Opérations sur des ressources hétérogènes dans un entrepôt de données à base d'ontologie

Ghoula, Nizar; de Ribaupierre, Hélène; Tardy, Camille; Falquet, Gilles

Opérations sur des ressources hétérogènes dans un entrepôt de données à base d'ontologie

Nizar Ghoula* — **Hélène de Ribaupierre*** — **Camille Tardy*** — **Gilles Falquet***

* *Centre universitaire d'informatique, Université de Genève*
7, route de Drize, CH-1227 Carouge, Suisse
{Prenom.Nom}@unige.ch

RÉSUMÉ. *Le nombre croissant d'ontologies disponibles sur le web a justifié l'apparition d'entrepôts d'ontologies. Cependant, peu d'entre eux intègrent également des ressources hétérogènes de type ontologique, textuelle, linguistique et terminologique. De plus, ils ne fournissent pas un ensemble complet d'opérateurs permettant la gestion et le traitement de ces ressources. Dans nos précédents travaux, nous avons proposé une approche pour la modélisation et la construction d'un entrepôt de ressources. Dans cet article, nous proposons une modélisation et une taxonomie d'opérateurs de gestion de connaissances et de ressources, et nous présentons deux scénarios d'utilisation. Nous pouvons combiner ces opérateurs afin de modéliser des processus complexes tels que l'intégration, l'annotation et l'alignement.*

ABSTRACT. *Knowledge engineering usually relies on knowledge resources, typically ontologies. A central point of our approach is to build a repository of knowledge resources. This repository is a collection of heterogenous resources that are based on different formalisms or models. In this paper we will focus on the description and use of some operators to manage and build new resources. As we demonstrate in two scenarios, these operations can be combined and implemented in different ways.*

MOTS-CLÉS : *Ontologie de ressources, Opérations, Terminologie, Alignement, Entrepôt de ressources*

KEYWORDS: *Ontology of Resources, Operations, Terminology, Alignment, Resources Repository*

1. Introduction

L'extraction et la représentation des connaissances est un problème largement exploré dont une des solutions est basée sur l'utilisation de ressources ontologiques, terminologiques et linguistiques. Ces connaissances existent actuellement sous forme de ressources de différents types tels que les terminologies, les bases de données terminologiques, les glossaires, les ontologies (générales ou de domaine), les dictionnaires multilingues ou encore les corpus de textes. Ces ressources sont représentées à l'aide de divers formalismes et langages (logique des prédicats, logique de description, réseaux sémantiques, graphes conceptuels, etc.). Dans le cadre d'une application qui nécessite l'usage d'un certain nombre de ressources externes, un concepteur est souvent amené à effectuer un travail laborieux de recherche et de pré-traitement (Sindt, 2003) afin de rassembler et de fabriquer des ressources adéquates aux besoin de ses applications.

Nous avons identifié trois problématiques : (i) les applications demandent de plus en plus de ressources couvrant des domaines multiples, ainsi, il faut rechercher et gérer ces ressources. (ii) Pour prendre en compte le contenu des documents, il faut utiliser des ressources permettant de les décrire et de les indexer, il est donc indispensable de rechercher et d'adapter des ressources externes pour cette tâche. (iii) Pour des besoins plus avancées, il faut souvent créer des collections de ressources complexes résultant de l'application d'un ou plusieurs processus (comme la désambiguïsation). De ce fait, il faut se doter de moyen d'intégration et de combinaison de ces ressources.

L'objectif principal de notre approche est de pouvoir générer de nouvelles ressources à partir de la composition des ressources existantes dans l'entrepôt. Ainsi, l'enrichissement des connaissances dans l'entrepôt s'effectue à chaque utilisation. D'autres approches semblables à TOK ont été proposées. Par exemple, BioPortal[1] (Noy *et al.*, 2004), offre la possibilité de rechercher et aligner manuellement des ontologies en exprimant des requêtes sur leurs entités. Dans un même contexte d'utilisation, le moteur de recherche d'ontologies Watson[2] (D'Aquin *et al.*, 2009) permet de repérer et d'indexer les ontologies du web sémantique en gardant des références vers leurs entités.

Contrairement à TOK, ces systèmes se focalisent sur un type spécifique des ressources et n'offrent pas la possibilité de référencer ou d'effectuer des processus complexes sur des ressources hétérogènes.

Dans cet article, nous présentons une modélisation et une taxonomie d'opérateurs de gestion de connaissances et de ressources, ainsi que deux scénarios d'utilisation de notre entrepôt : le premier scénario traite de l'importation de plusieurs corpus de textes, annotés par différents modèles de représentation et de l'alignement des différentes balises de métadonnées. Le deuxième scénario traite de la fusion de plusieurs ressources lexicales dans le but de créer un vocabulaire afin d'annoter des documents.

1. http://bioportal.bioontology.org/.
2. http://watson.kmi.open.ac.uk/WatsonWUI/.

