling and Management (EKAW-99), LNAI volume 1621, 26–29, pp. 139–156, 1999.

[Grefenstette et al. 2005] Grefenstette G., Semmar N., Gara F.: Modifying a Natural Language Processing System for European Languages to Treat Arabic in Information in Processing and Information Retrieval Application. ***In*** Proceedings of the ACL workshops, pp 31-38, 2005

.

[Grefenstette, 1998] Grefenstette G.: The Problem of Cross-language Information Retrieval. ***In*** Cross-Language Information Retrieval, Gregory Grefenstette (Ed.), pp. 1-9, 1998.

[Gronmo, 2002] Gronmo G.O.: Automagic Topic Maps. 2002. [On Line] : http://www.ontopia.net/topicmaps/materials/automagic.html

[Grosz et Sidner, 1986] Grosz B. J., Sidner C. L.: Attention, intentions, and the structure of discourse. ***In*** Computational Linguistics, 12(3):175–204, 1986.

[Gruber, 1993] Gruber T. : A translation approach to portable ontology specifications, In Knowledge Acquisition Journal, 5(2), Academic Press, pp. 199-220, 1993.

[Gruninger et Fox, 1995] Gruninger M., Fox M. S.: Methodology for the Design and Evaluation of Ontologies. ***In*** Proceedings of Workshop on Basic Ontological Issues in Knowledge Sharing, IJCAI-95, Knowledge Sharing, 1995.

[Gyun et Park, 2007] Gyun O. S., Park O., N.: Design and Users' Evaluation of a Topic Map-Based Korean Folk Music Retrieval System. ***In*** L. Maicher, A. Sigel, and L.M. Garshol (Eds.): Proceedings of TMRA 2006, LNAI 4438, pp. 74–89, 2007.

[Hahn et al. 2004] Hahn U., Markó K., Poprat M., Schulz S., Wermter J., Nohama P.: Crossing Languages in Text Retrieval via an Interlingua. ***In*** Proceedings of RIAO conference, pp. 100-115, 2004.

[Han et Karypis, 2000] Han E-H., Karypis G.: Centroid based document classification: Analysis and experimental results. ***In*** Proceedings of the 4th European Conference of Principles of Data Mining and Knowledge Discovery, pp. 424–431, 2000.

[Harrathi et al. 2005] Harrathi F., Calabretto S., Roussey C. : Indexation semi-automatique de corpus multilingues basée sur une ontologie. ***In*** Journal Sciences et techniques de l'information, ISSN 1762-8288 , pp. 203-219, 2005.

[Harrathi, 2009] Harrathi F. : Extraction de concepts et de relations entre concepts à partir de documents multilingues : Approche statistique et ontologique, Thèse de doctorat, INSA de Lyon, Septembre 2009.

[Hartmann et al. 2005] Hartmann J., Spyns P., Giboin A., Maynard D., Cuel R., Suarez-Figueroa M.C., Sure Y. : Methods for ontology evaluation. Knowledge Web D1.2.3 deliverable, 2005.

[Hearst, 1992] Hearst M.A.: Automatic acquisition of hyponyms from large text corpora. ***In*** Proceedings of the 14th conference on Computational linguistics, Volume 2, pp. 539 – 545, 1992

[Hearst, 1997] Hearst M. A.: Textiling: segmenting text into multi-paragraph subtopic passages. ***In*** Computational Linguistics, 23(1): 33–64, 1997.

[Heinecke, 2003] Heinecke J., Léger .A : MKBEEM – Web Sémantique pour le e-Commerce Multilingue, AFIA 2003. [On Line] http://heinecke.pagesperso-orange.fr/litlist/doc/AFIA-MKBEEM.pdf

[Hernandez et Mothe, 2006] Hernandez N., Mothe J.: D'un thesaurus vers une ontologie de domaine pour l'exploration d'un corpus. ***In*** Actes de la conférence Veille Stratégique Scientifique et Technologique VSST, 2006.

[Hernandez, 2005] Hernandez N. : Ontologies de domaine pour la modélisation du contexte en recherche d'information, Thèse de doctorat, Institut de recherche en informatique de Toulouse, université Paul Sabatier, 6 décembre 2005.

[Jacquemin et Bourigault, 2003] Jacquemin C., Bourigault D.: Term Extraction and Automatic Indexing. ***In*** MITKOV R. (Ed.): ***In*** The Oxford Handbook of Computational Linguistics, Oxford University Press, pp. 599-615, 2003.

[Jarke et al. 2000] Jarke M., Lenzerini M., Vassiliou Y., Vassiliadis P. : Fundamentals of Data Warehouse, Springer, ISBN 3-540-65365-1, 2000.

[Jiang et al. 2009] Jiang L., Liu J., Wu Z., Zheng Q., Qian Y.: ETM Toolkit: A development tool based on Extended Topic Map. ***In*** Proceedings of 13th International Conference on Computer Supported Cooperative Work in Design (CSCWD), pp. 528-533, 2009.

[Jiang et Conrath, 1997] Jiang J., Conrath D.: Semantic similarity based on corpus statistics and lexical taxonomy. ***In*** Proceedings of International Conference on Research in Computational Linguistics, 1997.

[Johnson et Shneiderman, 1991] Johnson B., Shneiderman B.: Tree-maps: A space-filling approach to the visualization of hierarchical information structures. ***In*** Proceedings of IEEE Visualization'91, pp. 284-291, 1991.

[Kalfoglou et Schorlemmer, 2003] Kalfoglou Y., Schorlemmer M.: Ontology mapping: the state of the art. ***In*** Knowledge Engineering Review, 18(1):1-31, 2003.

[Kan et al. 1998] Kan M. Klavans J. McKeown K. : Linear segmentation and segment significance; ***In*** Proceedings of 6th Workshop on Very Large Corpora, ACL SIG-DAT, pp. 197-205, 1998.

[Kaski et al. 1998] Kaski S., Honkela T., Lagus K., Kohonen T.: WEBSOM : self-organizing maps of document collections. Neurocomputing, volume 21, pp. 101-117, 1998.

[Kasler et al. 2006] Kasler L., Venczel Z., Varga L.Z. : Framework for Semi Automatically Generating Topic Maps. ***In*** Proceedings of the 3rd international workshop on text-based information retrieval TIR-06, Riva del Grada, pp 24-30, 2006.

