

Ontologie franco / anglaise du domaine informatique comme accès à un corpus de textes scientifiques

Gérald Kembellec

Equipe CITU-Paragraphe – Laboratoire Paragraphe
 Université Paris 8
 2, rue de la liberté
 93260 Saint-Denis
 gerald.kembellec@univ-paris8.fr
<http://geka.ei-10.eu>

Résumé :

Cet article présente une recherche visant à transformer un modèle d'organisation représentatif du domaine informatique en outil de recherche navigable intuitivement par un initié.

Mots-clés : Bibliothèque numérique, Ontologie de domaine, KBS.

1. Introduction

Dans le contexte universitaire français, il est courant d'observer des étudiants de 2^{ème} et 3^{ème} cycles éprouver de réelles difficultés à rassembler de la documentation sur leur domaine d'études ou de recherches. Dans notre optique, l'idée générale consiste à soutenir ces apprenants par l'outil, à compléter leur perception par de la connaissance de domaine « stockée » dans l'ontologie. Il est à envisager, à espérer même, qu'à terme la maîtrise de l'outil le rende obsolète, car source intrinsèque de connaissance. Dans notre contexte, il s'agit de recherche bibliographique dans le domaine informatique pour des personnes « initiées » mais non expertes, qui de plus sont perdues dans un corpus majoritairement anglophone.

Nous avons entamé notre recherche par une réflexion sur la pertinence de projeter le concept d'ontologie de domaine sur les notions de portail et de moteur de recherche. Cette démarche tend à générer une interface

homme/machine (IHM) afin d'améliorer l'approche utilisateur (le chercheur) de manière transparente et intuitive. La question que soulève cette réflexion est l'influence de la représentation de l'accès à l'information sur la recherche de connaissance. Son impact sur les résultats obtenus par rapport aux recherches plus traditionnelles est-il significatif ?

L'objectif final de notre travail de recherche est de mettre au point un système incrémental, idéalement autonome, d'indexation de documents scientifiques. Cette méthode s'appuie sur l'extraction des mots clés et le positionnement d'articles dans une ontologie de domaine informatique. L'autonomie du système serait un facteur non négligeable de réduction de coût, mais surtout un gain de temps. En effet cela éviterait à un groupe d'experts du domaine une veille technologique sans fin. Nous échapperions à la fatalité que chaque soubresaut technique ou idéologique provoquerait des débats sur l'opportunité d'indexation du nouveau concept à un emplacement donné de l'ontologie. Notre travail passe par la collecte des articles scientifiques relatifs à l'information que puis par une intégration à un corpus de documentation. Cette documentation sera représentée sous la forme d'une arborescence. Cela permettra l'émergence d'une visualisation globale du corpus de textes de la recherche informatique. Cette approche consentira également un accès facilité à l'information recherchée par moteur de recherche en langage naturel, par mots clés, contextualité, ou proximité sémantique. Ainsi, et c'est tout l'intérêt du concept, un utilisateur qui ne maîtrise pas encore l'ensemble du vocabulaire informatique (et la langue anglaise) pourrait trouver des articles pertinents en plusieurs langues, articles qu'il n'aurait pas su trouver seul par des méthodes de recherche traditionnelles.

Le présent article propose de mettre en œuvre la première partie de cette tâche, à savoir la construction de l'ontologie, le système de navigation pour la parcourir et un système de recherche dans un corpus scientifique.

2. Etat de l'art de la recherche d'informations par ontologies de domaine

Par ontologie de domaine nous entendons un ensemble de concepts hiérarchisés par un expert au sein d'une structure, et liés par des relations de proximité syntaxique ou sémantique.

La démarche classique d'utilisation des ontologies de domaine consiste à hiérarchiser les sous ensembles du domaine dans une optique de gestion. L'ontologie sert alors le plus souvent à hiérarchiser et classer les éléments composant le domaine ainsi qu'à décrire leurs relations. Une application courante est l'indexation de corpus spécialisés par ce biais.

Une utilisation plus novatrice de l'ontologie est d'inverser la démarche. Il est possible d'utiliser l'ontologie de domaine comme support de recherche dans un texte, un corpus, une bibliothèque numérique, ou même à l'extrême l'Internet.

Grâce à une combinaison de différentes technologies sémantiques, Stephan Bloehdorn a proposé une méthode intéressante de consultation de bibliothèques numériques [Bloehdorn *et al.* 2007]. Il a défini une approche par analyse de questions *structurées* en langage naturel avec une grammaire définie. Il s'agit pour le système de comprendre la question d'identifier les mots clés, les titres et les auteurs. Par exemple qui a écrit tel livre? Quel livre traite d'un sujet défini? Quel article fait partie de telle conférence et correspond à tels mots clés? Cette approche traduit le langage naturel en métadonnées, et reformule la question en langage SPARQL¹. Comme la réponse se trouve dans un fichier *Recherche Description Frameworke* (RDF²), les mises à jour en temps réel sont supportées, ainsi que l'hétérogénéité des formats et la location des ressources. Cette méthode permet de s'abstraire de toute base de données au sens commun du terme.