Nous prenons en considération la possibilité de combiner ces opérateurs afin de modéliser des processus complexes tels que l'intégration, l'annotation et l'alignement. Nous montrons à l'aide de ces deux scénarios l'utilité de notre approche, ainsi que la capacité de notre modèle et système à satisfaire des besoins différents dans le domaine de l'ingénierie de connaissances.

Dans la section suivante, nous décrivons une représentation formelle de notre entrepôt de ressources(TOK). Ensuite nous définissons quelques opérateurs fournis par TOK pour gérer des ressources de connaissances. Enfin, nous détaillons les deux scénarios à travers lesquels nous montrons l'utilité de notre système.

2. Entrepôt de ressources de connaissances

Nous avons proposé un modèle permettant d'unifier la représentation des ressources hétérogènes dans un formalisme pivot (Ghoula *et al.*, 2010). Il permet aux utilisateurs d'importer, de stocker leurs ressources dans un même espace et de choisir un modèle commun de représentation. Notre implémentation comporte une ontologie (TOK_Onto[3]), qui permet de décrire l'ensemble des ressources et de construire un entrepôt de contenu.

2.1. *Représentation formelle de l'entrepôt*

L'approche que nous proposons repose sur un modèle d'entrepôt de ressources constitué de trois niveaux : (i) ressource, (ii) représentation et (iii) définition. Lorsqu'une nouvelle ressource est importée dans le système nous en stockons une copie (niveau (i)). Une représentation de celle-ci est ensuite générée et stockée (niveau (ii)). Cette représentation joue deux rôles : 1) décrire globalement la ressource par des métadonnées et 2) décrire son modèle. Le niveau (iii) sert à importer le contenu de la ressource en utilisant les relations et les types de données du modèle de représentation utilisé. Dans le cadre de cet article, deux types de ressources sont ciblées : les ontologies formelles et les corpus de documents annotés.

Étant donné la diversité des ressources de connaissances terminologiques, ontologiques et linguistiques et la variété des formalismes et langages de représentation des connaissances, il serait vain de tenter de définir un modèle unifié capable de représenter le contenu de n'importe quelle ressource. L'approche que nous proposons consiste donc à définir un ensemble de modèles abstraits de contenu en fonction des besoins.

Définition 1 (Entrepôt de ressources) *Un entrepôt de ressources de connaissances terminologiques, ontologiques noté $\mathcal{TOK}$ est un tuple $\langle \mathcal{R}, \mathcal{T}_r, \mathcal{T}_e, \mathcal{M}, \mathcal{O}, \mathcal{P} \rangle$, tel que :*

– $\mathcal{R}$, est un ensemble de ressources de connaissances terminologiques, ontologiques et de connaissances ;

3. http ://cui.unige.ch/isi/onto/tok/OWL_Doc/

– $\mathcal{T}_r$ et $\mathcal{T}_e$, sont les ensembles respectifs des types de ressources (autonome, annotation, alignement) et des entités (concept, relation, terme, etc.) ;

– $\mathcal{M}$, est un ensemble non vide de modèles de représentation sous-jacents aux ressources stockées dans l'entrepôt ;

– $\mathcal{O}$, est un ensemble d'opérateurs sur les ressources ;

– $\mathcal{P}$, est un ensemble de processus définis à l'aide des opérateurs qui évolue au cours de l'utilisation de l'entrepôt.

Définition 2 (Modèle) *Un modèle de représentation d'une ou plusieurs ressources dans l'entrepôt $\mathcal{TOK}$ est un tuple $M = \langle Types, Rel, A \rangle$, tel que :*

– $Types = \{t_1, .., t_k\}$, est un ensemble non vide de types d'entités utilisées dans le modèle ;

– $Rel = \{rel_1, .., rel_m\}$, est un ensemble non vide de relations pré-définies propres au modèle ;

– $A = \{a_1, .., a_n\}$, est un ensemble de contraintes sur les relations du modèle ; Si $a_i \in A$ alors $a_i = \langle t_s, rel_j, t_l \rangle$ avec $t_s \in Types, t_l \in Types$ et $rel_j \in Rel$.

Lors de l'importation dans l'entrepôt, nous pourrons choisir les modèles de représentation nécessaires à l'exécution des tâches pour lesquelles la ressource est requise.

Propriété 1 (Instance d'un modèle) *Soit $M_i \in \mathcal{M}$ un modèle de représentation. Si $L_i \in \mathcal{L}$ est l'ensemble des liens entre les entités de la ressource $R_j \in \mathcal{R}$ relatif au modèle M_i, $M_i.Rel$ est l'ensemble des relations pré-définies propres au modèle M_i, $M_i.Types$ est l'ensemble de types d'entités utilisées dans le modèle M_i et n le nombre des entités de cette ressource, alors $L_i = \{\langle E_s, rel_v, E_t \rangle \mid E_s \in \mathcal{E}_j, E_t \in \mathcal{E}_j, rel_v \in M_i.Rel, E_s.type \in M_i.Types\}$. $L_i \in \mathcal{E}_j \times \mathcal{E}_j \times M_i$.*

Dans TOK, les modèles de représentations du contenu sont associés à un ensemble d'opérateurs. Cet ensemble comporte des implémentations des opérateurs abstraits en fonction des spécificités du modèle.