[Kefi et Reynaud, 2006] Kefi H., Reynaud C., Safar B.: Techniques structurelles pour l'alignement de taxonomies sur le web. ***In*** Atelier Fouille du Web - EGC 2006, 2006.

[Khan et Luo, 2002] Khan L., Luo F.: Ontology construction for information selection. ***In*** Proceedings of IEEE international conference on tools with IA, November 2002.

[Khan, 2000] Khan L., Ontology-based Information Selection, Thèse de doctorat, Faculty of the Graduate School, University of Southern California. August 2000.

[Kietz et al. 2000] Kietz J., Maedche A., Volz R.: A method for semi automatic Acquisition from Corporate Intranet. ***In*** Proceedings of Workshop "Ontology and text" (EKAW'2000), 2000.

[Kim et al. 2007] Kim J. M., Shin H., Kim H.J.: Schema and constraints-based matching and merging of Topic Maps. ***In*** Information Processing and Management, volume 43, pp. 930-945, 2007.

[Kong et al. 2006] Kong H., Hwang M., Kim P. : Efficient Merging for heterogeneous domain ontologies based on WordNet, ***In*** Journal JACIII, 10(5), pp 733-737, 2006

[Korthaus et al. 2004] Korthaus, A., Köhler, C. and Schader, M. : Semi-Automatic Topic Map generation from a conventional document index, ***In*** Proceedings of Int. Conf. Knowledge Sharing and Collaborative Engineering, KSCE 2004, US Virgin Islands, pp.101-108, 2004.

[Korthaus et al. 2006] Korthaus A., Aleksy M., Henke S., Schader M.: A Distributed Topic Map Architecture for Enterprise Knowledge Management. ***In*** Proceedings of the 1st IEEE/ACIS Workshop on Component-Based Software Engineering, Software Architecture and Reuse (COMSAR '06), 2006.

[Korthaus et al. 2009] Korthaus A., Markus A., Stefan H.: A distributed knowledge management infrastructure based on a Topic Map grid, ***In*** International Journal of High Performance Computing and Networking, Volume 6, Number 1, pp. 66 – 80, 2009.

[Labadié et Chauché, 2007] Labadié A., Chauché J. : Segmentation thématique par calcul de distance sémantique. ***In*** Actes de Extraction et Gestion des Connaissances EGC'07, 2007.

[Lammari et al. 2008] Lammari N., Du Mouza C., Métais E. : POEM: an Ontology Manager based on Existence Constraints. ***In*** Proceedings of International Conference on Database and Expert Systems Applications (DEXA'08), pp. 81-88, 2008.

[Lammari et Besbes Essanaa, 2009] Lammari N. et Besbes Essanaa S. : Rétro-Conception de Sites Web : Extraction du Contenu Informatif, ***In*** Vers l'Ingénierie des Evolutions, Revue des Sciences et Technologies de l'Information série Ingénierie des Systèmes d'Information (RTSI série ISI), 0(0), 2009.

[Lammari et Métais, 2004] Lammari N., Métais, E.: Building and Maintaining Ontologies: a Set of Algorithms, ***In*** Data and Knowledge Engineering, 48(2): 155-176, 2004.

[Landauer et al. 1998] Landauer T., Foltz P-W., Laham D.: Introduction to Latent Semantic Analysis. Discourse Processes, pp. 25: 259-284, 1998.

[Laublet, 2007] Laublet P. : Web Sémantique et Ontologies, ***In*** Nouvelles technologies cognitives et concepts des sciences humaines et sociales, Volume 1, Humanités Numériques, Hermès, Paris, 2007.

[Lavik et al. 2004] Lavik, S., Nordeng, T. W., Meloy, J. R. BrainBank Learning - building personal Topic Maps as a strategy for learning, ***In*** Proceedings of XML 2004, [On Line] : http://www.gca.org/xmlusa/2004/slides/lavik/BrainBankLearning-building-personal-topicmaps-as-strategy-for-learning.pdf

[Legrand et Soto, 1999] Legrand B., Soto M. : Navigation dans des structures hiérarchiques de grande dimension – Application à la supervision de réseaux, Journées Doctorales Informatique et Réseaux (JDIR'99), novembre 1999.

[Legrand et Soto, 2002a] Legrand B., Soto M.: Topic Maps et navigation intelligente sur le Web Sémantique, AS CNRS Web Sémantique, CNRS Ivry-sur-Seine, Octobre 2002.

[Legrand et Soto, 2002b] Legrand B., Soto M.: Visualization of the Semantic Web: Topic Maps Visualisation. ***In*** Proceedings of IEEE Sixth International Conference on Information Visualisation (IV'02), pp. 344, July 2002.

[Legrand et Soto, 2006] Legrand B., Michel S. : Visualisation exploratoire, généricité, exhaustivité et facteur d'échelle. ***In*** Numéro spécial de la revue RNTI Visualisation et extraction des connaissances, mars 2006.

[Legrand, 2001] Legrand B.: Extraction d'information et visualisation de systèmes complexes sémantiquement structurés. Thèse de doctorat. Université Pierre et Marie Curie, Décembre 2001.

[Li et Clifton, 2000] Li W., Clifton C.: SemInt: A Tool for Identifying Attribute Correspondences in Heterogeneous Databases Using Neural Network. ***In*** Data and Knowledge Engineering 33: 1, 49-84, 2000.

[Librelotto et al. 2004] Librelotto G.R. Ramalho J. C., Henriques P. R.: TM-Builder: An Ontology Builder based on XML Topic Maps. ***In*** Clei electronic journal, vol. 7, N° 2, paper 4, 2004.

[Librelotto et al. 2008] Librelotto G.R., Ramalho J. C., Henriques P. R.: A framework to specify, extract and manage topic maps driven by ontology. ***In*** Proceedings of SIGDOC, pp. 155-162, 2008.

[Lin et Qin, 2002] Lin, X., Qin, J.: Building a Topic Map Repository, 2002. [On Line]: http://www.knowledgetechnologies.net/proceedings/presentations/lin/xialin.pdf

[Lin, 1998] Lin D.: An information-theoretic definition of similarity. ***In*** Proceedings of 15th International Conference On Machine Learning, 1998.

[Lozano-Tello et Gomez-Perez, 2004] Lozano-Tello A., Gomez-Perez A.: Ontometric: A method to choose the appropriate ontology. ***In*** Journal of Database Management, Vol. 15, N°2, pp. 1-18, 2004.