3. Ontologie de domaine informatique, conception d'un modèle exploitable

Au départ, l'approche sera onomasiologique ou *top-down*, c'est-à-dire que le corpus sera classé à la volée dans une structure qui est donc un ensemble normalisé et fini. Ensuite nous ambitionnons d'enrichir éventuellement la structure si les corpus ajoutés en font émerger le besoin. L'ontologie de domaine est composée d'une arborescence de sujets allant d'une racine générique : le domaine (ici l'informatique), vers des feuilles de connaissance. Les arcs seront des relations de spécification / généralisation ou des liens de similarité. L'ontologie ne contient pas les articles, mais des mots clés dont l'héritage se fait de manière

¹ <http://www.w3.org/TR/rdf-sparql-query/>

² <http://www.w3.org/RDF/>

top-down (descendante), et qui permettent de générer une requête qui sera passée aux principales bibliothèques scientifiques en ligne.

Cette arborescence constitue le squelette externe ou *exo-squelette* du domaine, dont les premiers mots clés sont les mots constituant les intitulés des nœuds et des feuilles. Ces mots clés seront dits « natifs », par opposition aux autres mots clés ajoutés à posteriori, qui seront dits « ajoutés ».

3.1. Notion de pertinence utilisateur

L'informatique est un domaine très vaste, comprenant une multitude de sous-disciplines, et un outil puissant usité dans de nombreux domaines scientifiques. Il faudra donc autant que possible s'imprégner de *points de vues* pour la recherche, saisir le contexte d'étude de l'utilisateur.

Exemple : le terme de « stockage de données » n'aura pas le même sens pour un technicien en assemblage, un ingénieur système ou un documentaliste. Pour le technicien la représentation qui s'impose du stockage de données est le disque dur ou la clé USB. Le professionnel système et réseau aura lui une vision plus large de « stockage de données ». Il verra les concepts de périphériques mais aussi les méthodes de stockages tels les NAS, les redondances de données (niveau de RAID), la façon dont les informations sont partagées (Netbios, NFS, SMB³ ...) mais aussi les droits sur les données (lecture, écriture et exécution). Enfin, le documentaliste verra en ce terme principalement un progiciel de SIGB (Système intégré de gestion de bibliothèque) qui gère les prêts, les réservations, suivi des commandes ou encore l'état des livres. Ces trois professionnels, pointus en leur domaine font un usage différent du terme « stockage de données », cependant on ne peut pas parler ici de polysémie, mais plutôt de point de vue.

La question de la pertinence utilisateur se pose dans ce cas précis de la SRI. Cette observation a grandement influencé l'outil, en l'axant sur l'utilisateur et pas uniquement sur les données. Ce projet doit être une entité à l'utilisation souple, se mettant à la portée de l'utilisateur pour l'aider à maîtriser son domaine de connaissance.

³ Divers protocoles de partage de données en réseau

3.2. Perspectives d'évolution de l'ontologie

Par la suite, lors de la phase d'indexation de corpus, si un article semble « inclassable », nous proposons momentanément de le classer au plus proche dans une des branches temporaires de l'ontologie, comme *miscellanés* ou *general*. Puis une fois une taille suffisante atteinte il conviendra de les classer définitivement en créant une ou des nouvelle(s) branche(s) à l'ontologie là où la proximité sémantique est la plus forte. Il s'agit d'un des vecteurs de l'évolution de l'ontologie, qui n'est pas statique mais évolue avec le corpus et le travail des usagers et experts.

Les extensions qui pourront être ajoutées à l'ontologie doivent être anticipées lors de son élaboration. Il doit être possible d'ajouter de nouveaux concepts sans avoir à toucher aux fondations de l'ontologie. Par exemple dans la branche ayant le plus de mots clés en commun ferait une racine convenable. Peut être même suffirait il de nommer la branche en résumant les concepts de l'article en un label en langage naturel.

4. Méthodes de recherches proposées et présomptions de modèles exploitables

Mots clés, Langage naturel, parcours de graphe de domaine, proximité sémantique. Nous sommes partis sur la définition du domaine de recherche avec un *exo-squelette* ontologique minimal. Cette partie du travail nécessite de trouver des approches taxonomiques représentant le plus finement et le plus exhaustivement possible le vaste domaine de l'informatique. Ensuite, pour conceptualiser ce domaine il va être nécessaire de segmenter les intitulés de chaque branche. Cette phase de spécification passe par une étape de construction des « grappes » de mots clés relatifs à chaque branche, grâce aux lemmes des mots extraits des intitulés.