3. Taxonomie des opérations sur les ressources de connaissances

Nous avons défini un ensemble d'opérateurs atomiques sur les modèles de ressources fournis par l'entrepôt TOK tels que l'alignement, l'annotation, la fusion, la sélection, le mapping entre modèles, la jointure et la composition (Falquet *et al.*, 2008). Dans cette partie, nous présentons trois opérateurs qui seront exploités dans les scénarios ci-dessous.

3.1. *Opérations de représentation*

À chaque modèle de représentation du contenu on associe un ou plusieurs opérateurs d'importation et d'exportation spécifiques.

3.1.1. *Importation ou abstraction*

Il n'est pas toujours nécessaire de préserver tout le contenu d'une ressource lors de son importation. Par exemple, dans le cadre du deuxième scénario, nous n'avons pas besoin d'importer les axiomes. Ainsi, un modèle de représentation de type WordNet (figure 1) suffit pour représenter les entités qui nous intéressent. Cette opération transforme une ressource écrite dans une langage L en une ressource représentée par un modèle M.

Dans ce type de modèle, nous partons des entités conceptuelles, liées entre elles par des relations sémantiques, vers les entités terminologiques reliées avec les concepts par des relations de description. Les termes sont décrits par des formes lexicales qui sont des entités linguistiques permettant de désigner un terme dans une langue donnée.

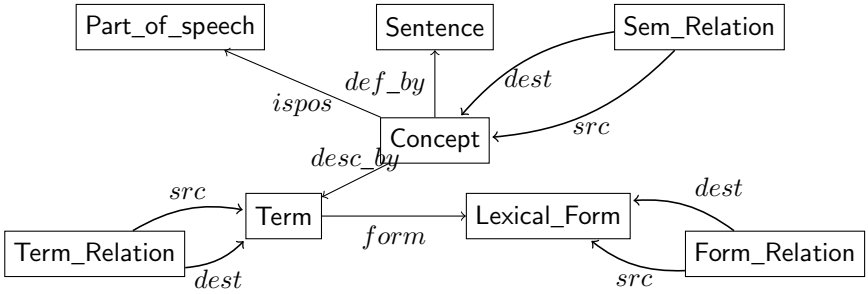

Figure 1. *Partie de la description du modèle WordNet_Like (WN_Like)*

3.1.2. *Exportation ou réification*

Cet opérateur permet de passer de la représentation abstraite de la ressource dans l'entrepôt dans un modèle M vers la génération d'une ressource indépendante dans un langage ou formalisme L.

On dénote par i_{LM} l'opération qui à partir d'une ressource exprimée dans un langage L, produit une représentation de son contenu dans un modèle M et par e_{ML} l'opération d'exportation correspondante. La composition $e_{ML}(i_{LM}(R))$ ne permet pas toujours de retrouver la ressource d'origine R car l'opération i_{LM} peut faire perdre de l'information. Par contre, il est possible, et désiré, que $i_{LM}(e_{ML}(S)) = S$ soit valide où S est une représentation dans un modèle LM de la ressource R.

3.2. *Opérations d'enrichissement*

Les opérateurs d'enrichissement permettent de générer de nouvelles ressources complémentaires ou indépendantes des celles existantes. Ils sont basés sur des algorithmes spécifiques (Euzenat *et al.*, 2003) et des ressources auxiliaires telles que WordNet ou des dictionnaires multilingues.

3.2.1. *Alignement*

L'alignement permet d'exprimer explicitement des relations entre des ressources (Kalfoglou *et al.*, 2003). Une méthode d'alignement est composée de la définition d'une distance entre les entités d'une ressource et du calcul de la meilleure correspondance entre elles en réduisant au minimum la distance de mesure de similarité (Euzenat *et al.*, 2007).

Définition 3 (Opérateur d'alignement) *Un opérateur d'alignement prend en entrée deux ressources R_i et R_j par rapport à un modèle de représentation M_1 et un ensemble de ressources auxiliaires représentées dans d'autres modèles $M_2, \ldots$ pour produire une ressource d'alignement représentée dans un modèle M_{al}.*

La signature de cet opérateur est de la forme :

$$M_1, M_1 \rightarrow \mathsf{Align}(M_1, M_{al})[M_2, \ldots] \rightarrow \mathsf{R}_{\mathsf{Align}}(M_{al})$$

M_{al} est un modèle qui regroupe les relations d'alignement utilisées pour représenter les correspondances, $\mathsf{Op}_{\mathsf{Align}}$ est l'opérateur d'alignement utilisé et $\mathsf{R}_{\mathsf{Align}}$ est la ressource d'alignement générée.

