Ontologies pour l'aide à l'exploration d'une collection de documents.

Nathalie HERNANDEZ (*,**), **Josiane MOTHE** (*,**)
hernandez@irit.fr , mothe@irit.fr

(*)IRIT,
118 route de Narbonne
31062 Toulouse-Cedex 4, France

(**)ERT34 - IUFM,
56 av. de l'URSS,
31078 Toulouse Cedex 4, France

Mots clefs :
ontologie, veille, exploration, visualisation d'ontologies
Keywords:
ontology, technological watch, exploration, ontology visualisation
Palabras clave :
Escudriñar científico y tecnológico, ontologoas, visualisacion de ontologoas

Résumé

Dans cet article nous présentons un système à base d'ontologies pour l'aide à une activité de recherche, d'analyse et d'exploration de corpus documentaires relatifs à un domaine scientifique. Les ontologies utilisées visent à représenter un domaine à la fois à travers le vocabulaire de ce domaine, mais également au travers de l'ensemble des méta-données qui peuvent être utiles dans des activités d'accès à l'information et de veille. Le modèle de représentation des documents repose donc sur deux ontologies différentes : une ontologie du domaine de la veille (liée à la tâche à réaliser) et une ontologie du domaine abordé dans le corpus. Le cadre applicatif que nous avons choisi est l'astronomie. Nous présentons donc des échantillons de l'ontologie du domaine de la veille et de l'ontologie du domaine de l'astronomie validées par des astronomes. Le prototype de l'interface de ce système est également présenté. Cette interface permet à un utilisateur de parcourir les données et d'explorer la collection de documents au travers de la représentation de la connaissance liée à la tâche de veille et au domaine de l'astronomie.

1 Introduction

Les systèmes de recherche d'information visent à restituer (tous) les documents pertinents (et seulement ceux là) par rapport à un besoin d'information exprimé par un utilisateur. Les systèmes d'exploration visent quant à eux à fournir de l'information élaborée à partir de l'analyse d'un ensemble d'information relative à un thème. Les index jouent un rôle primordial dans ces deux types de systèmes en définissant les descripteurs (mots ou groupements de mots) qui représentent le contenu des documents et à partir desquels les documents peuvent être accédés ou analysés. L'indexation des documents n'est pas le seul moyen utilisé pour structurer et organiser une collection. Un modèle de structuration complémentaire consiste à utiliser les méta-données explicitement associées aux documents. Ces méta-données sont des données factuelles qui contiennent de l'information sur l'information comme par exemple le nom des auteurs, la date de publication, les mots clés choisis par les auteurs… Ces méta-données peuvent être combinées aux descripteurs issus de l'indexation pour fournir une représentation plus complète des documents de la collection. Cette représentation basée sur les méta-données et les descripteurs correspond à de la connaissance relative au corpus.

Quel que soit le mode de représentation interne des informations par le système, un des problèmes auquel l'utilisateur doit faire face est d'exprimer son besoin en information. Généralement quand un utilisateur sait ce que contient la collection, comment elle est structurée, ce qu'il recherche et comment il peut le décrire, il n'a pas de problème pour formuler sa requête. En fournissant les méta-données, le système peut aider l'utilisateur à évaluer le potentiel de la collection, cerner ses besoins et définir ses requêtes. Se pose cependant le problème de choisir quelles méta-données fournir à l'utilisateur et quelle organisation choisir pour représenter cette connaissance. Les ontologies semblent être la solution la plus adaptée à cette problématique.

Le modèle que nous proposons se base donc sur des ontologies pour l'aide à une activité de recherche ou d'exploration d'un domaine. Ces ontologies permettent la représentation d'un domaine à la fois à travers son vocabulaire spécifique, mais également au travers de l'ensemble des méta-données qui sont utiles dans de telles activités. Dans la partie 2, nous présenterons un état de l'art sur l'utilisation des ontologies en Recherche d'Information (RI) et en veille. Nous définirons ensuite les objectifs auxquels nous souhaitons répondre ainsi que le cadre applicatif que nous choisissons (partie 3). Nous détaillerons dans la partie 4 notre solution basée sur la définition d'une ontologie du domaine de la veille et d'une ontologie du domaine abordée dans le corpus. Dans une dernière partie, nous présenterons l'interface du système permettant l'exploration d'une collection documentaire au travers de la navigation de ces deux ontologies.

2 Ontologies pour la recherche d'information et la veille

2.1 Définitions

Le terme « ontologie » a été emprunté au domaine de la philosophie dans lequel il signifie « l'essence de l'essentiel ». Dans le domaine de la gestion de connaissance, le sens de ce mot est différent. Gruber [11] introduit la notion d'ontologie comme "une spécification explicite d'une conceptualisation". Cette définition a été légèrement modifiée par Borst [4]. Une combinaison des deux définitions peut être résumée ainsi : « une spécification explicite et formelle d'une conceptualisation partagée ». Cette définition s'explique ainsi [23] : *explicite* signifie que le « type des concepts et les contraintes sur leurs utilisations sont explicitement définies », *formelle* se réfère au fait que la spécification doit être lisible par une machine, *partagée* se rapporte à la notion selon laquelle une ontologie « capture la connaissance consensuelle, qui n'est pas propre à un individu mais validée par un groupe », *conceptualisation* se réfère à « un modèle abstrait d'un certain phénomène du monde basé sur l'identification des concepts pertinents de ce phénomène ».