[Lu et al. 2008] Lu H., Feng B., Zhao Y., Zheng Q., Liu J.: Distributed Knowledge Management Based on Extended Topic Maps. ***In*** Proceedings of the International Conference on Computer Science and Software Engineering, Volume 01, pp. 649-652, 2008.

[Lu et al. 2009] Lu H., Feng B., Zhao Y., Zheng Q., Liu J.: Distributed Knowledge Fusion Based on Extended Topic Maps. ***In*** Journal of Jilin University (Science Edition), 2009.

[Lu et al. 2010] Lu H., Feng B., Li X.: Similarity Algorithm of Extended Topic Maps for Multi-Resource Knowledge Fusion. ***In*** Journal of Xi'an Jiaotong University, 2010.

[Lu et Feng, 2009] Lu H., Feng B.: An intelligent topic map-based approach to detecting and resolving conflicts for multi-resource knowledge fusion. ***In*** Information Technology Journal, vol. 8, pp. 1242-1248, 2009.

[Mackinlay et al. 1991] Mackinlay J. D., Robertson G. G., Card S. K., The Perspective Wall: Detail and Context Smoothly Integrated. ***In*** Proceedings of Human Factors in computing Systems CHI'91 Conference, pp. 173-179, 1991.

[Madhavan et al. 2001] Madhavan, J., Bernstein, P. A., Rahm, E.: Generic matching with Cupid. ***In*** International Journal of Very Large Data Bases (VLDB), 10(4): 334-350, 2001.

[Maedche et Staab, 2000] Maedche A., Staab S.: Mining ontologies from text. ***In*** Proceedings of the 12th European Workshop on Knowledge Acquisition, Modeling and Management. Lecture Notes In Computer Science (LNCS), Vol. 1937 pp. 189-202, 2000.

[Maedche et Staab, 2001] Maedche A., Staab S.: Ontology learning for the semantic Web. ***In*** IEEE Intelligent Systems, Special Issue on Semantic Web, 16(2): 46-53, 2001.

[Maedche et Staab, 2002] Maedche A., Staab S.: Measuring similarity between Ontologies. ***In*** Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW'02), LNCS/LNAI 2473, pp. 251-263, 2002.

[Maedche et Staab, 2004] Maedche A., Staab S.: Handbook on Ontologies, chapter Ontology Learning. ***In*** Springer: Handbook in Information Systems., pp. 173–190, 2004

[Maicher et Witschel, 2004] Maicher L., Witschel H. F.: Merging of Distributed Topic Maps based on the Subject Identity Measure (SIM) Approach. ***In*** Proceedings of Berliner XML tags, pp. 301–307, 2004.

[McGregor et al. 1999] McGregor R., Chalupsky H., Moriarty D., et Valente A.: Ontology Merging with OntoMorph, 1999, [On Line] : http://reliant.teknowledge.com/HPKB/meetings/meet040799/Chalupsky/index.htm, 1999.

[McGuiness et al. 2000] McGuiness D.L., Fikes R., Rice J., Wilder S.: An environment for Merging and Testing Large Ontologies. ***In*** Proceedings of Seventh International Conference on Principles of Knowledge Representation and Reasoning (KR'2000), April 2000.

[Meadche et al. 2003] Meadche A., V Pekar., S. Staab: Ontology Learning Part One- On discovering Taxonomic Relations from the web. ***In*** N. Zhong and J. Liu and Y. Yao: Web Intelligence, Springer-Verlag, 2003.

[Melnik et al. 2002] Melnik S., Garcia-Molina H., Rahm E.: Similarity Flooding: A Versatile Graph Matching Algorithm. ***In*** Proceedings of ICDE 2002, 2002.

[Moore, 2001] Moore G. : RDF and Topic Maps: an exercise of convergence, *XML Europe 2001*, [On Line] : http://www.topicmaps.com/topicmapsrdf.pdf.

[Morris et Hirst, 1991] Morris J., Hirst G.: Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structuree of Text. Computational Linguistics, vol. 17, n° 1, pp. 21-48, 1991.

[Muller et al. 2004] Muller H.-M., Kenny E. E., Sternberg P. W.: Textpresso: an ontology based information retrieval and extraction system for biological literature, PLoS Biology, 2(11):1984–1998, 2004.

[Munzner, 1997] Munzner T.: H3: Laying Out Large Directed Graphs in 3D Hyperbolic Space. ***In*** Proceedings of IEEE Symposium on Information Visualization, 1997

[Nassr et Boughanem, 2002] Nassr N., Boughanem M. : Croisement de langues en recherche d'information : traduction et désambiguïsation de requêtes par phrases alignées. ***In*** Actes du XX^{ème} Congrès INFORSID, Juin 2002.

[Neshatian et Hejazi, 2004] Neshatian K., Hejazi, M. R.: Text categorization and classification in terms of multi-attribute concepts for enriching existing ontologies. ***In*** Proceedings of 2nd Workshop on Information Technology and its Disciplines, pp. 43–48, 2004.

[Névéol et Ozdowska , 2005] Névéol A., Ozdowska S.: Extraction bilingue de termes médicaux dans un corpus parallèle. ***In*** Actes des 5ème journées Extraction et Gestion des Connaissances (EGC), pp. 655-666, 2005.

[Newcomb, 2002] Newcomb S. R., Hunting S., Algermissen J., Durusau P.: The Reference Model for Topic Maps (RM4TM), [On Line] : http://www.isotopicmaps.org/rm4tm/RM4TM-official.html.

[NISO, 2004] National Information Standards Organisation (NISO), Understanding Metadata, 2004, [On Line]: http://www.niso.org/standards/resources/UnderstandingMetadata.pdf

[Nobécourt, 2000] Nobécourt J.: A method to build formal ontologies from text. ***In*** Proceedings of EKAW 2000, 2000.

[Noy et Musen, 2000] Noy N. F., Musen M.: PROMPT: Algorithm and Tool for Automated Ontology Merging and Alignment. ***In*** Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI-2000), pp 450-455, 2000.

[Noy et Musen, 2001] Noy N. F., Musen M. A.: Anchor-Prompt: Using non-local context for semantic matching. ***In*** Proceedings of Workshop on Ontologies and Information Sharing, International Joint Conferences on Artificial Intelligence (IJCAI-01), 2001.

