Une ontologie OWL pour le CDM-fr

Nicolas Delestre, Nicolas Malandain, Boulares Ouchenne

Une ontologie OWL pour le CDM-fr

Application à la Liaison entre Offres de FormAtion et Ressources pédagogiques

Nicolas Delestre, Nicolas Malandain et Boulares Ouchenne

LITIS, Normandie Université, INSA Rouen
Avenue de l'Université, 76801 Saint-Étienne-du-Rouvray, France
{nicolas.delestre,nicolas.malandain,boulares.ouchenne}@insa-rouen.fr

Résumé Le projet SemUNIT propose une version web sémantique du schéma de métadonnées SupLOMfr. Ce travail n'a pas été réalisé pour le CDM-fr, le schéma de métadonnées décrivant les parcours pédagogiques. Dans la première partie de l'article, nous élaborons une version web sémantique du CDM-fr en réutilisant au maximum des ontologies OWL ou schémas RDF préexistants. L'exposition de ces données ouvertes, dorénavant structurées, permet la création de nouveaux services dans le domaine des TICE. Ceci est l'objet de la deuxième partie de l'article qui présente un moteur de recherche permettant de trouver des ressources compatibles avec un cours ou un parcours pédagogique.

Mots-clé: Web Sémantique, données ouvertes structurées, parcours pédagogiques, CDM-fr, ressources pédagogiques, SupLOMfr

1 Introduction

Grâce aux travaux de Tim Berners-Lee, nous pouvons consulter facilement des documents multimédia indépendemment de leur position géographique. Toutefois, au regard du nombre exponentielle de documents disponibles sur le web, il est nécessaire pour les consulter, de les retrouver et donc au préalable de les indexer. Il y a eu des moteurs de recherche basés sur une indexation manuelle, mais ceux que nous utilisons aujourd'hui effectuent tous une indexation automatique. Malheureusement les documents du Web sont pour la plupart des documents qui nécessitent une interprétation « intelligente » pour être convenablement indexés (par exemple pour les textes, il faut gérer l'ambiguïté des langues naturelles, pour les images, il faut associer un sens aux formes, etc.), tâche qui ne peut pas être effectuée par les robots des moteurs de recherche. Dès la deuxième partie des années 90, le W3C définit le Web pour les machines : le web des données et le web sémantique. Les mots clés de cette évolution sont les métadonnées, leur représentation (triplets RDF), un langage formel d'expression (RDF) et la réutilisation de schémas de métadonnées (décrites en RDFS ou OWL) [3].

Dès le début des années 2000, dans le domaine de la pédagogie, des travaux ont été initiés pour élaborer des normes ou standards de schémas de métadonnées décrivant d'une part les ressources (par exemple LOM, LOMfr et SupLOMfr) et

d'autre part les parcours avec le CDM et sa déclinaison française, le CDM-fr. Toutefois, à l'exception du projet SemUnit [4] qui a proposé une version web sémantique du schéma de métadonnées SupLOMfr (que l'on va nommé dans cet article SemUNT), les autres schémas ne sont pas modélisés et représentés pour le Web Semantique. C'est par exemple le cas pour le schéma de métadonnées CDM-fr. La figure 1 présente le cadre de notre travail : partir des schémas de métadonnées pour créer des schémas OWL et les instancier. Dans notre cas nous utilisons SemUNIT pour les ressources indexées d'UNIT [1] et nous créons SemINSA pour les cours du département ASI de l'INSA de Rouen. Les zones grisées correspondent aux travaux décrits dans cet article.

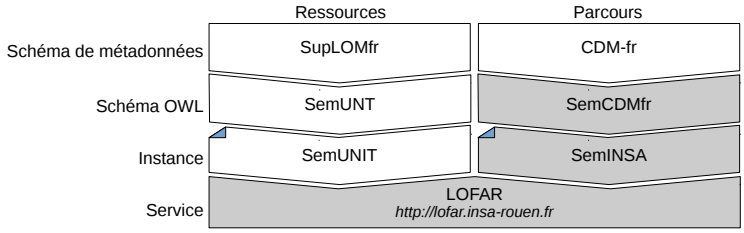

Figure 1. Positionnement des modèles et du service LOFAR

Après une description de l'existant, nous allons proposer un schéma OWL pour le CDM-fr. Ensuite nous présenterons une application, le projet LOFAR (Liaison entre Offres de FormAtion et Ressources pédagogiques) qui va tirer parti d'une part de ces descriptions de parcours pédagogiques et d'autre part des descriptions de ressources pédagogiques proposées par le serveur *SemUNIT*. Enfin, après une validation expérimentale, nous présenterons l'architecture logiciel de notre prototype, librement accessible via le Web.