Une ontologie fournit une base solide pour la communication entre les machines mais aussi entre humains et machines en définissant le sens des objets tout d'abord à travers les symboles (mots ou expressions) qui les désignent et les caractérisent et ensuite à travers une représentation structurée ou formelle de leur rôle dans le domaine. Différents niveaux sémiotiques sont à considérés dans une

ontologie [5]. Le *niveau lexical* recouvre l'ensemble des termes utilisés pour transcrire le sens des concepts. Le *niveau conceptuel* représente les concepts et les relations conceptuelles qui les relient.
De nombreux types de structure de connaissance se cachent derrière le mot ontologie (taxonomie, thesaurus,…). Ces structures de données peuvent être à la fois terminologiques (elles contiennent un ensemble de termes) et conceptuelles (elles définissent généralement des concepts). Cependant, ces structures peuvent différer par leur contenu (connaissances générales : WordNet [http://www.cogsci.princeton.edu/~wn/], Cycl, connaissances d'un domaine : MeSH [http://www.nlm.nih.gov/mesh/MBrowser.html],…), par le type des relations sémantiques entre les concepts (relations taxonomiques, méronymiques, …), et par le niveau de formalisation (représentation logique, représentation dans un langage dédié aux ontologies tels que DAML+OIL, OWL …).

2.2. Utilisation des ontologies en RI

De plus en plus de travaux portent sur la formulation de requêtes et l'indexation des contenus de documents par des mots clés à l'aide d'une ontologie.
Une telle approche sémantique semble très prometteuse dans l'optique d'améliorer les performances des systèmes de RI [15]. Ainsi, une ontologie doit permettre d'affiner les réponses d'un système en augmentant les chances de formuler une requête à partir des termes ou descripteurs représentant au mieux le besoin en information.
L'expansion de requête, utilisée pour limiter le silence (le silence fait référence aux documents pertinents mais qui ne sont pas retrouvés par le système) , peut reposer sur des liens entre les concepts présents dans la requête et ceux présents dans l'ontologie. De façon complémentaire, les termes de la requête peuvent être désambiguïsés en se basant sur les définitions et les liens entre concepts présents dans l'ontologie, réduisant ainsi les risques de bruit (le bruit fait référence aux documents non pertinents retrouvés par le système). Dans OntoBroker et $(Ka)^2$ [3] [9] par exemple, les pages web sont annotées manuellement par des concepts d'une ontologie. Pour une requête donnée, tous les concepts liés aux termes de la requête sont inférés et ajoutés à la requête.
Une ontologie peut également servir à l'indexation des documents. Dans ce cas, les descripteurs ne sont plus choisis directement dans les documents mais au sein même de l'ontologie. Les textes sont alors indexés par des concepts qui reflètent leur sens plutôt que par des mots bien souvent ambiguës. Il convient dans ce cas d'utiliser une ontologie reflétant le ou les domaines de connaissance abordés dans la collection documentaire. Il est en effet nécessaire de retrouver dans l'ontologie les concepts présents dans la collection pour indexer les documents à partir de toutes les thématiques abordées.
L'indexation du contenu des documents à partir d'une ontologie présente de plus les avantages suivants :
- Aider l'utilisateur à formuler sa requête. En présentant l'ontologie à l'utilisateur, il est possible de le guider dans le choix des termes de sa requête. Plusieurs interfaces de visualisation et d'exploration d'ontologie ont été proposées comme KAON [6], Graphlet [13], IRAIA [17] .
- Faciliter la RI au sein de collections hétérogènes en indexant tous types de document à partir des mêmes concepts.
Dans le contexte de la RI, une ontologie n'est généralement pas représentée logiquement. Le formalisme utilisé sert habituellement à faciliter la gestion des concepts en tant qu'objets, leur classification, la comparaison de leurs propriétés et la navigation au sein de l'ontologie en accédant à un concept et à ceux qui lui sont reliés. Son utilisation est donc réduite à l'accès à la connaissance pour permettre une meilleure indexation ou stockage des informations et faciliter la recherche. Nous verrons dans la section 3 que les besoins en information peuvent être plus complexes et en particulier, nécessiter l'élaboration d'information à partir des informations stockées. Les ontologies représentent la connaissance et permettent à partir de règles d'inférences de déduire de nouvelles connaissances. Nous proposons donc d'exploiter également cette caractéristique des ontologies.

2.3. Utilisation des ontologies pour la veille

Les activités d'analyse prennent en compte les descripteurs issus des documents ainsi que leurs meta-données. De nombreux travaux visent à intégrer de la connaissance dans le processus de veille pour