D'un point de vue technique, pour une plus grande facilité de manipulation, nous intégrerons l'ontologie et ses mots clés dans une base de données, ce qui permettra de traduire de manière complète l'ontologie en *Extensible Markup Language* (XML⁴) en tenant compte de ses évolutions en temps réel.

Pour notre phase de test, le corpus de recherche sera composé des intitulés d'articles parus depuis 1945 et référencés dans la *DataBase systems and Logic*

⁴ <http://www.w3.org/XML/>

Programming (DBLP) par Michael Ley⁵ de l'Université allemande de Trier. Il s'agit à l'origine d'un document XML, d'environ un million d'entrées au format BibTEX⁶ (format de description bibliographique de LaTeX⁷). Notons que les papiers sont rédigés dans diverses langues. Nous proposerons également des méta-requêtes vers les bibliothèques Computer Science Bibliography (CSBIB)⁸ et Association for Computing Machinery (ACM).

5. Construction du modèle

5.1. Choix d'une référence de classification informatique

D'un point de vue technique, pour une plus grande facilité de manipulation, nous intégrerons l'ontologie et ses mots clés dans une base de données, ce qui permettra de traduire de manière complète l'ontologie en XML en tenant compte de ses évolutions en temps réel. Nous allons dans un premier temps chercher un organisme spécialiste des questions informatiques proposant un système de représentation du domaine que nous ambitionnons de modéliser. Pour des raisons de simplicité, l'encyclopédie en ligne Wikipédia se détache dans un premier temps. En effet le domaine informatique y est classé selon une hiérarchie interne et un corpus abondamment pourvu est immédiatement disponible en XML et RDF. Cependant, à l'heure actuelle la caution scientifique de Wikipédia n'est pas démontrable. Nous allons donc nous concentrer sur la Computing Classification System⁹, dont la légitimité n'est plus à prouver. De plus, pour ne rien gâcher l'ACM possède sa propre bibliothèque numérique d'articles scientifiques, eux-mêmes indexés selon le modèle Computing Classification System (CCS).

Dans le contexte, le CCS n'est pas exploitable en l'état. Le CCS semble *a priori* plus être plus une taxonomie qu'une ontologie. Dans une taxonomie, le vocabulaire est organisé sous une forme hiérarchique. Cette hiérarchisation

correspond souvent à une spécification. Une taxonomie est une forme d'ontologie dont la grammaire n'a pas été formalisée. Dans le CCS cette grammaire a été définie par des rapports clairs de descendance et d'ascendance mais aussi des liens transversaux de proximité sémantique. Ainsi on peut un lien de ce type entre B.8 *Performance and reliability* et C.4 *Performance of systems* (cf. Figure 1).

Cependant, d'après Gruber, un des aspects importants d'une ontologie (en sus de la clarté, de la cohérence, d'un engagement minimal, et de la déformation) est l'extensibilité [Gruber 1993]. Il convient donc d'effectuer un traitement pour permettre d'anticiper les évolutions de l'ontologie. En effet le système d'identifiant du CCS ne s'applique qu'aux nœuds et pas aux feuilles. Cela empêche de conserver l'esprit de référencement si pratique proposé par l'ACM en cas de spécialisation d'une feuille. On ne peut pas imaginer une relation de spécification étendant un élément non référencé. C'est pourquoi nous choisissons de donner de manière arbitraire un identifiant aux feuilles pour les transformer en nœuds potentiels. Par respect pour le travail initial et pour distinguer nos évolutions du travail initial, nous avons choisi de donner comme identifiant des feuilles l'identifiant du père auquel s'ajoutera une lettre de l'alphabet. Notons que la notation CCS utilisant déjà le « m » pour *divers/miscellaneous* et le « g » pour *general/general*, nous avons ôté ces deux lettres de notre processus d'identification des nœuds et feuilles.