2 Le Schéma de Métadonnées CDM-fr

Début des années 2000, la Norvège a proposé un schéma de métadonnées pour décrire les établissements d'enseignements supérieurs et les parcours proposés par ces derniers (CDM pour *Course Description Metadata*). En 2004 la SDTICE [2] a constitué un groupe de travail qui a proposé en 2005 un profil d'application français de ce schéma : le CDM-fr (Cf.[8]).

Un conteneur d'information CDM-fr regroupe l'ensemble des informations relatives à l'enseignement supérieur. Il est constitué de descriptions d'entités organisationnelles, de personnes, de programmes d'études et de cours. L'implantation technique de CDM-fr est structurée sous la forme d'un schéma XML comportant quatre éléments constitutifs principaux :

1. Université Numérique Ingénierie et Technologie, `http://www.unit.eu/`
2. Sous-direction des technologies de l'information et de la communication pour l'éducation

1. **OrgUnit :** contient la description et les coordonnées des établissements d'enseignement ou des composantes responsables de l'organisation et du déroulement des programmes d'études et des cours.
2. **Person :** contient la description et les coordonnées des acteurs qui interviennent dans la gestion, l'organisation et le déroulement des programmes d'études et/ou des cours.
3. **Program :** contient les unités d'enseignement préparant à un examen ou aboutissant à un diplôme, un titre, une qualification, une certification.
4. **Course :** contient les informations relatives à une unité d'enseignement (nom, nombre de crédits, prérequis, etc).

L'étude de ce schéma de métadonnées met en exergue plusieurs problèmes, dont les deux principaux sont d'une part l'absence de modèle conceptuel de données (seul un schéma XML est proposé) et d'autre part la mauvaise formalisation de certaines informations, via l'utilisation des éléments XML `infoBlock` et `subBlock`. Ces éléments peuvent être inclus à tous les éléments principaux (`OrgUnit`, `Person`, `Program` et `Course`) et leurs descendants. Les éléments `infoBlock` et `subBlock` permettent d'inclure des informations non formalisées, en langue naturelle (multilingue) qui peuvent de plus être mises en forme (utilisation de balises provenant du HTML). Or ces éléments sont très utilisés. Par exemple un module du logiciel Scolpédagogie de la suite Coktail (système d'information universitaires) utilise abondamment ces deux éléments, excluant de fait un post traitement efficace des fichiers XML produits. La figure 2 montre un exemple de code XML, où l'on voit que pour la description d'un cours, le même élément XML `subBlock` est utilisé pour indiquer l'URL des ressources proposées par ce cours, la langue utilisée pour dispenser ce cours et enfin les heures de cours magistraux ou de travaux dirigés. Ces informations sont ici proposées en anglais, mais dans le même fichier XML on peut les retrouver en français.

```
<formOfTeaching>
  <subBlock>
  site web : http://XXXXX/course/view.php?id=68 </subBlock>
  <altLangBlock language="en-gb">
    <subBlock>
    Lectures : 21h </subBlock>
    <subBlock>
    Exercises : 42h</subBlock>
    <subBlock>
    language : french</subBlock>
    <subBlock>
    http://XXXX/course/view.php?id=68 </subBlock>
  </altLangBlock>
</formOfTeaching>
```

Figure 2. Exemple de code XML produit par le logiciel *Scolpédagogie*

On en déduit que les données contenues dans les XML CDM et CDM-fr sont moins structurées que les données stockées dans le système d'information :

les utiliser comme source pour des outils d'indexation et de recherche serait dès lors moins performant. Dans la pratique ces fichiers XML sont uniquement utilisés en entrée de feuilles XSLT pour générer des pages HTML des sites Web d'établissement, présentant l'offre de formation [3].

C'est à la suite de ce constat que nous avons décidé de proposer un schéma de métadonnées de description de parcours pédagogiques compatible avec les modèles et outils du Web Sémantique.