exploiter ces données. Cette connaissance peut être représentée sous forme de ressources de plusieurs natures : ressources terminologiques (vocabulaire contrôlé, thésaurus) ou hiérarchies de concepts.
Dans [21] des ressources terminologiques issues de la microbiologie et de la génomique sont utilisées pour améliorer la fouille des données textuelles du domaine de la biologie moléculaire. Les ressources terminologiques utilisées sont formées des termes caractéristiques du domaine. L'indexation des documents se fait à partir de ce vocabulaires contrôlé. Cette indexation a pour intérêt d'améliorer la classification et la cartographie des données par rapport à un système n'utilisant pas de lexique et basant l'extraction des index sur la recherche de patrons.
Un autre type d'approche consiste à indexer les documents à partir de hiérarchies de concepts. Une hiérarchie de concepts est formée des concepts d'un domaine organisés du plus générique ou plus spécifique. Un concept peut être défini à partir d'un ou plusieurs termes. Cette approche est suivie dans [10]. Les documents sont annotés à partir d'une hiérarchie représentant les problèmes rencontrés par un utilisateur de télécommunication . Ces annotations permettent ensuite d'organiser les données et de les catégoriser en fonction des problèmes rencontrés. Cette approche est également utilisée dans [1]. Plusieurs hiérarchies sont alors utilisées pour indexer des documents suivant les différentes facettes du domaine abordé dans un corpus. Dans le cadre de ces travaux, des documents portant sur l'économie sont ainsi catégorisés suivant trois hiérarchies de concepts relatives aux régions, aux industries et aux indices économiques. Dans [2] les ontologies servent de support à la construction de hiérarchies de concepts. Chaque document de la collection est associé à un ensemble de concepts d'une ou plusieurs hiérarchies. Ces hiérarchies sont ensuite utilisées pour l'exploration graphique d'une collection.
Dans la majorité des cas, la connaissance est représentée par des ressources terminologiques. Dans le cas ou la connaissance est représentée à un niveau conceptuel (hiérarchie de concepts), la seule relation utilisée pour organiser les concepts est la relation de spécificité/généricité. Les travaux présentés dans cet article repose sur une ontologie dans laquelle les niveaux lexical et conceptuel sont élaborés (concepts définis à partir de plusieurs termes, relations de plusieurs types). Ce type de représentation de connaissance a pour intérêt de permettre un niveau d'analyse plus fin à partir des relations entre les concepts sémantiquement plus riches. Les relations sont également formalisées. Des inférences peuvent alors permettre d'extraire de nouvelles connaissances. Par exemple, si la relation transitive R lie les concepts A et B puis les concepts B et C, une inférence permettra d'établir que A et C sont également reliés par R.

3. Objectifs du système

Les travaux présentés dans cet article, visent à proposer un système à base d'ontologies pour l'aide à une activité de recherche, d'analyse et d'exploration de corpus documentaires relatifs à un domaine scientifique ; les ontologies jouant un rôle central dans la gestion des informations. L'objectif du système OntoExplo est d'offrir plusieurs fonctionnalités.
Une première fonction du système consiste à *indexer les documents* pour permettre à un utilisateur ou à un module d'analyse de les retrouver. OntoExplo propose donc un module d'indexation des documents scientifiques à base d'ontologies. Ce module assure la catégorisation automatique des documents à partir des concepts d'une ontologie du domaine abordé dans le corpus (les concepts étant considérés comme des catégories). La catégorisation automatique de documents, dont le but est d'affecter automatiquement des documents à des catégories pré-définies [22], est généralement basée sur des approches statistiques ou syntaxiques [12]. L'ontologie permet tout d'abord de définir l'ensemble des catégories ou thématiques du domaine, mais également d'ajouter de la sémantique dans un processus de catégorisation à partir des relations qu'elle établit entre les concepts. La méthode que nous utilisons est décrite dans [18].
Une deuxième fonction consiste à *explorer l'ensemble des informations* disponibles dans la collection afin de découvrir de nouvelles connaissances ou des informations utiles mais non connues à priori. Ainsi, le système doit à partir d'ontologie supporter l'aide à une activité d'analyse en extrayant de l'information implicitement présente dans la collection à partir de règles d'inférence définies sur la connaissance qu'elle représente.
Une troisième fonction du système consiste à proposer une *interface permettant la recherche, l'exploration et la visualisation* de données hétérogènes. L'interface se base sur le résultat de la

catégorisation des documents suivant l'ontologie et sur les informations élaborées lors de l'étape précédente.
Afin de mettre en place un système répondant aux attentes précédemment énoncées, nous avons défini un modèle de représentation des données basé sur une ontologie du domaine de la veille et une ontologie du domaine abordé dans le corpus. Ces deux ontologies sont présentées dans la section suivante.
La cadre applicatif de ce système est l'astronomie. De grandes masses de données, réparties et hétérogènes sont produites en astronomie telles que les archives des observatoires au sol et spatiaux, les catalogues et la cartographie numérique du ciel, les articles dans les journaux électroniques. La base de données Simbad [http://simbad.u-strasbg.fr/] regroupe par exemple des informations sur plus de 1,5 millions d'étoiles et 1,25 millions d'objets non stellaires (galaxies, novae, supernovae,…). Les données stockées sur ces objets peuvent être de plusieurs types : mesures, observations, bibliographie… Le serveur bibliographique ADS [http://ads.harvard.edu/unavailable.html] contient plus de 3 millions de résumés d'articles publiés dans les domaines de l'astronomie, l'astrophysique, l'instrumentation, la physique et l'astrophysique. Le Centre de Données astronomiques de Strasbourg possède plus de cinq milles catalogues répertoriant les observations faites sur les étoiles, galaxies et autres objets astronomiques. Le projet MDA [16] vise à permettre une utilisation scientifique optimale de ces informations. Les objectifs définis pour le système OntoExplo s'inscrivent dans le cadre de ce projet. L'astronomie est donc le domaine choisi pour valider notre approche. Les méthodes développée sont cependant génériques et peuvent s'adapter à d'autres domaines par le choix d'ontologies relatives aux nouveaux domaines d'application.