⁵ <http://dblp.uni-trier.de/>

⁶ <http://www.bibtex.org/>

⁷ <http://www.latex-project.org/>

⁸ <http://iinwww.ira.uka.de/bibliography/index.html>

⁹ CCS de l'ACM <http://www.acm.org/class/1998/>



Top Two Levels of The ACM Computing Classification System (1998)

- **A. General Literature**
 - A.0 GENERAL
 - A.1 INTRODUCTORY AND SURVEY
 - A.2 REFERENCE (e.g., dictionaries, encyclopedias, glossaries)
 - A.m MISCELLANEOUS
- **B. Hardware**
 - B.0 GENERAL
 - B.1 CONTROL STRUCTURES AND MICROPROGRAMMING (C.3.2)
 - B.2 ARITHMETIC AND LOGIC STRUCTURES
 - B.3 MEMORY STRUCTURES
 - B.4 INPUT/OUTPUT AND DATA COMMUNICATIONS
 - B.5 REGISTER-TRANSFER-LEVEL IMPLEMENTATION
 - B.6 LOGIC DESIGN
 - B.7 INTEGRATED CIRCUITS
 - B.8 PERFORMANCE AND RELIABILITY **new!** (C.4)
 - B.m MISCELLANEOUS
- **C. Computer Systems Organization**
 - C.0 GENERAL
 - C.1 PROCESSOR ARCHITECTURES
 - C.2 COMPUTER-COMMUNICATION NETWORKS
 - C.3 SPECIAL-PURPOSE AND APPLICATION-BASED SYSTEMS (J.7)
 - C.4 PERFORMANCE OF SYSTEMS
 - C.5 COMPUTER SYSTEM IMPLEMENTATION
 - C.m MISCELLANEOUS
- **D. Software**
 - D.0 GENERAL
 - D.1 PROGRAMMING TECHNIQUES (E)
 - D.2 SOFTWARE ENGINEERING (K.6.3)
 - D.3 PROGRAMMING LANGUAGES
 - D.4 OPERATING SYSTEMS (G)
 - D.m MISCELLANEOUS
- **E. Data**

Figure 9 : CSS d'ACM

5.2. Travail sur l'ontologie en franais

Notons que l'ontologie devra  tre enti rement traduite en franais, mais que les mots anglais ne devront pas  tre lemmatis s. Il est donc n cessaire de faire le distinguo de la langue pour les intitul s. Par exemple NFS, l'acronyme de « Network File System » n'a pas d' quivalent franais, le concept sera traduit dans la version franaise par « Syst me de fichiers en r seau » pour donner un aperu du concept. Cependant, l'emploi de l'acronyme sera   n'en pas douter utilis  dans les articles. L'informatique  tant une science majoritairement anglophone, les articles franais comporteront de toute faon par d faut des mots franais et anglais. Il serait d'ailleurs judicieux de proposer des relations de synonymie entre

des mots  tant indiff remment employ s en anglais ou en franais par les professionnels, les enseignants et les chercheurs sp cialis s dans le domaine. Cependant en toute objectivit , ce travail entraine n cessairement une sp cification de la conceptualisation du domaine. Il y aura in vitablement des inexactitudes, qui seront corrig es *a posteriori*.

5.3. Traduction de l'ontologie en franais

D'apr s la lettre ouverte   l'Agence d' valuation de la recherche et de l'enseignement sup rieur (AERES) par quelques milliers de chercheurs franais, il est largement admis que la *lingua franca* de la recherche scientifique est aujourd'hui l'anglais. Pourquoi traduire les intitul s des branches de l'ontologie en franais alors que le corpus est majoritairement anglais, la langue scientifique? Nous attirons l'attention sur le fait que si l'utilisateur final ma trise peut  tre la lecture de textes techniques et scientifiques, il peut se sentir plus   l'aise en franais pour effectuer sa recherche, quitte   lire les articles en anglais avec un bon dictionnaire sous la main.

Le choix le plus simple et le plus  conomique pour automatiser une traduction anglo-franaise est l'utilisation d'un outil de traduction en ligne. Les outils qui ont attir  notre attention ont  t  Babelfish de Yahoo et Google translate de la suite Google. Nous avons conu et utilis  une Interface de Programmation Applicative (API) de *wrapping*¹⁰ pour g n rer une version franaise de l'ontologie bas e sur un de ces outils. Notons au passage que ce type d'application en ligne gagnerait   poss der sa propre API.

Une fois cette  tape termin e, nous avons rapidement compris que rien ne remplace une traduction manuelle, c'est pourquoi nous int grons une notion de *folksonomy* par flux RDF Site Summary¹¹ (RSS) dans l'outil, cela permettra   l'utilisateur final de signaler une erreur de traduction, ou une impr cision, au comit  de gestion. Ce groupe sera form  des chercheurs des laboratoires de l'unit  de recherche et de formation et validera ou non la proposition. Selon Thomas Vander Wal, la valeur du marquage ext rieur de la *folksonomy* vient des usagers en utilisant leurs propres mots ce qui ajoute une dimension explicite, qui va  tre une inf rence de l'objet [Vander Wal 2006]. Le syst me de nommage

¹⁰ Technique qui consiste   cr er un programme informatique permettant   deux autres programmes de communiquer.