Actuellement, la majorité des algorithmes d'alignement peuvent aligner des ontologies en OWL, mais ils n'utilisent pas toute la sémantique exprimée par ce formalisme. Ils sont basés sur les labels textuels attachés à chaque classe dans la structure de l'ontologie. La structure est généralement un graphe représentant la hiérarchie des classes et les propriétés qui font le lien entre deux classes : e.g. Il y a un axiome de la forme $Class_1 \sqsubseteq property$ **only/some** $Class_2$.

Dans ce cas, il est plus approprié de représenter une ontologie en OWL par un graphe de structure que par le modèle de logique de description. Les algorithmes d'alignement seront plus faciles à écrire et vont permettre d'aligner plusieurs types d'ontologies pouvant être représentées par un graphe étiqueté.

3.2.2. *Annotation*

L'opérateur d'annotation permet de décrire des entités d'une ressource R_1 avec les entités d'une ressource R_2 à travers un élément d'annotation de type relation. Cet élément est définit dans un modèle spécifique représenté dans TOK.

La signature d'un opérateur d'annotation est de la forme :

$$M_1, M_2 \rightarrow \mathsf{An}(M_1, M_2, M_{ann}) \rightarrow \mathsf{R}_{\mathsf{An}}(M_{ann})$$

M_1 est le type (modèle) du contenu de la ressource à annoter, M_2 est le type de contenu de la ressource qui sert comme références dans l'annotation, M_{ann} est le modèle d'annotation utilisé et R_{An} est la ressource d'annotation résultante. Par exemple, *word sense disambiguation (WSD)* est un type d'opérateur d'annotation. Partant d'un texte en langage naturel et d'une ontologie lexicale, cet opérateur produit un ensemble de correspondances entre les mots dans le texte et leur sens (les concepts dans l'ontologie). La signature de cet opérateur est de la forme :

$$\textit{WSD : Text, WN\_Like} \rightarrow \mathsf{An}(\textit{Text, WN\_Like})$$

À titre d'exemple, nous allons utiliser TEI[4], DC[5], la DTD de pubMed[6] et d'autres formalismes, comme des ressources d'annotation. Elles seront importées et représentées dans TOK à l'aide d'un modèle M_{ann}. Ce modèle, représente les entités de ces formalismes comme des types et utilise la relation $rdf{:}type$ comme relation d'annotation.

3.2.3. *Fusion*

La fusion consiste à intégrer deux ressources pour produire une nouvelle ressource regroupant les concepts, les relations et les instances des ressources originales (Stumme *et al.*, 2001) (Pinto *et al.*, 2001). La fusion s'effectue par rapport à une ressource d'origine, donc si on a deux entités alignées la ressource résultante va contenir celle de la ressource de référence.

En fonction du modèle de la ressource, cette opération peut prendre des implémentations différentes. Par exemple, l'opérateur de fusion sur deux ontologies dans le modèle DL (logique de description) se réduit à implémenter l'opération d'union de leurs vocabulaires et axiomes (Noy *et al.*, 2001) (Klein, 2001).

Cet opérateur s'applique sur une liste de ressources représentées dans le même modèle et utilise notamment des ressources d'alignement entre elles. Pour fusionner des ressources d'alignement ou d'annotation, il faut que celles-ci soient relatives aux mêmes ressources d'origine. En premier lieu, pour chaque ressource R_i à fusionner, il faut regrouper et fusionner tous les alignements ayant comme source R_i sous le même modèle d'alignement M_{al}. Une ressource d'alignement ayant entrées et sorties multiples est construite et représentée avec le modèle M_{al}. L'ensemble des ressources et l'alignement construit vont permettre de créer une nouvelle ressource qui sera la fusion de toutes les ressources en fonction de l'alignement. La signature d'un opérateur de fusion est :

---

4. http ://www.tei-c.org/index.xml
5. http ://dublincore.org/
6. http ://www.ncbi.nlm.nih.gov/entrez/query/static/PubMed.dtd

$$M_1, M_1 \rightarrow \mathsf{Merge}(M_1, M_1)[M_{al}] \rightarrow \mathsf{R}_{\mathsf{Mg}}(M_1)$$

4. Scénarios d'utilisation de l'entrepôt TOK

En se basant sur l'espace de stockage élaboré, les traitements sur les connaissances devront permettre l'utilisation, la génération, l'intégration de connaissances et la production de nouvelles ressources dans différents formalismes. Dans cette section, nous présentons deux scénarios d'utilisation reflétant les différentes possibilités d'exploitation du système TOK (figure 2).