[Noy et Musen, 2003] Noy N. F., Musen M. A.: The PROMPT suite: Interactive tools for ontology merging and mapping. ***In*** International Journal of Human-Computer Studies, 59(6):983–1024, 2003.

[Oard et Dorr, 1996] Oard W.O, Dorr B.: A Survey of Multilingual Text Retrieval. Report UMIACS-TR-96-19 CS-TR-3615, 1996. [On Line]: http://www.clis.umd.edu/dlrg/filter/papers/mlir.ps

[Oard et Hackett, 1997] Douglas W. Oard, Hackett P. : Document Translation for Cross-Language Text Retrieval at the University of Maryland, ***In*** The Sixth Text Retrieval Conference, pp. 687-696, 1997.

[Ogievetsky, 2001] Ogievetsky N.: XML Topic Maps through RDF glasses. ***In*** Extreme Markup Languages 2001, Montréal. [On Line]: http://www.cogx.com/rdfglasses.html

[Ouziri, 2006] Ouziri M.: Semantic integration of Web-based learning resources: A Topic Maps-based approach. ***In*** Proceedings of the Sixth International Conference on Advanced Learning Technologies (ICALT2006), 2006.

[Parekh et al. 2004] Parekh V., Gwo J-P., Finin T.: Mining Domain Specific Texts and Glossaries to Evaluate and Enrich Domain Ontologies. ***In*** Proceedings of International Conference of Information and Knowledge Engineering, 2004.

[Passonneau et Litman, 1993] Passonneau R. J., Litman D. J. ,: Intention-based segmentation: human reliability and correlation with linguistic cues, ***In*** Annual Meeting of the ACL archive, Proceedings of the 31st annual meeting on Association for Computational Linguistics, pp. 148 – 155, 1993

[Pepper, 2000] Pepper S.: The TAO of Topic Maps: Finding the way in the age of infoglut. ***In*** Proceedings of XML Europe 2000 Conference, June 2000.

[Pepper, 2002a] Pepper S.: Topic Map Erotica RDF and Topic Maps "in flagrante". 2002. [On Line]: http://www.ontopia.net/topicmaps/materials/MapMaker_files/frame.htm

[Pepper, 2002b] Pepper S.: Methods for the Automatic Construction of Topic Maps. 2002. [On Line]: http://www.ontopia.net/topicmaps/materials/autogen-pres.pdf

[Pepper, 2008] Pepper S.: Topic Maps, Article for the Encyclopedia of Library and Information Sciences, 2008, [On Line]: http://www.ontopedia.net/pepper/papers/ELIS-TopicMaps.pdf.

[Pirkola, 1998] Pirkola A.: The Effects of Query Structure and Dictionary Setups in Dictionay-Based Cross-language Information Retrieval, ***In*** Proceedings of ACM-SIGIR'98, 1998.

[Pivano et al. 2004] Pivano B., Calabretto S., Roussey C., Pinon J.M. : SyDoM: un système de recherche d'information multilingue basé sur les connaissances. ***In*** Actes IC'2004. pp. 103-114, 2004.

[Popov et al. 2004] Popov B., Kiryakov A., Ognyanoff D., Manov D., Kirilov A.: KIM: A Semantic Platform For Information Extraction and Retrieval Journal of Natural Language Engineering, Vol. 10, Issue 3-4, , pp. 375-392, September 2004.

[Rada et al. 1989] Rada R., Mili H., Bicknell E., Blettner M.: Development and application of a metric on semantic nets, ***In*** IEEE Transaction on Systems, Man and Cybernetics, 19(1), pp 17-30, 1989.

[Radwan, 1994] Radwan K. : Vers l'accès multilingue en langage naturel aux bases de données textuelles. Thèse de doctorat, Université Paris-sud, Centre d'Orsay, 1994.

[Rahm et Bernstein, 2001] Rahm E., Bernstein P. A. : A survey of approaches to automatic schema matching. ***In*** Journal VLDB, Volume 10, 2001, pp 334-335.

[Rasgado et Guzman, 2006] Rasgado A. D. C., Guzman A. A. : A language an Algorithm for Automatic Merging of Ontologies. ***In*** Proceedings of Conference on Computing (CIC), pp 180-185, 2006.

[Resnik, 1995] Resnik, P. : Using Information Content to Evaluate Semantic Similarity in a Taxonomy, ***In*** proceedings of International Joint Conference for Artificial Intelligence (IJCAI-95), pp. 448-453, 1995.

[Reynar, 2000] Reynar J., Topic Segmentation: Algorithms and applications. Thèse de doctorat, Université de Pennsylvania, Seattle, WA, 2000.

[Reynolds et Kimber, 2002] Reynolds J., Kimber W.E.: Topic Map Authoring With Reusable Ontologies and Automated Knowledge Mining. ***In*** Proceedings of XML'2002, 2002.

[Rijsbergen, 1979] Rijsbergen C. V.: Information Retrieval. Butterworth – Heinemann, Newton (MA), 1979.

[Roberson et Dicheva, 2007] Roberson, S., Dicheva, D.: Semi-Automatic Ontology Extraction to Create Draft Topic Maps, ***In*** Proceedings of 45th ACM Southeast Conference, pp. 23-24, March 2007.

[Roitman et Gal, 2006] Roitman Haggai and Gal Avigdor: OntoBuilder: Fully Automatic Extraction and Consolidation of Ontologies from Web Sources Using Sequence Semantics. ***In*** Lecture Notes in Computer Science, Volume 4254, pp. 573-576, 2006.

[Rousselot et al. 1996] Rousselot F., Frath P., Oueslati R.: Extracting concepts and relations from Corpora. ***In*** Proceedings of the Workshop on Corpus-oriented Semantic Analysis, European Conference on Artificial Intelligence (ECAI'96), 1996.

[Roussey et al. 2001] Roussey C., Calabretto S., et Pinon J.-M.: SyDoM: A Multilingual Information Retrieval System for Digital, ***In*** proccedings of International Conference ICCC/IFIP On Electronic Publishing (ELPUB'2001), pp. 150-164, July 2001.

[Roussey et al. 2002] Roussey C., Calabretto S., et Pinon J.-M. : SyDoM: un outil d'annotation pour le Web sémantique. ***In*** Journées scientifiques Web sémantique, octobre 2002.

[Roux et al. 2000] Roux C., Proux D., Rechermann F., Julliard L.: An ontology enrichment method for a pragmatic information extraction system gathering data on genetic interactions, ***In*** Proceedings of the ECAI2000 Workshop on Ontology Learning, 2000.