4. Ontologie du domaine de la veille et ontologie du domaine abordé dans le corpus

4.1. Définitions

Le modèle de représentation des données proposé est à base d'ontologies. Ces ontologies doivent permettre la représentation d'un domaine à la fois à travers le vocabulaire de ce domaine, mais également au travers de l'ensemble des méta-données qui peuvent être utiles dans des activités d'accès à l'information et d'exploration. Le modèle repose donc sur deux ontologies différentes : une ontologie du domaine de la tâche à réaliser (ici la veille) et une ontologie du domaine abordé dans le corpus. Ces deux types d'ontologie font appel à des concepts différents.
Une **ontologie du domaine de la tâche** organise les méta-données, à savoir les rôles des connaissances dans la réalisation de la tâche.
Une ontologie référant aux activités de veille en astronomie est présentée figure 1 ; elle a été construite en coopération avec des astronomes [19]. Cette ontologie a pour objectif de permettre une cartographie du domaine en présentant les acteurs du domaine (auteurs, laboratoires, pays d'affiliation des auteurs), l'aspect temporel, mais également les journaux du domaine dans lesquels les articles scientifiques sont publiés. L'ontologie du domaine de la veille est construite semi-automatiquement à partir des méta-données connues et utiles pour la tâche visée. C'est-à-dire que les concepts sont choisis à priori, mais des techniques automatiques permettent d'extraire leur valeur de façon automatique. Nous nous appuyons pour cela sur les travaux de [7].
L'**ontologie du domaine abordé dans le corpus** quant à elle représente la connaissance liée au domaine traité dans le contenu des documents du corpus. Dans le cadre des travaux présentés dans cet article, cette ontologie s'attache à définir la connaissance liée au domaine de l'astronomie (objets astronomiques, techniques, phénomènes physiques, théories). Deux ontologies de l'astronomie existent [14] [8]. Une partie de l'ontologie de l'astronomie issue de [8] est représentée dans la figure 2.
Une ontologie de domaine est généralement construite de façon manuelle, mais certains travaux visent à automatiser certaines étapes de la construction afin de limiter le coût de construction et de mise à jour. La construction d'une ontologie de domaine ne fait pas partie de ce projet. D'autant plus que deux ontologies de l'astronomie sont disponibles gratuitement [14] [8]. Le système présenté se veut générique et peut reposer sur n'importe quelle ontologie liée au domaine visé.

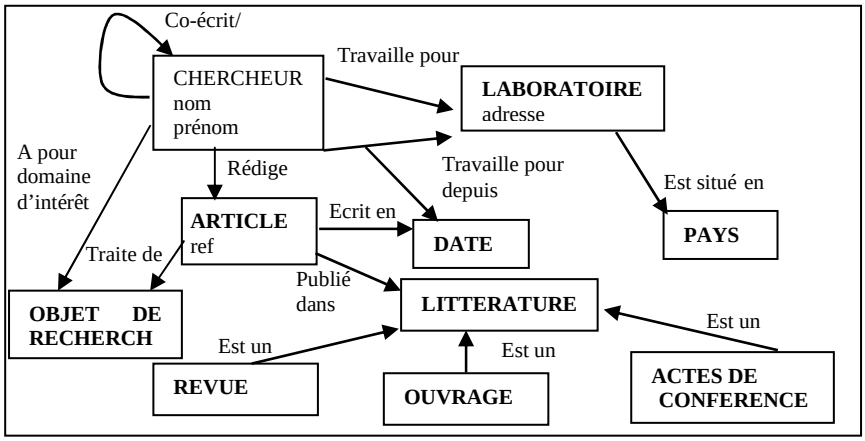

Figure 1. Ontologie du domaine de la veille

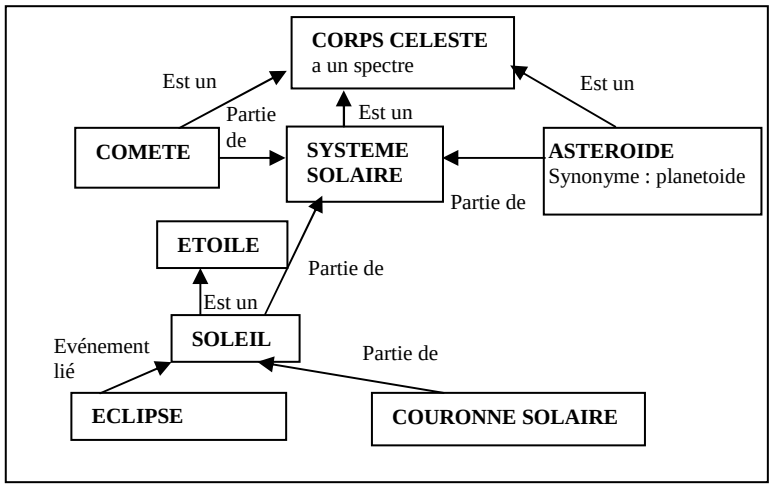

Figure 2. Ontologie de domaine

4.2. Intérêts de ces ontologies

Le premier intérêt d'utiliser des ontologies dans notre modèle est que les relations utilisées dans les deux ontologies définies dans la section précédente, sont sémantiquement plus nombreuses et mieux définies que celles présentes dans les hiérarchies de concepts traditionnellement utilisées en RI et en veille (cf partie 2). Ceci permet de catégoriser les documents en se basant sur une sémantique plus riche. Les inférences qui pourront être déduites seront plus précises et les documents pourront être restitués par des procédés mieux adaptés. Les deux sections suivantes présentent l'intérêt de l'utilisation de chacune de ces deux ontologies.