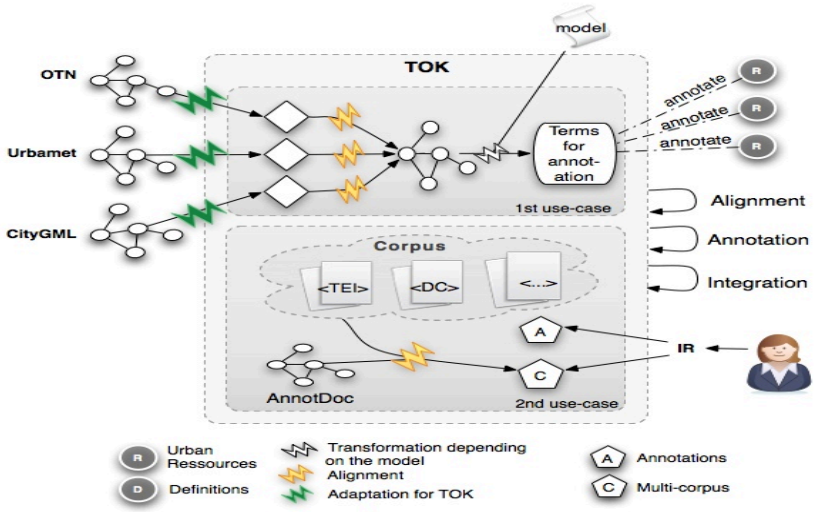

Figure 2. *Utilisation de l'entrepôt TOK dans des scénario de gestion de connaissances*

4.1. *Alignement des formalismes de représentation*

Dans un grand nombre de cas, tels que la recherche d'information et la classification, nous avons besoin d'un corpus annoté permettant de faire des tests, de construire un classifieur, de rechercher l'information en fonction de certains critères, etc. (figure 3)

Le modèle présenté (figure 3) est une partie du modèle d'annotation de documents scientifiques que nous sommes en train de développer dans le cadre d'une autre recherche (de Ribaupierre *et al.*, 2010). Les informations de types bibliographiques

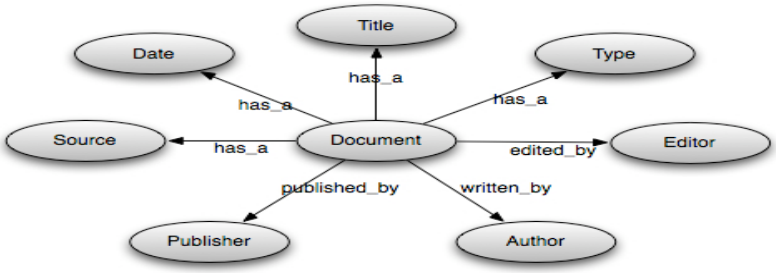

Figure 3. *Modèle des métadonnées du document*

sont utilisées comme caractéristiques pour décrire un document. Ces informations permettent d'identifier le document, mais aussi, de faire de la recherche d'informations précises suivant des critères avancés.

Un corpus annotée est une ressources textuelle enrichie par des informations linguistiques ou structurelles selon un ou plusieurs modèles d'annotation. Le problème de l'hétérogénéité se pose dès que l'on veut utiliser plus d'un corpus annoté par des formalismes différents. Il faut donc aligner ces formalismes. Il existe un grand nombre de formalisme, tels que Dublin Core[7] , TEI , DTD de PubMed, PRISM[8], etc.

Pour aligner ces différentes balises, nous allons utiliser notre modèle générique (indépendant de l'approche TOK) que nous avons défini pour représenter les documents scientifiques. Dans cet article, nous nous concentrons seulement sur la partie qui concerne l'en-tête d'un document.

Supposons que l'on dispose de trois corpus, C_1 en TEI, C_2 en PubMed et C_3 en DublinCore (DC), il est difficile de rechercher des informations dans ces trois corpus par le moyen d'une seule requête. Pour remédier à ce problème, nous allons aligner les modèles d'annotation. En premier lieu, ces formalismes sont importées en utilisant un modèle de description propre à TOK (M_{ann}). Ensuite, nous effectuons des alignements entre les représentations de ces modèles et la représentation de notre modèle d'annotation H (figure 3).

Nous rencontrons un certain nombre de problèmes pour aligner ces représentations. Le premier est courant en alignement, le nom des balises peut avoir le même sens sémantique, mais utilisent des termes différents. Par exemple, pour décrire un auteur dans Dublin Core, la W3C recommande d'utiliser DC.creator, alors que dans TEI, la balise utilisée est <author>. Nous pouvons considérer que ces deux balises ont la même sémantique, c'est-à-dire désigner l'auteur ou les auteurs du document.