[Sagot et Fišer 2008] Sagot Benoît, Fišer Darja: Construction d'un wordnet libre du français à partir de ressources multilingues. ***In*** Proccedings of TALN 2008, France, 2008.

[Salton, 1983] Salton G., Mcgill M.: Introduction to Modern Information Retrieval. New York : McGraw-Hill, 1983.

[Salton, 1988] Salton G., Buckley C.: Term-weighing approaches in automatic text retrieval. ***In*** Information Processing & Management, 24(5):513-523, 1988.

[Salton, 1989] Salton G.: Automatic text processing, Addisson-Wesley, Reading, MA, USA, 1989.

[Sarkar et Brown, 1994] Sarkar M., Brown M.: Graphical Fisheye Views. Communications of the ACM, Vol. 37, N°12, pp. 73-84, 1994.

[Schauble et Braschler, 2000] Schauble P, Braschler M.: Experiments with the Eurospider Retrieval System for CLEF 2000 Workshop of the Cross language Evaluation Forum, CLEF 2000, LNCS, volume 2069, 2000.

[Schnurr et al. 2000] Schnurr H., Sure Y., Studer R.: On-To-Knowledge Methodology-Baseline version Deliverable. On-To-Knowledge EU-IST-1999-10132 Project D15. August 2000.[On Line] : http://www.ontoknowledge.org/del.shtml

[Sheridan et Ballerini, 1996] Sheridan P., Ballerini J. P.: Experiments in multilingual information retrieval using SPIDER system. ***In*** Proceedings of ACM SIGIR'96, pp. 58-65, 1996.

[Shvaiko et Euzenat, 2004] Shvaiko, P., Euzenat, J.: A survey of Schema-based Matching Approaches. Technical Report DIT-04-087. Informatica e Telecomunicazioni, University of Trento, 2004.

[Shvaiko et Euzenat, 2005] Shvaiko P., Euzenat J.: A Survey of Schema-based Matching Approaches. ***In*** Journal on Data Semantics, vol. N°, pp. 146-171, 2005.

[Sisaïd-Cherfi et al. 2002] Sisaïd-Cherfi S., Akoka J., Comyn-Wattiau I. : Conceptual Modeling Quality – From EER to UML Schemas Evaluation. ***In*** Proceedings of ER 2002, pp 414-428, 2002.

[Sisaïd-Cherfi et al. 2006] Sisaïd-Cherfi, S., Akoka J., Comyn-Wattiau I. : Use Case Modeling and Refinement: A Quality-Based Approach. ***In*** Proceedings of ER 2006, pp 84-97, 2006.

[Smith, 2001] Smith B., Wetly C.: Ontology: Towards a New Synthesis. ***In*** ACM Proceedings of FOIS'01, pp. 3-9, October 2001.

[Soergel et al. 2004] Soergel D., Lauser B., Liang A., Fisseha F., Keizer J. and Katz S.: Reengineering thesauri for new applications: The Agrovoc Example, ***In*** Journal of digital information, volume 4, issue 4, article n° 257, 2004.

[Sowa, 1984] Sowa J. F. : Conceptual Structures : Information Processing in Mind and Machine. Ed. Addison-Wesley, 1984.

[Sparck-Jones, 1972] Sparck-Jones K. : Statistical interpretation of term specificity and its application in retrieval. ***In*** Journal of Documentation, 28(1), pp. ll-20, 1972.

[Stefanova et Risch, 2010] Stefanova S., Risch T. : SPARQL queries to RDFS views of Topic Maps Source. ***In*** International Journal of Metadata, Semantics and Ontologies, Volume 5, N° 1, pp. 1-16, 2010.

[Storey et Sandeep, 2004] Veda Storey C., Sandeep P. : Understanding Relationships: Classifying Verb Phrase Semantics, Conceptual Modeling. ***In*** Berlin-Heidelberg: Springer-Verlag, Lecture Notes in Computer Science (LNCS), Volume 3288/2004, pp. 336-347, 2004.

[Stuckensschmidt et al. 2004] Stuckensschmidt H., Harmelen F. V., Serafini L., Bouquet P., Giunchiglia F. : Using C-OWL for the Alignment and Merging of Medical Ontologies. ***In*** Proceedings of the First International WS on Formal Biomedical K. R. (KRMed), 2004.

[Studer et al. 1998] Studer R., Benjamins V.R. et Fensel D. : Knowledge Engineering: Principles and Methods. ***In*** IEEE Transactions on Data and Knowledge Engineering, 25(1&2): 161-197, 1998.

[Stumme et al. 2006] Stumme G., Hotho A., Berendt B. : Semantic Web mining: State of the art and future directions. ***In*** Journal of Web Semantics, 4(2):124–143, June 2006.

[Stumme et Madeche, 2001] Stumme G., Madeche A. : FCA-Merge: Bottom-Up Merging of Ontologies. ***In*** B. Nebel (Ed.): Proceedings of 17th International Joint Conference on Artificial Intelligence, pp. 225-230, 2001.

[Supekar, 2006] Supekar K.: A peer-review approach for ontology evaluation. ***In*** Proceedings de la 8ème conference Internationale Protégé, 2006.

[Sure et al. 2003] Y Sure., Akkermans H., J Broekstra., J Davies., Ding Y., Duke A., R. Engels, D Fensel., I Horrocks., V Iosif., A Kampman, Kiryakov A., Klein M., Lau T., Ognyanov D., Reimer U., Simov K., Studer R., Meer J., Van Harmelen F.: On-To-Knowledge: Semantic Web Enabled Knowledge Management. ***In***: N. Zhong and J. Liu and Y. Yao: Web Intelligence, Springer-Verlag, 2003.

[Suryanto et Compton, 2001] Suryanto H., Compton P.: Discovery of ontologies from Knowledge Bases. ***In*** Proceedings of First International conference on knowledge Capture, October, 2001.

[Thanh Le et al. 2004]. Thanh Le, B., Dieng-Kuntz R., Gandon F.: On Ontology Matching Problems for building a Corporate Semantic Web in a Multi-Communities Organization. ***In*** Proceedings of ICEIS 2004, pp. 236-243, 2004.