4.2.1. Utilisation de l'ontologie du domaine de la veille

Dans notre système, les documents sont indexés à partir de l'ontologie du domaine de la veille. Cette première indexation permet de prendre en compte les différentes meta-données définies dans l'ontologie de la veille et retrouvées dans les documents. L'indexation consiste à créer une instance des concepts de l'ontologie pour chaque valeur de méta-données retrouvée dans le document. La détermination automatique des valeurs des meta-données se fait à partir des travaux issus de [7]. Cette indexation a pour intérêt de permettre de nombreuses activités de veille. L'ensemble des méta-données issues du corpus est représenté par des instances de l'ontologie donnant accès aux relations sémantiques entre ces meta-données. Ceci permet d'intégrer de la sémantique dans les analyses qui seront effectuées sur ces données. L'originalité des travaux présentés dans cet article porte sur cette ontologie de veille car elle permet de représenter la connaissance utile dans les activités d'analyse.
Dans notre cas, un exemple d'indexation est la création de l'instance *Chercheur ayant pour Nom Dupond et Prénom Jean* et de l'instance *Chercheur ayant pour Nom Durant et Prénom Michel* reliés tous deux par la relation *co-écrit* si les deux chercheurs sont co-auteurs d'un même article. Après une analyse des réseaux et groupes de chercheurs fortement corrélés la relation *travaille avec* pourra

ensuite éventuellement être établie entre ces deux chercheurs. De la même façon l'ensemble des concepts retrouvés dans le document est instancié (publication, organisme de rattachement d'un chercheur, …). A partir de cette indexation, les activités d'analyse peuvent porter par exemple sur l'étude des collaborations entre chercheurs dans un organisme au cours du temps, le nombre d'articles par pays, l'influence de l'arrivée d'un chercheur dans un organisme ….

4.2.2. Utilisation de l'ontologie du domaine abordé dans le corpus

Les documents sont également indexés par rapport à l'ontologie du domaine abordé dans le corpus. Le contenu de chaque document est analysé et les concepts de l'ontologie de domaine jugés les plus représentatifs sont retenus pour l'indexer.
Cette indexation permet de représenter le contenu des documents à partir de la connaissance du domaine abordé dans le corpus. Ceci présente de nombreux avantages comme explicité dans la section 2.2.

4.3. Liens entre les deux ontologies

La création de ces deux ontologies distinctes mais accessibles l'une par l'autre permet d'optimiser l'aide à la réalisation d'une tâche. La connaissance représentée à travers ces deux ontologies peut ainsi être combinée pour que de nouvelles connaissances soient déduites de ces deux représentations. Dans notre cas, afin d'optimiser les tâches d'exploration et de veille, il est intéressant de prendre en compte les méta-données relatives aux documents, la connaissance lié au domaine d'application mais aussi la connaissance établie à partir des deux ontologies. Les valeurs de certaines méta-données sont retrouvées à partir de la connaissance du domaine abordé dans le corpus, de la même façon les méta-données permettent d'affiner et de stocker les informations issues du domaine abordé dans le corpus.
Le lien entre les deux ontologies est présenté figure 3.
L'ontologie du domaine de l'astronomie permet de déterminer le contenu de méta-données qui ne font pas partie du document au préalable. Dans la figure 3, l'instance du concept *objet de recherche* de l'ontologie de veille est trouvé dans l'ontologie de l'astronomie. Le « domaine d'intérêt » d'un chercheur est établi automatiquement à travers la catégorisation des publications dont il est auteur. Les concepts de l'ontologie de l'astronomie jugés les plus représentatifs de ces articles permettent d'établir la relation.
L'ontologie de la veille permet quant à elle de stocker à partir des instances des *articles* les thématiques abordées et qui sont issues du domaine de l'astronomie. Les thématiques sont mémorisées comme des instances du concept *objet de recherche*. Ces instances sont déterminées à partir des concepts jugés représentatifs des articles dans l'ontologie de l'astronomie.
Dans l'optique d'activités de veille, les liens établis entre les deux ontologies permettent la mise en place de nouvelles analyses. Les thématiques de recherche des organismes pourront être analysées en se basant sur les centres d'intérêts des chercheurs appartenant à cet organisme ainsi que l'évolution de ces thématiques au cours du temps à partir des dates de publication des articles, et aussi l'influence d'un chercheur sur les axes de recherche à partir de la date à laquelle il a rejoint l'organisme.