7. http ://dublincore.org/
8. http ://www.idealliance.org/industry_resources/intelligent_content_informed_workflow/prism

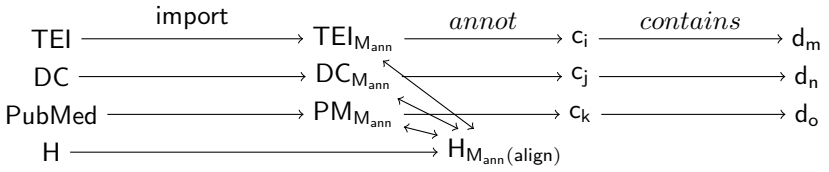

Figure 4. *Importation et alignement des modèles de représentation et des ressources annotées dans TOK*

Le deuxième problème réside dans le fait que la sémantique de certaines balises peut être très proche, voir similaire, à d'autres balises dans le même formalisme. C'est le cas, par exemple, dans Dublin Core où DC.creator, désigne l'auteur, mais, ou DC.contributor désigne plus ou moins la même chose. Dans le cas d'article scientifique, nous décidons que DC.creator a la même sémantique que DC.contributor, et dans le cas où les deux auraient été utilisées, DC.creator est le premier auteur et les auteurs sous DC.contributor sont les 2^{eme} au N^{ieme} auteurs.

Enfin, les corpus sont importés et leurs annotations restent représentées par les modèles relatifs aux formalismes d'origine. À travers le modèle global et ses alignements avec les autres modèles, nous pouvons interroger les corpus avec une même requête.

Le processus peut être décrit de la manière suivante (en assumant que l'implémentation de l'opérateur d'importation par rapport aux modèles décrit est déjà fournie). Les modèles de représentation pour les alignements et les annotations sont : M_{ann}, un modèle décrivant les types des entités et les relations pour l'annotation, et M_{al}, un modèle représentant les relations d'équivalences entre des concepts. Nous supposons que l'importation des corpus est directe par rapport à la représentation de leur formalisme d'origine.

1) importer TEI , DC, $PubMed$ et le modèle H (figure 3) dans le modèle M_{ann} content model :

$$TEI_{M_{ann}} = i_{XML,M_{ann}}(TEI)$$
$$DC_{M_{ann}} = i_{RDF,M_{ann}}(DC)$$
$$PM_{M_{ann}} = i_{XML,M_{ann}}(PM)$$
$$H_{M_{ann}} = i_{OWL,M_{ann}}(H)$$

2) aligner $TEI_{M_{ann}}$, $DC_{M_{ann}}$ et $PM_{M_{ann}}$ à $H_{M_{ann}}$

$$Al_{H_{TEI}} = Al_{M_{ann},M_{ann}}(H_{M_{ann}}, TEI_{M_{ann}})[M_{al}]$$
$$Al_{H_{DC}} = Al_{M_{ann},M_{ann}}(H_{M_{ann}}, DC_{M_{ann}})[M_{al}]$$
$$Al_{H_{PM}} = Al_{M_{ann},M_{ann}}(H_{M_{ann}}, PM_{M_{ann}})[M_{al}]$$

3) importer C_1 dans $TEI_{M_{ann}}$, C_2 dans $DC_{M_{ann}}$ et C_3 dans $PM_{M_{ann}}$:

$$C_{1_{TEI}} = i_{XML,TEI_{M_{ann}}}(C_1)$$
$$C_{2_{TEI}} = i_{XML,DC_{M_{ann}}}(C_2)$$
$$C_{3_{TEI}} = i_{XML,PM_{M_{ann}}}(C_3)$$

4) sélectionner les entités de C_1 dont les tags sont alignés avec $H_{M_{ann}}$:

$$C_{1_H} = \mathsf{select}_{\mathsf{rel\_type}}(H_{M_{ann}}, Al_{H_{TEI}})(C_{1_{TEI}})$$

Cette description est indépendante du nombre de ressources et de leurs évolutions. Dans le cas de changement de type de ressources, il suffit de modifier le modèle de représentation. Les ressources sont représentées à l'intérieur de l'entrepôt. Dans le scénario suivant, les ressources sont stockées indépendamment de TOK. Un composant intermédiaire assure la transformation de ces ressources dans des modèles de représentation.

4.2. *Intégration des ressources*

Pour la planification du développement durable d'une ville, il est important de pouvoir regrouper différentes ressources, provenant de différents domaines pour assister une prise de décision. En effet, l'urbanisme est un domaine interdisciplinaire qui partage certains de ses concepts et termes avec des domaines tels que l'économie, la politique, la sociologie, etc. (Lacasta *et al.*, 2007).

On ne peut attendre d'un utilisateur qu'il soit expert dans toutes ces disciplines et qu'il puisse ainsi être à même de connaître un vocabulaire pertinent pour l'annotation de documents. Les experts du domaine, quand à eux, peuvent certifier la pertinence et la justesse des différents vocables. C'est avec l'intention de combler cet écart que nous proposons un système d'aide à l'annotation. Il sera composé d'espaces de stockage (bibliothèque numérique sémantique) qui contiendront : des ressources lexicales annotées par les experts (couche conceptuelle) mais aussi des documents annotés par les utilisateurs à l'aide des vocabulaires définis par ces ressources (couche ressource).