[Uchida, 2004] Uchida H. : The Universal Networking Language (UNL) Specifications Version 3 Edition 3. UNL Center, UNDL Foundation, December 2004, 43 pages. [On Line] : http://www.undl.org/unlsys/unl/UNLSpecs33.pdf

[Utiyama et Isahara, 2001] Utiyama M., Isahara H.: A Statistical model for domain-independent text segmentation. ***In*** Proceedings of ACL'2001, pp. 491–498, 2001.

[Valtchev, 1999] Valtchev, P. : Construction automatique de taxonomies pour l'aide à la représentation de connaissance par objets. Thèse de doctorat, Université de Grenoble 1, 1999.

[Van Hage et al. 2005] Van Hage W., Katrenko S., Schreiber G.: A Method to combine Linguistic Ontology-Mapping Techniques. ***In*** Proceedings of International Semantic Web Conference, pp. 732-744, 2005.

[Van Wijk et Van De Wetering, 1999] Van Wijk J. J., Van De Wetering H. : Cushion Treemaps: visualisation of hierarchical information. ***In*** Proceedings of IEEE Symposium on Information Visualization, pp. 73-78, 1999.

[Velardi et al. 2001] Velardi P., Missikof M., Fabriani P.: Using text processing techniques to automatically enrich a domain ontology. ***In*** Proceedings of ACM- FOIS, 2001.

[Weerdt et al. 2006] Weerdt D. D., Pinchuk R., Aked R., Orus J. J., Fontaine B.: TopiMaker - An Implementation of a Novel Topic Maps Visualization. ***In*** Proceedings of TMRA 2006 - International Workshop on TopicMap Research and Applications, LNAI volume 4438, October, 2006

[Wills, 1999] Wills, G. J. : NicheWorks: Interactive Visualization of Very Large Graphs. ***In*** Journal of Computational and Graphical Statistics, vol. 8, pp 190-212, 1999.

[Wu et al. 2006] Wu X. F., Zhou L., Zhang L., and Ding Q.L. : TOM algorithm in distributed topic maps merging. ***In*** Engineering Journal of WuHan University, 39(5):131-136, 2006.

[Wu et Palmer, 1994] Wu Z., Palmer M. : Verb semantics and lexical selection. ***In*** Proceedings of 32nd Annual Meeting of The Association for Computational Linguistics, pp. 133-138, 1994.

[Xu et al. 2002] Xu F., Kurz D., Piskorski J., Schmeier, S. : A domain adaptive approach to automatic acquisition of domain relevant terms and their relations with bootstrapping. ***In*** Proceedings of the 3rd international conference on language resources and evaluation, 2002.

[Yang et al. 2006] Yang Z., Zhan D., Ye C. : Evaluation Metrics for Ontology Complexity and Evaluation Analysis. ***In*** Proceedings of IEEE International Conference on e-Business Engineering (ICEBE06), 2006.

[Zacklad et al. 2003a] Zacklad M., Caussanel J., Cahier J.P. : Un méta-modèle basé sur les Topic Maps pour la structuration et la recherche d'information, 2003. [On Line] http://enssibal.enssib.fr/autres-sites/RTP/websemantique/octobre/octobre4/zacklad.pdf

[Zacklad et al. 2003b] Zacklad M., Cahier J.P., Pétard X. : Du Web Cognitivement Sémantique au Web Socio-Sémantique. Journée « Web Sémantique et SHS ». 7 mai 2003. [On Line] : http://www.lalic.paris4.sorbonne.fr/stic/as5.html

[Zaher et al. 2006] Zaher L.H., Cahier J.-P., Zacklad M. : The Agoræ / Hypertopic approach. ***In*** Proceedings of the International Workshop IKHS - Indexing and Knowledge in Human Sciences, 2006.

[Zampa, 2005] Zampa V. : Utilisation de l'analyse sémantique latente pour tenter d'optimiser l'acquisition par exposition à une langue étrangère de spécialité, Vol. 8, n° 2: spécial Atala, pp. 135-146, 2005. [On Line] : http://alsic.revues.org/index339.html#tocto1n2

[Zargayouna, 2005] Haïfa Z. : Indexation sémantique de documents XML. Thèse de doctorat, Université Paris XI, 2005.

[Zipf, 1949] Zipf G. K. : Human behaviour and the principle of least effort, Cambridge, Mass, Addison-Wesley, 1949.

Annexes

Annexe A Algorithmes de segmentation thématique

Dans notre travail, npous avons utilisé TextTiling pour la segmentation thématiques des documents sources, nous proposons dans cet annexe de présenter les algorithmes de segmentation les plus connus parmi lesquels ***DotPlotting*** [Reynar, 2000], ***C99*** [Choi, 2000] et ***Segmenter*** [Kan et al. 1998]. Les quatre méthodes s'appuient sur la notion de cohésion lexicale, c'est à dire la répétition des mots comme indicateur d'homogénéité thématique.

Segmenter : chaînes lexicales

Segmenter [Kan et al. 1998] effectue une segmentation linéaire basée sur les chaînes lexicales présentes dans le texte. Ces chaînes relient les occurrences des termes dans les phrases. Une chaîne est rompue si le nombre de phrases séparant deux occurrences est trop important. Ce nombre dépend de la catégorie syntaxique du terme considéré. Une fois tous les liens établis, un poids leur est assigné en fonction de la catégorie syntaxique des termes en jeu et de la longueur du lien. Un score est ensuite donné à chaque paragraphe en fonction des poids et des origines des liens qui le traversent ou qui y sont créés. Les marques de segmentation sont alors apposées au début des paragraphes ayant les scores maximaux.

Etant donné qu'un concept peut être désigné par un ensemble de mots, le concept de chaînes lexicales a été élargi aux chaînes conceptuelles à l'aide de WordNet ou d'autres ressources sémantiques. Les auteurs de [Kan et al. 1998] montrent que l'amélioration est très peu significative.

DotPlotting : répétition de termes

L'algorithme DotPlotting a été proposé par Reynar en 2000 [Reynar, 2000]. Il se base sur une représentation graphique du texte par les positions des occurrences des termes du texte à segmenter. Lorsqu'un terme apparaît à deux positions du texte x et y, les quatre points (x, x), (x, y), (y, x) et (y, y) sont représentés sur un graphe, ce qui permet de déterminer visuellement les zones du texte où les répétitions sont nombreuses.