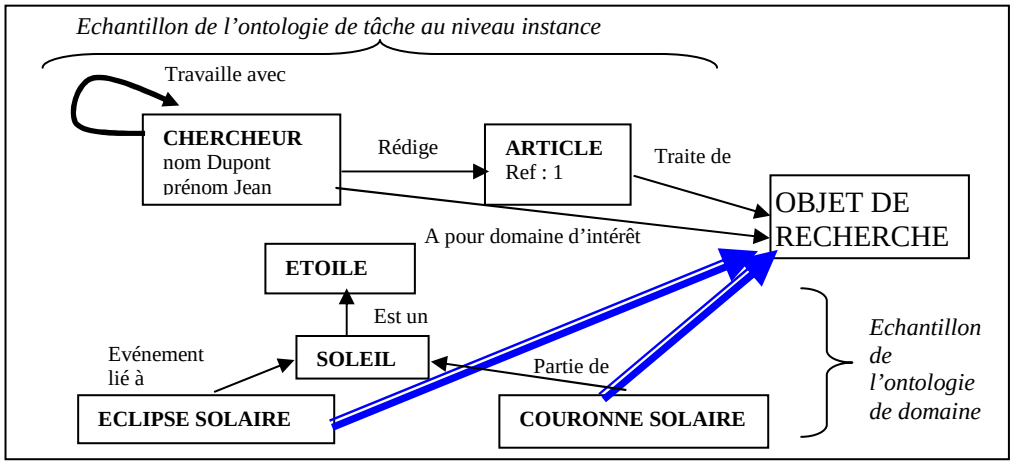

Figure 3. Liens entre l'ontologie de tâche et l'ontologie de domaine

5. Interface de visualisation du corpus

Une interface basée sur ce modèle est en cours de développement. Une première copie d'écran de cette interface est présentée dans la figure 4. L'interface permet de visualiser à la fois l'ontologie du domaine de la veille et l'ontologie du domaine de l'astronomie instanciées par rapport au corpus à analyser. Les liens entre les deux ontologies sont présentés à l'utilisateur. Pour cela, l'écran est partagé en deux fenêtres, la fenêtre de gauche permettant de visualiser l'ontologie de veille et la fenêtre de droite l'ontologie de domaine liée au corpus traité. Différents icônes permettent d'ouvrir les deux ontologies, charger le corpus et de zoomer sur certaines parties des ontologies.
Deux types d'exploration du corpus sont possibles à partir de cette interface. L'exploration à partir de l'ontologie du domaine de la veille est présentée dans la première section, l'exploration à partir de l'ontologie du domaine de l'astronomie est présentée dans la seconde section.

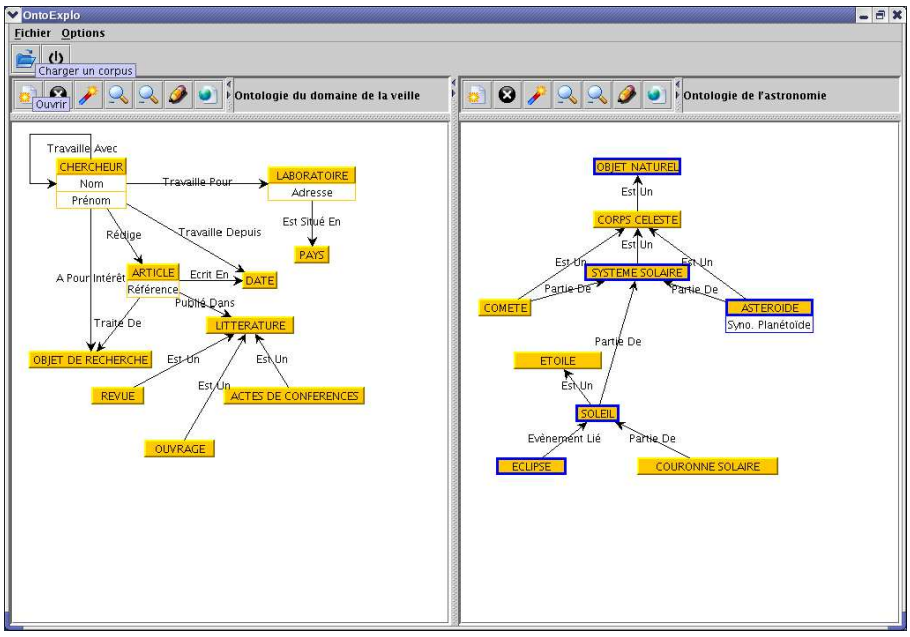

Figure 4. Interface d'exploration du corpus.

5.1 Exploration à partir de l'ontologie du domaine de la veille

L'utilisateur peut naviguer au sein de l'ontologie du domaine de la veille et parcourir les différentes instances des concepts de cette ontologie (figure 5). Comme présenté dans la figure 6, en sélectionnant l'instance sur laquelle porte sa recherche (comme par exemple le *chercheur Dupont)*, l'utilisateur visualise toutes les relations connues entre cette instance et les autres instances de la base (chercheurs avec qui il travaille : M Durant,…, articles : ref 124,148,769, laboratoire de rattachement : CSTB, …). Les instances du concept *objet de recherche* (domaine d'intérêt d'un chercheur ou bien thème d'un article) retrouvés dans l'ontologie de l'astronomie, sont présentées à l'utilisateur dans la partie droite de l'interface (les domaines d'intérêt de Dupont sont les systèmes solaires et les éclipses). L'utilisateur peut ainsi situer ces concepts dans leur contexte. Les concepts sont en effet présentés au sein d'échantillons de l'ontologie ,ceci permettant de visualiser les concepts auxquels ils sont reliés.