Dans la littérature, les annotations sont souvent employées au sein du web sémantique, pour fournir à des documents web, des tags sémantiques dérivés d'ontologies (Amann *et al.*, 2000). Cependant ces exemples restent limités à des documents structurés, par exemple HTML, ou sont limités par un vocabulaire non hiérarchique ou très dépendant de leur contexte (Hendler *et al.*, 2008). Nous allons donc essayer d'appliquer ici des annotations sur des documents non structurés, dépendant de plusieurs domaines et basées sur une ontologie alignée.

La couche conceptuelle du système est donc composée d'ontologies comme Hydrontology (Vilches Blazquez et al., 2007, cité dans (Lacasta *et al.*, 2007)), OTN ((Lo-

renz et al., 2005), cité dans (Lacasta *et al.*, 2007)), CityGML[9], de thésaurus comme Urbisoc (Alvaro-Bermejo, 1988, cité dans (Lacasta *et al.*, 2007)), Urbamet[10], Agrovoc[11], etc.

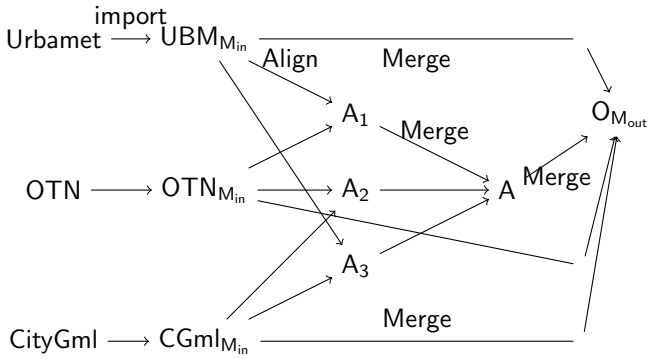

Figure 5. *Alignement et intégration de ressources hétérogènes dans le domaine de l'urbanisme*

Dans le cadre de cet article, nous nous concentrons sur les étapes d'alignements et d'annotations. Les ressources, les documents et les algorithmes de traitement sont stockés dans la bibliothèque du système et non par TOK, comme décrit dans le scénario précédent. Lors de l'ajout d'un nouveau document, les algorithmes, le modèle ainsi que les ressources de la couche conceptuelle nécessaires, sont envoyés à TOK, au sein d'un scénario :

$$S = D + \{O1, ..., On\} + F + M$$

Où S est un scénario, D un domaine de référence donné par l'utilisateur, O_i est une ressource lexicale, F une formule de traitement pour l'alignement et M un modèle pour la transformation de la ressource résultant de l'alignement, afin de permettre l'annotation des documents.

Nous pouvons définir trois étapes lors de l'ajout d'un document par l'utilisateur. Les experts peuplent tout d'abord la couche conceptuelle à l'aide d'ontologies, de thesaurus, etc. qu'ils annotent. Cette annotation consiste à définir les différents domaines et dimensions qui les composent. Les dimensions d'un domaine sont des concepts généraux utilisés pour exprimer les thèmes du domaine (Radhouani, 2008). Un concept est décrit par un ou plusieurs labels (traduction, synonymes), et par une description qui, en quelques mots, le définit et permet ainsi de le désambiguïser. Seules les classes mères seront annotées avec une ou plusieurs dimensions, leurs sous-classes héritent alors de ces propriétés.

9. http ://kugel.bv.tu-berlin.de/typo3-igg/index.php ?id=1524
10. http ://www.urbamet.com/thesaurus/thesaurusurbamet.htm
11. http ://www.fao.org/agrovoc/

L'utilisateur choisit ensuite, lors de l'ajout d'un document le ou les domaines auxquels il peut appartenir. TOK aligne les différentes parties des ressources concernées par ces dimensions et/ou domaines choisis et renvoie au système une nouvelle ressource sous forme d'une liste de concepts utilisant le modèle externe à TOK décrit dans la figure 6.

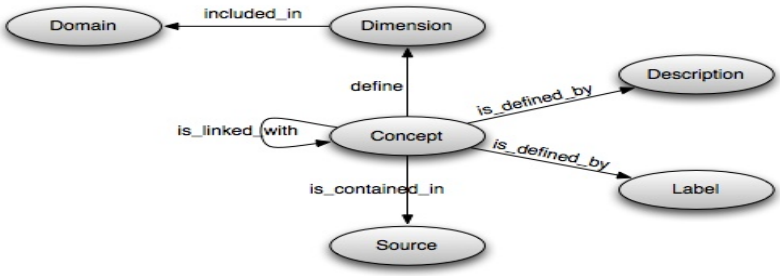

Figure 6. *Modèle de la ressource intégrée pour l'annotation des documents*

Enfin, l'utilisateur annote son document en sélectionnant les concepts qui le décrivent le mieux, parmi ceux retournés par TOK.