¹¹ <http://web.resource.org/rss/1.0/>

français des nœuds de l'ontologie automatisé dans un premier temps, poursuivi et développé par les utilisateurs anglophones sera validé par des experts au besoin sans avoir eu recours à un traducteur professionnel, ni mobilisé un expert à plein temps. Cette procédure permet un évident gain de temps pour les chercheurs du groupe et une économie financière non négligeable.

L'aspect technique de cette démarche devra être simplifié au maximum pour l'utilisateur afin de ne pas le décourager de faire une proposition. L'opération ne doit également pas lui prendre plus de quelques secondes. Une fois la proposition faite, un flux RSS est généré et restera actif jusqu'à vérification par au moins deux membres du comité. Nous ambitionnons ainsi de corriger la partie française de l'ontologie sur une période de temps encore indéterminée.

Le procédé permet aussi de tenir compte des mutations terminologiques inhérentes aux évolutions du domaine *Information Technology* (IT). L'utilisateur final bénéficie grâce à son interaction avec le système d'un enrichissement de sa connaissance du pôle de connaissances tout en participant à son évolution.

5.4. La génération des mots clés et l'émergence de proximité sémantique

Considérons le corpus formé par les intitulés composant l'ontologie de domaine IT du point de vue de l'infométrie statistique. Selon Le Coadic [Le Coadic 2006], si l'on considère un ensemble d'articles scientifiques, il faut s'intéresser aux mots significatifs et à leur cooccurrence pour dégager des proximités sémantiques significatives. Ainsi lorsqu'un *n-uplet* de mots associés apparaît simultanément dans plusieurs intitulés de nœuds, il est probable que les sujets traités soient associés. Bien sûr, dans le cas précis nous n'utiliserons cette approche que sur des intitulés, mais gageons que les labels ACM sont suffisamment précis pour être représentatifs de l'ensemble des articles, tant du point de vue général, que particulier.

Ainsi, les mots les plus représentatifs du label seront ajoutés comme mots clés de l'article et de la branche, les autres notés dans la proximité sémantique. Ultrieurement, lors de la phase d'indexation d'une bibliothèque numérique, si un article semble pouvoir être indexé à deux endroits, il sera proposé de créer un lien de proximité entre deux branches de l'ontologie.

5.5. L'interface avec le corpus

Nous tenterons à terme une méthode compilation (ou *clustering*) de différentes bases d'articles en ligne comme CSBIB, DBLP, ACM entre autres... Le *clustering* passera par une phase de prétraitement. Chaque bibliothèque scientifique possédant sa propre interface d'interrogation, nous allons essayer de trouver le document RDF qui décrit chaque base. Notons que si chaque site de ce type fournissait des services de description de données comme RDFa¹², cette démarche serait grandement simplifiée.

Une base de données décrivant « la bibliothèque scientifique du domaine informatique » sera constituée, décrivant chaque article par son titre, son contexte de publication, son année et ses auteurs. Cette base de données, mise à jour automatiquement chaque semaine de manière incrémentale, permettrait idéalement de générer à la volée un document unique RDF décrivant le pseudo corpus. C'est sur ce document que s'appuieront les recherches internes à l'outil.

Le terme à la volée signifie qu'en théorie pour chaque requête, un cliché du corpus sera constitué en RDF par interrogation de la base et traité pour tenir compte des mises à jour hebdomadaires. Cette démarche, bien que souhaitable est techniquement utopique. Il est tout de même possible, voire souhaitable au vu de la masse de données (dans une optique d'économie de ressource système) de conserver un cliché en cache. Ce cliché de la base deviendrait fichier RDF et serait alors la représentation du pseudo corpus.

Le corpus d'articles scientifiques n'est pas hébergé localement sur la machine hôte de l'ontologie pour des raisons légales, mais également de capacité de stockage. C'est pour cette raison que nous préférons dans le contexte utiliser le terme de pseudo-corpus plutôt que corpus. En effet les labels, et éventuellement les résumés indexés dans les bibliothèques numériques ne constituent pas à proprement parler un corpus.

5.6. Choix d'un type de représentation

Pour rendre le corpus plus accessible, nous optons de faciliter la représentation de l'ontologie de domaine sous la forme d'une carte navigable. L'arborescence doit permettre un focus sur la branche contenant une formalisation du concept recherché.