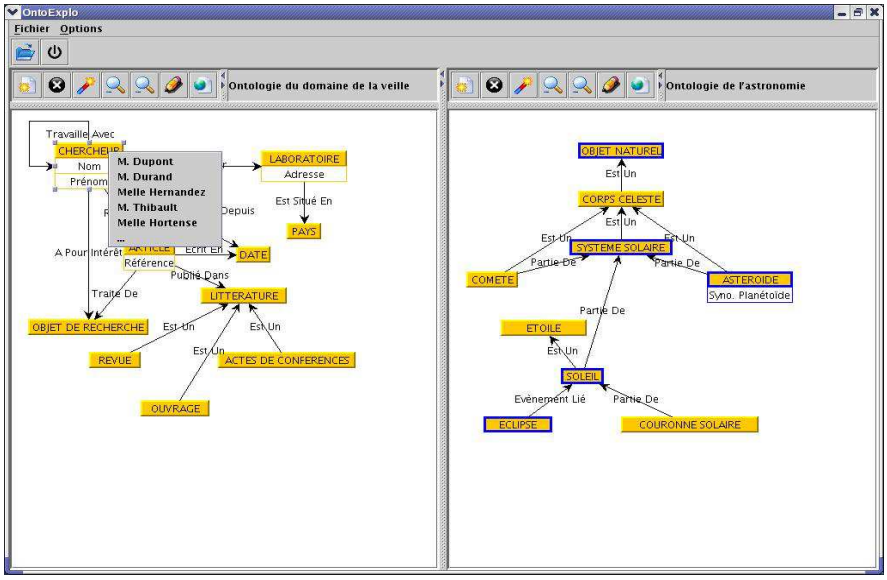

Figure 5. Visualisation des instances de l'ontologie de la veille

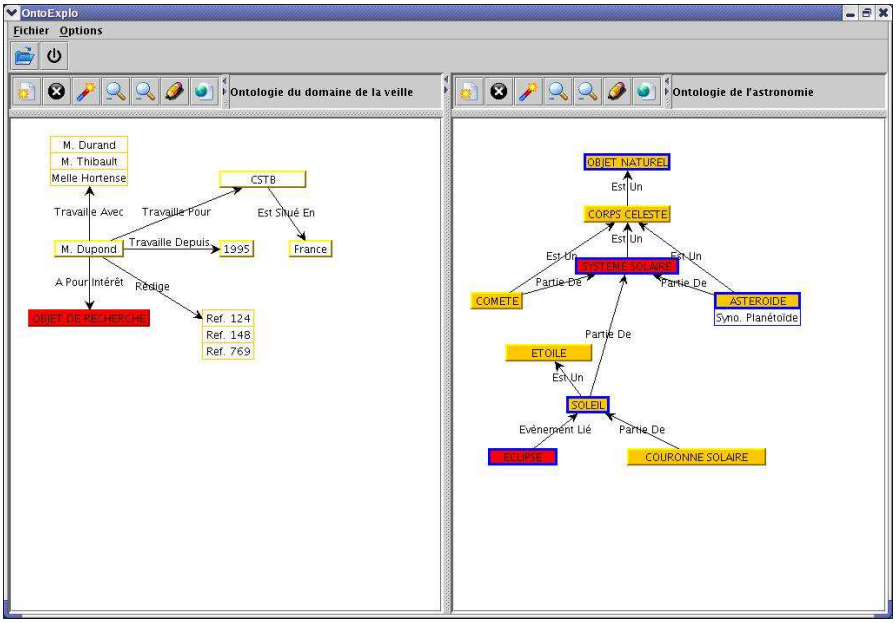

Figure 6. Visualisation de la connaissance établie pour une instance de l'ontologie de la veille

5.2 Exploration à partir de l'ontologie du domaine abordée dans le corpus

L'utilisateur peut aussi choisir de naviguer au sein de l'ontologie du domaine de l'astronomie. Cette possibilité est présentée dans la figure 7. Les concepts présents dans le corpus sont alors mis en évidence. A partir d'un clic sur un concept, l'utilisateur peut retrouver quels sont les articles abordant ce concept ainsi que les chercheurs s'intéressant à cette thématique. L'exploration du corpus se fait donc à la fois à travers la connaissance relative à la tâche de veille mais également à travers une représentation de la connaissance du domaine de l'astronomie.

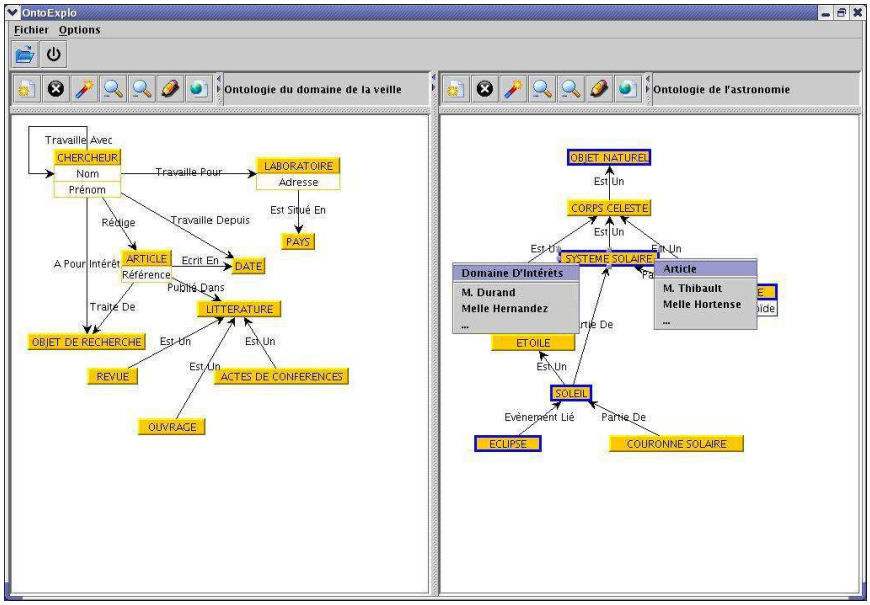

Figure 7. Navigation au sein de l'ontologie de l'astronomie

6.Conclusion

Le système que nous présentons a pour objectif d'aider non seulement une activité de RI mais également une activité de veille et d'exploration interactive d'un domaine par le biais d'ontologies. Plus précisément, il doit permettre une exploitation optimale des grandes masses de données relatives au domaine étudié. Nous avons proposé une ontologie du domaine de la veille contenant les méta-données liées à cette tâche et un échantillon d'une ontologie du domaine (ici l'astronomie). L'ontologie du domaine de la veille est construite semi-automatiquement. Elle est composée de méta-données connues du domaine et enrichie à partir de liens inférés de l'ontologie du domaine de l'astronomie (tels que les domaines d'intérêt des chercheurs, les thématiques des laboratoires…). Un prototype d'interface est présenté. L'utilisateur peut naviguer au sein des données issues du corpus à travers la représentation de la connaissance liée à la veille ainsi que celle liée au domaine de l'astronomie.