5. Conclusion

Notre principal objectif est de démontrer, à travers les deux cas d'utilisation, les différentes possibilités du système TOK.

Dans le premier cas, nous avions besoin d'aligner sur notre modèle les différentes métadonnées des corpus importés dans TOK, permettant ainsi d'homogénéiser les données bibliographiques, dans le but de produire un multi-corpus. Dans le deuxième cas, tout en gardant le stockage des ressources indépendant de TOK, nous avons aligné les différentes ressources lexicales en vue d'une aide à l'annotation de documents.

La nouveauté par rapport aux systèmes existants est de pouvoir intégrer des ressources hétérogènes dans un même système avec une flexibilité dans la représentation, et d'offrir une palette d'opérateurs permettant la gestion et la combinaison de ces ressources. Le stockage des ressources dans l'entrepôt TOK est effectuée sous forme d'une base de données RDF et l'interrogation est assurée par le langage SPARQL.

Une prochaine étape du travail consiste à définir des règles et des axiomes permettant d'associer à chaque tâche l'ensemble des ressources à utiliser, la représentation correspondante, les opérateurs disponibles ou la combinaison des opérateurs nécessaire à cette tâche. Pour assurer la réalisation de cette perspective nous devons encore : (i) définir un modèle pour chaque tâche de traitement de connaissances utilisant les ressources TOK ; (ii) définir et appliquer un ensemble d'heuristiques pour la déduction des correspondances afin de construire des alignements entre ces ressources.

6. Bibliographie

Amann B., Fundulaki I., Scholl M., « Integrating ontologies and thesauri for RDF schema creation and metadata querying », *International Journal on Digital Libraries*, vol. 3, p. 221-236, 2000.

D'Aquin M., Schlicht A., Stuckenschmidt H., Sabou M., « Criteria and Evaluation for Ontology Modularization Techniques », p. 67-89, 2009.

de Ribaupierre H., Falquet G., « Recherche d'information dans des corpus scientifiques basée sur la structure du discours », 2010.

Euzenat J., Shvaiko P., *Ontology Matching*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2007.

Euzenat J., Valtchev P., « An integrative proximity measure for ontology alignment », *Proc. ISWC-2003 workshop on semantic information integration, Sanibel Island (FL US)*, p. 33-38, 2003.

Falquet G., Jiang C.-L. M., Guyot J., « Un modèle et une algèbre pour les systèmes de gestion d'ontologies », *EGC*, p. 697-702, 2008.

Ghoula N., Falquet G., Guyot J., « TOK : A Meta-model and Ontology for Heterogeneous Terminological, Linguistic and Ontological Knowledge Resources », *Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM International Conference on*, vol. 1, p. 297-301, August, 2010.

Hendler J., Golbeck J., « Metcalfe's law, Web 2.0, and the Semantic Web », *Web Semant.*, vol. 6, p. 14-20, February, 2008.

Kalfoglou Y., Schorlemmer M., « Ontology mapping : the state of the art », *Knowl. Eng. Rev.*, vol. 18, n° 1, p. 1-31, 2003.

Klein M., « Combining and relating ontologies : an analysis of problems and solutions », *in* A. Gomez-Perez, M. Gruninger, H. Stuckenschmidt, M. Uschold (eds), *Workshop on Ontologies and Information Sharing, IJCAI'01*, Seattle, USA, 2001.

Lacasta J., Nogueras-Iso J., Zarazaga-Soria F., Muro-Medrano P., « Generating an urban domain ontology through the merging of cross-domain lexical ontologies », *Ontologies for urban development : conceptual models for practitioners*, COST, p. 69-83, October, 2007.

Noy N. F., Musen M. A., « Anchor-PROMPT : Using Non-Local Context for Semantic Matching », *Workshop on Ontologies and Information Sharing at the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI-2001)*, Seattle, WA, 2001.

Noy N. F., Musen M. A., « Specifying Ontology Views by Traversal », *The Semantic Web – ISWC 2004*, p. 713-725, 2004.

Pinto H. S., Martins J. P., « A methodology for ontology integration », *K-CAP'01 :Proceedings of the 1st international conference on Knowledge capture*, ACM, New York, NY, USA, p. 131-138, 2001.

Radhouani S., Un modèle de Recherche d'Information orienté précision fondé sur les dimensions de domaine, PhD thesis, University of Geneva, University Joseph Fourier, 2008.

Sindt T., « Formal Operations for Ontology Evolution », *procedings of the International Conference on Emerging Technologies*, 2003.

Stumme G., Maedche A., « FCA-MERGE : Bottom-Up Merging of Ontologies », *IJCAI*, p. 225-234, 2001.