Cette méthode a été adaptée par [Reynar, 2000] à la segmentation thématique de textes. Les positions de début et de fin des zones les plus denses du graphe sont les limites des

segments thématiquement cohérents. La densité est calculée pour chaque unité d'aire en divisant le nombre de points d'une région par l'aire de cette région. A partir de là, deux algorithmes peuvent déterminer les frontières thématiques : identifier les limites en maximisant la densité au sein des segments, ou repérer la configuration qui minimise la densité des zones entre les segments.

C99 : mesure de similarité

Cet algorithme proposé par Choi dans [Choi, 2000] utilise une mesure de similarité entre chaque unité textuelle. L'idée de base de cette méthode est que les mesures de similarité entre des segments de textes courts sont statistiquement insignifiantes, et que donc seul des classements locaux sont à considérer pour ensuite appliquer un algorithme de catégorisation sur la matrice de similarité.

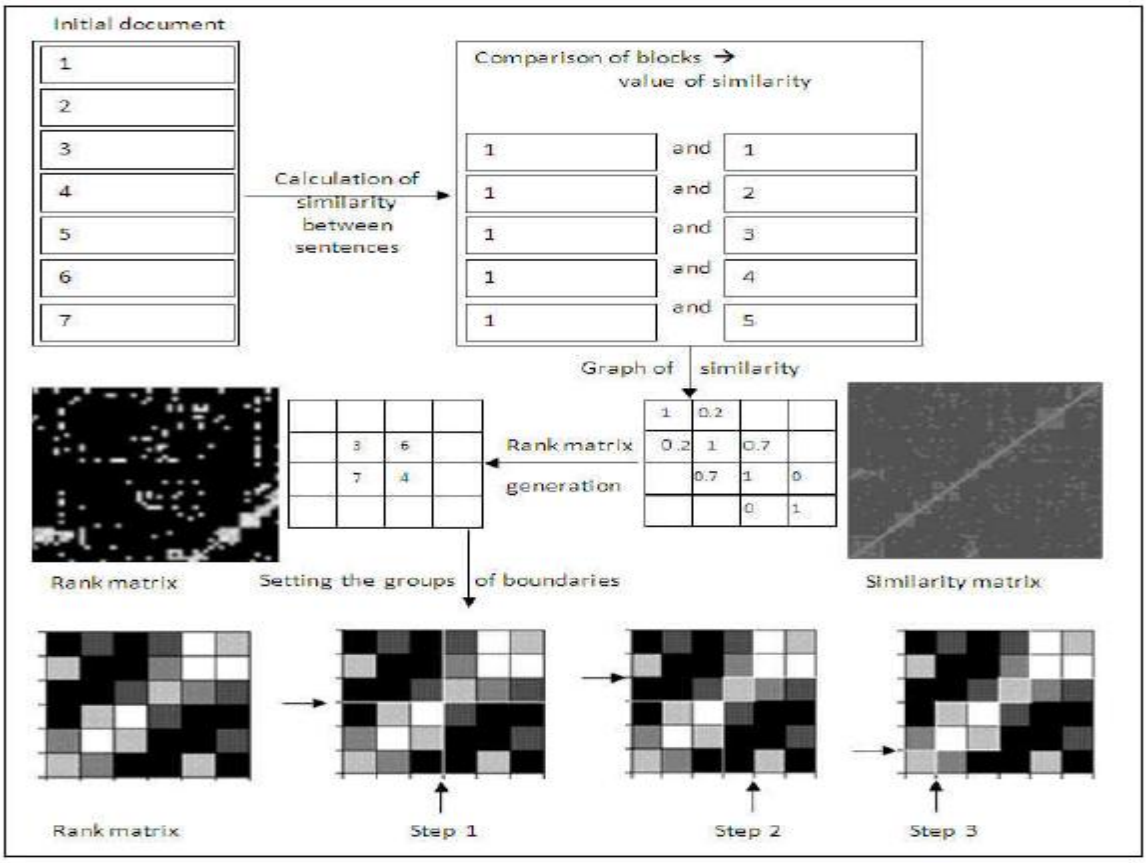


Figure A.1 Illustration des étapes de l'algorithme de Choi

Dans un premier temps, une matrice de similarité est donc construite, représentant la similarité entre toutes les phrases du texte à l'aide de la mesure de similarité proposée par Rijsbergen [Rijsbergen, 1979], calculée pour chaque paire de phrases du texte, en utilisant

chaque mot commun entre les phrases, et après « nettoyage » du texte : suppression des mots vides et lemmatisation.

On effectue ensuite un « classement local », en déterminant pour chaque paire d'unités textuelles, le rang de sa mesure de similarité par rapport à ses **m × n − 1** voisins, **m × n** étant le masque de classement choisi. Le rang est le nombre d'éléments voisins ayant une mesure de similarité plus faible, conservé sous la forme d'un ratio r afin de prendre en compte les effets de bord.

r = rang / nombre de voisins dans le masque

Enfin, la dernière étape détermine les limites de chaque segment de la même manière que l'algorithme ***Dotplotting*** emploie la maximisation. En effet, l'algorithme cherche à déterminer quelle configuration offre la plus grande densité, en recherchant une nouvelle limite thématique à chaque étape. Les segments sont alors représentés par des carrés le long de la diagonale de la matrice de similarité modifiée avec les classements locaux. Pour chaque segment de la répartition proposée à une étape de la segmentation on considère son aire notée $\mathbf{a_k}$ et son poids $\mathbf{s_k}$ qui est la somme des tous les rangs des phrases qu'il contient. La densité D de la configuration est calculée avec :

$$D = \frac{\sum_{k=1}^{m} s_k}{\sum_{k=1}^{m} a_k}$$

L'algorithme s'arrête lorsque la densité de la meilleure répartition proposée est suffisamment faible, ou si le nombre de frontières thématiques est déjà déterminé, lorsqu'il est atteint.

La procédure de Choi est ainsi composée de **trois** étapes. Tout d'abord, le document à segmenter est découpé en unités textuelles minimales, habituellement les phrases. Les mots composant ces phrases sont soumis à différents traitements comme la suppression de mots peu informatifs sur le thème du texte (article, pronom, verbes très fréquents, ...) et une lemmatisation. Ensuite, une mesure de similarité entre toutes les paires d'unités prises deux à deux est calculée. Enfin, le document est segmenté de façon récursive en fonction des frontières entre les unités textuelles qui maximisent la somme des similarités moyennes à l'intérieur des segments ainsi constitués.