¹² <http://www.w3.org/TR/xhtml1-rdfa-primer/>

Il existe un certain nombre de manières de visualiser les ontologies, mais toutes ne sont pas propres à la navigation, en tous cas pas à une navigation intuitive. Dans notre contexte, l'outil de représentation doit se conformer à un certain nombre de règles exposées par Christophe Tricot et Christophe Roche [Tricot et al. 2007] suite à un certain nombre d'observations. A minima, pour être efficace le système de visualisation doit respecter les règles suivantes :

- Offrir une vue globale de l'ontologie. Cela permettra à l'utilisateur d'identifier tous les concepts du domaine.
- Utiliser une approche "focus + contexte" pour permettre à l'utilisateur de se concentrer sur certains éléments tout en ayant accès aux autres;
- Utiliser la géométrie plane, pour éviter de déranger la perception naturelle de manipulation dans le plan. Ce point en particulier, ne sera pas respecté, car au vu de la masse de données à afficher et de la volonté de respecter les points précédents, il est complexe, voire impossible de combiner un affichage en arborescence et la géométrie euclidienne.

Du retour d'expérience de C. Tricot, nous noterons également qu'il émerge deux types d'utilisateurs: "novices" et "experts". Les novices comprennent le domaine et ses concepts sans pour autant saisir la finesse de l'organisation et les interactions. Les experts quant à eux saisissent parfaitement la globalité du domaine tant du point de vue des concepts que des rapports qui les lient. Dans notre contexte d'utilisation, les utilisateurs ont un profil qui peut être hybride pour un étudiant en MASTER ou un docteur qui se renseigne sur un sujet transversal à ses travaux, jusqu'à un profil « expert » pour un spécialiste de domaine. Nous essayerons donc de trouver un compromis de représentation du domaine offrant des accès directs au contexte sur l'élément en focus. Dans l'article de C. Tricot, il semble que le modèle de représentation par *radial tree* soit le plus approprié pour des experts et l'*eye tree* pour des novices.

La visualisation en *eye tree* permet une vision globale du domaine ainsi que la possibilité d'un grand angle focalisé (*fisheye polar*) sur un point de détail autour duquel s'articule le domaine. Son principal défaut, dans le contexte, est de se borner au plan ce qui empêche la mise en perspective autorisée par les *cones trees*. Le *radial tree* est assez similaire à l'*eye tree* combinant la vision globale du domaine et le *fisheye polar*. Cependant une plus grande place est faite au contexte et au focus au sein même du graphe. Il semble que ce qui fait l'intérêt du *radial tree* (le focus + contexte) cause également une perte de contact avec l'objectif premier qui est de

conserver la vue globale. De plus, un *radial tree* décrivant l'ACM serait parfaitement illisible du fait même de la taille de l'ontologie.

Au vu de ladite taille, une visualisation par «grappe d'informations» émerge grâce à la combinaison d'ontologie et de la technologie dite *Topic Mapper*, grâce à l'applet open source hypergraph¹³. Bien que n'étant pas spécialement préconisé pour représenter efficacement une ontologie, le *Topic Mapper* est une représentation de type *hyperbolic tree* qui consiste à cartographier l'ontologie et d'y naviguer à volonté. Nous allons l'adapter pour faire émerger des points de vues, mais aussi des focus avec leurs contextes. Il s'agira ainsi d'une approche hybride entre l'*eye tree* et l'*hyperbolic tree*.

5.7. Génération d'une méta-requête naviguée

La première étape de génération de requête est le filtrage du «bruit» sur le label grâce à des *stop-words* (une dans chaque langue) qui va éliminer les mots vides comme les articles, les pronoms, ainsi que les substantifs trop communs pour avoir un sens significatif lors du positionnement de l'utilisateur dans le navigateur de l'ontologie. Une étape similaire préalable est effectuée lors de l'utilisateur du moteur de recherche en langue naturelle.

La deuxième étape consiste à une lemmatisation des mots français, puis un calcul de proximité statistique de l'ensemble des mots dégagés avec la grappe de mots clés d'une des branches de l'ontologie. Peut être est-il judicieux de donner une valeur aux mots clés dans le contexte (arc value)? Ce point sera une perspective à approfondir.