Plusieurs perspectives sont à envisager. Tout d'abord l'interface devrait prendre en compte la possibilité de visualiser des ontologies contenant de nombreux concepts et de nombreuses instances. Les ontologies de l'astronomie, par exemple, sont formées de plusieurs centaines de concepts, une visualisation adaptée à l'utilisateur doit donc être définie pour lui permettre une vision générale mais aussi précise de la connaissance qui lui est présentée. La création de profils utilisateur est envisagée, ces profils permettraient de personnaliser l'interface en fonction de caractéristiques choisies par l'utilisateur (niveau de profondeur des concepts visualisés, choix des relations à afficher …).. Ensuite, les analyses des corpus peuvent conduire à la proposition de nouveaux concepts et relations à ajouter aux deux ontologies, l'interface devrait donc prendre en compte ce cas de figure. Finalement, une autre perspective est de proposer des mesures permettant d'évaluer l'adéquation entre une ontologie de domaine et un corpus. Ces mesures permettraient de choisir l'ontologie la plus adaptée à l'analyse d'un corpus en évaluant également ses lacunes et les modifications à y apporter.

Remerciements

Nous tenons à remercier Laurent Cardoner, stagiaire à l'IRIT, qui a réalisé l'implantation de l'interface.

Bibliographie

[1] Augé, J., Englmeier, K., Hubert, G., & Mothe, J. (2001). Classification automatique de textes basée sur des hiérarchies de concepts. Veille stratégique, scientifique et technologique, Barcelone, 291-300.

[2] Aussenac-Gilles, N., & Mothe, J. (2004, April). Ontologies as background knowledge to explore document collections. In Coupling approaches, coupling media and coupling languages for information retrieval (pp. 129-142). LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE.

[3] Benjamins R., Fensel D., Decker D., Gomez Perez A., *(KA)2 : building ontologies for the internet : a mid-term report*, International Workshop on ontological engineering on the global information infrastructure, pp 1-24, 1999.

[4] Borst P., *Construction of Engineering Ontologies for Knowledge Sharing and Reuse*, Rapport de thèse, Tweente University, 1997. http://www.ub.utwente.nl/webdocs/inf/1/t0000004.pdf

[5] Bourigault D., Aussenac-Gilles N., *Construction d'ontologies à partir de textes*, tutorial TALN, 2003.

[6] 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*, EC-Web 2002, pp 304-313, 2002.

[7] Dkaki, T., Dousset, B., & Mothe, J. (1997). Mining information in order to extract hidden and strategic information. In Proceedings of the 5th International Conference on computer-assisted information searching on Internet (RIAO'97.

[8] FACT GURU http://www.site.uottawa.ca:4321/astronomy/index.html

[9] Fensel D., *Ontologies: a silver bullet for Knowledge Management and Electronic Commerce*, Berlin, Springer Verlag, ISBN 3-540-00302-9, 2001.

[10]Grivel L., Guillemin-Lanne S., Coupet P., Huot C., *Analyse en ligne de l'information : une approche permettant l'extraction d'informations stratégiques basée sur la construction de composants de connaissance*, VSST, pp 149-161, 2001

[11]Gruber R. T., *A Translation Approach to Portable Ontology Specification*, Knowledge Acquisition (5), pp 199-220, 1993.

[12]Hernandez N., *Etude de l'utilisation de syntagmes nominaux pour la catégorisation automatique de documents*, INFORSID, pp 53-68 , 2003.

[13]Himsolt M., *The graphlet system*, Graph Drawing, volume 1190 of Lecture Notes in Cornputer Science, Springer- Verlag, pp 233-240, 1996.

[14]IAU http://msowww.anu.edu.au/library/thesaurus/

[15]Masolo C., *Ontology driven Information retrieval: Stato dell'arte*, Rapport de IKF (Information and Knowledge Fusion) Eureka Project E!2235. LADSEB-Cnr, Padova (I), 2001.

[16]MDA, http://cdsweb.u-strasbg.fr/MDA/mda.html

[17]Mothe J., Chrisment C., Dousset B., Alaux J., *DocCube: Multi-Dimensional Visualisation and Exploration of Large Document Sets*, Journal of the American Society for Information Science and Technology, Special topic section: web retrieval and mining, Guest Editor: Hsinchun Chen, 54 (7), pp 650-659, 2003.

[18]Hubert G., Augé J., Englmeier K., Mothe J., *Catégorisation automatique de textes basée sur des hiérarchies de concepts*, Journées Bases de Données Avancées, pp 69-87, Octobre 2003.

[19]Mothe J., Egret D., Chrisment C., Englmeier K., *Exploring Bibliographic Collections using Concept Hierarchies*, Library and Information Services in Astronomy IV, B. Corbin, E. Bryson, and M. Wolf (eds.), 2002.

[20]Piatetsky-Shapiro G., *Knowledge Discovery in Databases: 10 years after* , SIGKDD Explorations, Vol 1, No 2, 2000.

[21]Royaute J., Francois C., Besagnu D., *Apport d'une méthodologie de recherche de termes en corpus dans un processus de KDD : application de veille en biologie moléculaire*, VSST, pp 49-62, 2001

[22]Sebastiani F.,. *Machine Learning in Automated Text Categorization*, ACM Computing Surveys, vol 34, No. 1, pp 1-47, 2002.

[23]Studer R., Benjamins R., Fensel D., *Knowledge Engineering: Principles and Methods, Data and Knowledge Engineering*, 25(1-2) pp 161-197, 1998.