L'étape la plus importante est celle qui calcule la similarité entre toutes les paires de phrases prises deux à deux. La procédure initialement proposée par Choi reposait sur la

métrique du cosinus appliquée aux vecteurs représentant les paires de phrases. Pour être déclarés cohérents, deux passages doivent contenir des mots communs. Il s'agit d'une conception très restrictive de la cohésion lexicale. Afin de dépasser cette limitation, dans un travail ultérieur, Choi et al. ont proposé d'utiliser l'analyse sémantique latente (ASL) pour estimer la similarité entre deux phrases. Pour ce faire, on applique la métrique du cosinus non aux vecteurs bruts, mais aux vecteurs pondérés par les dimensions sémantiques dérivées par l'analyse sémantique latente. Les étapes suivantes sont identiques quelle que soit la méthode employée pour calculer les similarités.

Comparaison entre les algorithmes de segmentation

Chaque méthode utilise des caractéristiques différentes pour réaliser la segmentation, les différences se portent sur le choix de catégories de vocabulaire. Tout d'abord, l'utilisation de la structure existante du texte ou non, la prise en compte d'un contexte global ou local voir les deux, et enfin le calcul de la similarité.

Étant donné toutes ces différences, il est difficile de dire qu'une méthode est mieux qu'une autre, bien que des comparaisons ont été effectuées sur les mêmes corpus, car la composition d'un corpus dépend du vocabulaire utilisé, du domaine, et de la structure. De plus, depuis quelques années, les procédures de validation sont devenues de moins en moins exigeantes. Alors que les premières études confrontaient la segmentation automatique aux jugements de lecteurs, la procédure actuelle est de soumettre à l'algorithme une série de textes concaténés artificiellement et mesurer sa capacité à trouver les concaténations des différents textes.

Et enfin, toutes ces méthodes ont été testées en anglais, une seule évaluation en français a été effectuée. Les expériences conduites sur les documents écrits en français montrent que l'algorithme TextTiling est le plus efficace. Les expériences ont confirmé le fait que le type de document que l'on segmente, son thème, sa taille et la variation de taille des segments à repérer, sont autant de caractéristiques à prendre en compte pour la segmentation.

Annexe B Liste des publications

Ellouze Nebrasse, Ben Ahmed Mohamed, Métais Elisabeth : State of the Art on Topic Maps Building Approaches, ***In*** Proceedings of MBSDI 2008, Model Based Software and Integration Systems MBSDI 2008, R. D.Kutsche and N. Milanovic (Eds.), pp.102-112, 2008.

Ellouze Nebrasse, Métais Elisabeth, Ben Ahmed Mohamed : State of The Art of Topic Maps Building Approaches, ***In*** International Journal of Computer Science and Software Technology, Vol. 1, No. 1, International Sciences Press, pp. 51-57, January-June 2008.

Ellouze Nebrasse, Lammari Nadira, Métais Elisabeth, Ben Ahmed Mohamed : CITOM: Approche de construction incrémentale d'une Topic Map multilingue, ***In*** Actes du Workshop RISE Recherche d'information sémantique (associé à INFORSID 2009), Catherine Roussey et Jean-Pierre Chevallet, pp. 65-85, Mai 2009.

Ellouze Nebrasse, Lammari Nadira, Métais Elisabeth, Ben Ahmed Mohamed : Usage-Oriented Topic Map Building Approach, ***In*** Proceedings of Third International Conference on Metadata and Semantics Research MTSR 2009, Fabio Sartori, Miguel Angel Sicilia et Nikos Manouselis, volume 46, pp. 13-23, 2009.

Ellouze Nebrasse, Lammari Nadira, Métais Elisabeth, Ben Ahmed Mohamed : CITOM: Incremental Construction of Topic Maps, ***In*** Proceedings of 14th International Conference on Applications of Natural Language to Information Systems NLDB 2009, pp. 49-61, June 2009.

Article soumis

Ellouze Nebrasse, Lammari Nadira, Métais Elisabeth, Ben Ahmed Mohamed : CITOM: Incremental Construction of Topic Maps, Article accepté suite à NLDB 2009 pour soumission dans le journal DKE.

Nebrasse ELLOUZE

Approche de recherche intelligente fondée sur le modèle des Topic Maps

Résumé

Dans le cadre de cette thèse, nous avons conçu et développé une approche que nous avons nommée ACTOM pour « **A**pproche de **C**onstruction d'une **TO**pic Map **M**ultilingue ». Cette dernière sert à organiser un contenu multilingue composé de documents textuels. Elle a pour avantage de faciliter la recherche d'information dans le contenu. Notre approche est incrémentale et évolutive, elle est basée sur un processus automatisé, qui prend en compte des documents multilingues et l'évolution de la Topic Map selon le changement du contenu en entrée et l'usage de la Topic Map. Elle prend comme entrée un référentiel de documents que nous construisons suite à la segmentation thématique et à l'indexation sémantique de ces documents et un thésaurus du domaine pour l'ajout de liens ontologiques entre Topics. Pour enrichir la Topic Map, nous nous basons sur deux ontologies générales et nous proposons d'explorer toutes les questions potentielles relatives aux documents sources. Dans ACTOM, en plus des liens d'occurrences reliant un Topic à ses ressources, nous catégorisons les liens en deux catégories: (a) les liens ontologiques et (b) les liens d'usage. Nous proposons également d'étendre le modèle des Topic Maps défini par l'ISO en rajoutant aux caractéristiques d'un Topic des méta-propriétés servant à mesurer la pertinence des Topics plus précisément pour l'évaluation de la qualité et l'élagage dynamique de la Topic Map.

Abstract

In this thesis, we have proposed ACTOM, a Topic Map building approach based on an automated process taking into account multilingual documents and Topic Map evolution according to content and usage changes. To enrich the Topic Map, we are based on a domain thesaurus and we propose to explore all potential questions related to source documents on order to represent usage in the Topic Map. In our approach, we extend the Topic Map model that already exists by defining the usage links and a list of meta-properties associated to each Topic, these meta-properties are used in the Topic Map evolution process. In our approach ACTOM, we propose also to precise and enrich semantics of Topic Map links so, except occurrences links between Topics and resources, we classify Topic Map links in two different classes, those that we have called "ontological links" and those that we have named "usage links". We have defined an automated and evolutive pruning process to manage the Topic Map evolution and handle all possible changes related to content and usage of the Topic Map.