¹³ <http://sourceforge.net/projects/hypergraph/>

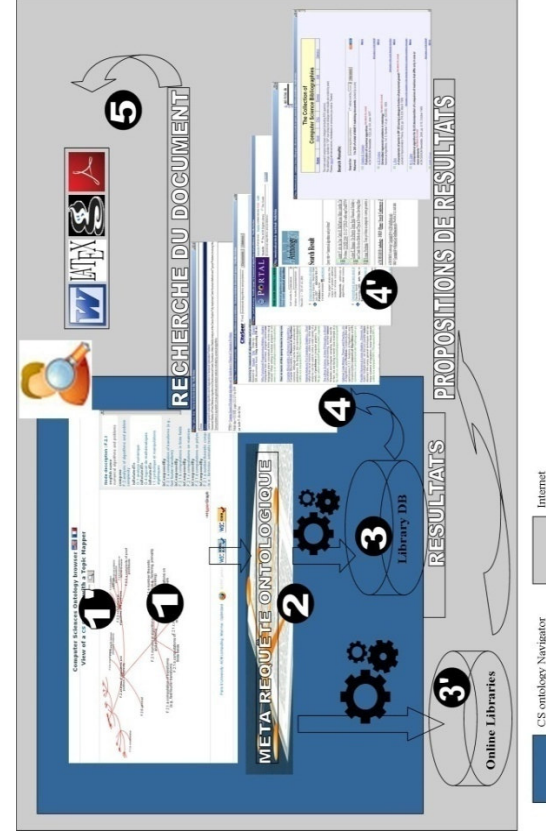


Figure 10 : mod elisation conceptuelle

Description r ecapitulative:

- 1 Et 1' : Possibilit  de se positionner dans l'ontologie par navigation ou par une requ te en langage naturel.
- 2 Le positionnement permet de cerner un point de vue utilisateur et des centres d'int rets,
- 3 Et 3' : ce qui va d gager des m ta-donn es et constituer des requ tes vers le RDF interne ou les biblioth ques num riques en ligne.
- 4 Et 4' : Les intitul s des articles correspondant   la requ te et trouv s dans le RDF ou les bases de connaissances scientifiques sont propos s.
- 5 Les articles sont cherch s sur le net si via Google scholar si l'Uniform Resource Identifier¹⁴ (URI) est absente de la base, ou directement propos s sur les

¹⁴ <http://www.w3.org/2004/11/uri-iri-presentation.html>

biblioth ques num riques. Si l'on utilise les biblioth ques num riques, l'acc s aux documents est direct.

6. Essais de recherche navigu e

6.1. Navigation dans l'ontologie

La premi re  tape de la recherche par navigation consiste   descendre dans l'arborescence jusqu'au n ud le plus repr sentatif du concept recherch .

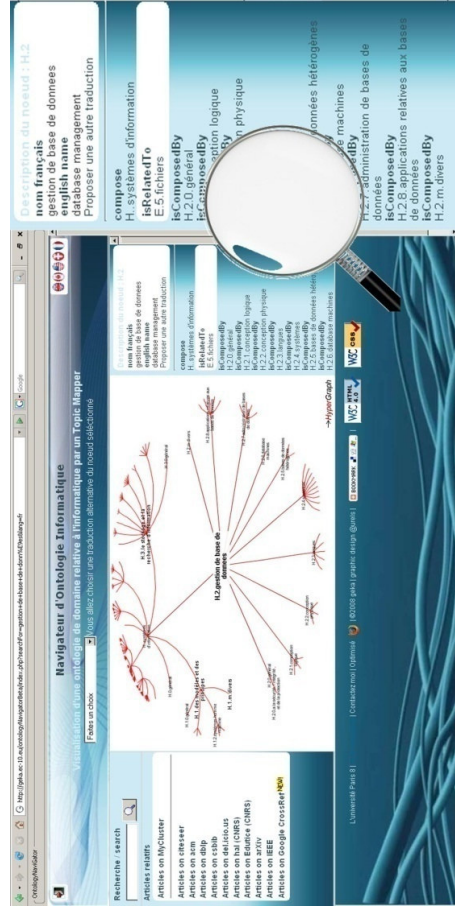


Figure 11: Recherche d'articles scientifiques par navigation de l'ontologie et zoom sur le contexte en focus.

Ici, la d marche de navigation pour atteindre le n ud «Gestion de base de donn es» a  t  de passer par la racine «Informatique» puis par « Syst mes d'information » et enfin de s lectionner n ud « Gestion de base de donn es ».

6.2. Exemple de recherche contextuelle d'articles

The image shows a workflow for finding scientific articles. It starts with a Google Scholar search for "Managing Document Taxonomies in Relational Databases". The search results show a book by Ido Millet, published in 2005. A red magnifying glass highlights the book title. An arrow points to a preview of the book's abstract, which discusses the challenges of applying relational database technologies to document management. The abstract text is: "This chapter addresses the challenge of applying relational database technologies to document management. It first describes how normalizing the data model within a relational database can be used to address the data management challenges of document management. It then discusses how document management can be used to address the challenges of document management." The preview also shows the book's title, author, and publisher information.

Figure 12: Articles scientifiques proposés par le système

Le bloc rouge du contexte (cf. figure 3) propose un accès direct aux articles des bibliothèques numériques en lignes comme CSBIB, DBLP, ou ACM en

générant des requêtes contextuelles vers ces sites (cf. figure 2). Mais l'outil propose également d'interroger la base interne d'intitulés d'articles. Dans l'exemple (cf. figure 4) une recherche sur « gestion de base de données » est générée et propose quelques dizaines de résultats. Nous choisissons « *Managing document taxonomies in relational databases* », la base nous donne l'auteur principal. L'outil vérifie la présence d'une URI relative à l'article dans la base, et en l'absence de celle-ci une requête est générée vers Google Scholar qui nous offre un accès direct à l'article.

Des tests ont été effectués vers les bases classiques, mais les résultats portant exactement sur le contexte de recherche sont encore trop rares. Le mécanisme de génération de requêtes est encore au stade d'heuristique, mais de bonnes perspectives sont ouvertes.

7. Limites et perspectives

Comme l'essai de l'outil l'a montré, la justesse actuelle des méta-requêtes générées est critiquable et de fait les résultats sont parfois approximatifs sur les bases externes. Il devrait cependant apparaître que, plus l'ontologie s'étofferait d'articles, plus la recherche et l'indexation seront précises. A cette fin, nous tenterons d'indexer un corpus préexistant de taille respectable. A cette fin, nous nous fixons comme ambition de créer un script d'extraction incrémental de contenu sur la bibliothèque en ligne DBLP. Ce travail en cours d'automatisation devrait affiner la pertinence d'indexation et de recherche par ontologie.

Une autre limite est l'accès physique aux articles qui est souvent soumis à un abonnement payant, quand ce n'est pas un paiement à l'unité. C'est pourquoi cette solution trouvera plus facilement une place dans les locaux d'un laboratoire universitaire ou une bibliothèque. Cependant, l'utilisation de proxy devrait permettre d'étendre l'accès aux bibliothèques numériques à tout un campus. L'outil va être mis à disposition des étudiants de second cycle du département informatique de l'Université Paris 8, au centre de calcul. Nous proposerons un formulaire en ligne pour enregistrer les retours en suivant les parcours utilisateurs.

Dans un avenir proche nous envisageons d'étendre l'application avec une ontologie basée sur les personnes au format *friend of a friend* (FOaF)¹⁵ pour mieux cerner les groupes de travail, équipe, et laboratoires ainsi que les liens de

¹⁵ <http://xmlns.com/foaf/spec/>

transversalit  disciplinaire. Un autre objectif que nous ambitionnons est de rendre le syst me le plus autonome possible.  ventuellement le syst me de navigation de type *hyperbolic tree / eye tree* sera d laiss  si un autre type de visualisation plus navigable ou ergonomique  merge.

8. Conclusion

Au cours de ce travail nous avons tent  de r aliser un outil de recherche pour les chercheurs dont les travaux sont li s   l'informatique. Cette interface consultable en ligne permet de lier un contexte de recherche ontologique   des biblioth ques scientifiques en ligne. Cette ontologie bas e sur la CCS de l'ACM a  t  traduite en fran ais de mani re automatique pour proposer aux chercheurs francophones une alternative en langue maternelle. Notre solution propose aux chercheurs de trouver des articles relatifs   un contexte d' tude bas  gravitant autour d'un n ud du domaine ontologique IT. Cette requ te est g n r e par navigation graphique du domaine ou langage naturel. Une fois le contexte de recherche d gag , un travail automatis  permet de trouver des articles en relation dans la base de donn es interne ou de proposer des m ta-requ tes vers les biblioth ques num riques scientifiques en ligne. Le syst me de g n ration de m ta-recherche  tant uniquement bas  sur les intitul s des n uds de l'ontologie, ce projet en l' tat a rapidement montr  ses limites. Cependant les r sultats sont encourageants, des perspectives d'am lioration ont donc  t  envisag es dans un futur proche ou sont d j    l' tude.

Bibliographie

- Bloehdorn S., Cimiano P., Duke A., Haase P., Heizmann J., Thurlow I., V lker J., *Ontology-based Question Answering for Digital Libraries*. EC2DL, *Lecture Notes in Computer Science*, Vol. 4675, pp. 14-25, 2007.
- Dragan G., Marek H., *Searching Web Resources Using Ontology Mappings*. *Integrating Ontologies*, CEUR Workshop Proceedings, Vol. 156, CEUR-WS.org, 2005.
- Griber T. R., *Toward Principles for the Design of Ontologies Used for Knowledge Sharing*, 1993.
- Le Coadic, Y.F., *Math matique et statistique en science de l'information et en science de la communication*. IBICT, 2006.

Trioot C., Roche C., *Visualisation of Ontology: a focus and context approach*, InSciT2006, 2006.

Vander Wal T., *Understanding Folksonomy (Tagging that Works)*, dConstruct, 2006.