3 SemCDMfr : un modèle OWL pour le schéma de métadonnées CDM-fr

Pour définir le modèle OWL de ce schéma de métadonnées CDM-fr, nous sommes partis des schémas XML (XSD) existant tout en appliquant les bonnes pratiques proposées dans [1]. Nous avons suivi les étapes suivantes :

Étape 1 Création d'un modèle conceptuel de données à partir des schémas XML. L'objectif est de partir des éléments et attributs XML pour en extraire les concepts et relations formant le modèle. Cependant, nous avons rencontré beaucoup de difficultés pour interpréter convenablement certains éléments XML. En effet, le peu de documentation et la redondance d'information dans certains éléments ont compliqué énormément notre travail. En outre, comme nous le signalions précédemment, une grande partie des informations n'est pas formalisée et apparaît dans le contenu d'éléments de type `infoBlock` (un mélange de texte avec des informations semi-structurées et non structurées).

Étape 2 Identification de schémas RDF ou OWL pré existants qui proposeraient des classes et propriétés équivalentes à celles identifiées à l'étape précédente. Afin de proposer une ontologie qui soit le plus interopérable, nous avons retenu des ontologies identifiées par le W3C [4] ou fréquemment utilisées, et donc référencées dans des entrepôts d'ontologies, comme par exemple *Linked Open Vocabularies* [5]. Ainsi nous avons utilisé les schémas suivants :

- **foaf** ce schéma permet de décrire une organisation et/ou une personne et ses relations (`http://xmlns.com/foaf/0.1/`).
- **vcard** ce schéma permet de décrire les coordonnées d'une personne ou d'une entité organisationnelle (`http://www.w3.org/TR/vcard-rdf/`).
- **aiiso** ce schéma permet de décrire la structure organisationnelle interne d'une institution académique (`http://vocab.org/aiiso/schema`).

Étape 3 Enrichissement de l'ontologie en ajoutant des caractéristiques à certaines relations, comme par exemple le fait qu'une propriété soit transitive, réflexive, symétrique, ...

3. C'est le cas pour l'INSA de Rouen : `http://formations.insa-rouen.fr/cdm/`
4. `http://www.w3.org/wiki/Good_Ontologies`
5. `http://lov.okfn.org/dataset/lov/`

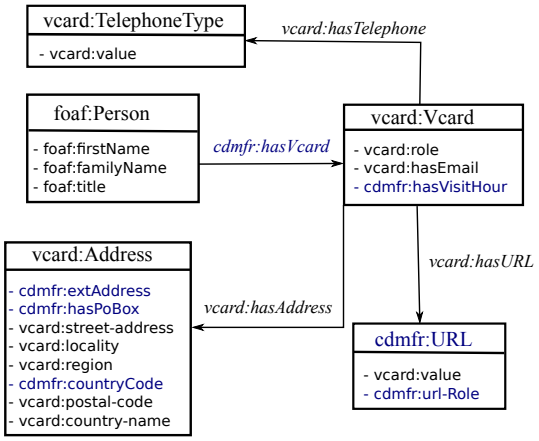

Figure 3. Modèle OWL de l'élément `Person`.

Afin d'être plus lisible, à l'image du schéma CDM-fr initial, notre ontologie peut être décomposée en quatre parties : les personnes, les établissements ou organismes de formation, les parcours et enfin les cours. Dans la suite de cet article, chaque partie est présentée par un diagramme de classes UML, tel que les classes UML représentent des classes OWL, les attributs de classe UML représentent des *OWL data properties* (les types de données n'ont pas été ajoutés afin de ne pas alourdir la présentation) et les relations d'association représentent des *OWL object properties*. Chaque identifiant est préfixé par un nom de domaine. Seuls ceux préfixés par `cdmfr` (en bleu) sont réellement décrits dans notre schéma, les autres sont des réutilisations des trois schémas présentés précédemment. Enfin, lorsque deux identifiants d'*OWL object properties* sont associés à une relation bidirectionnelle, cela signifie que ces deux propriétés sont symétriques (très souvent l'un commençant par `is` et l'autre par `has`).

3.1 Partie décrivant les personnes

La classe `Person` permet de décrire les différentes personnes intervenant dans le cycle d'apprentissage. La figure 3 représente le schéma OWL de cette première partie de l'ontologie. On peut constater que nous avons réutilisé au maximum les schémas `foaf` et `vcard`, nous avons juste ajouté six propriétés et une classe.

3.2 Partie décrivant l'établissement

La classe `Organization` est la classe centrale de l'ontologie qui décrit une entité organisationnelle, celle qui gère ou propose des unités d'enseignement et des programmes d'études. La structure d'une entité organisationnelle peut être de type hiérarchique avec des entités organisationnelles subordonnées (par exemple pour les universités : les facultés), d'où les relations `is-sub-unit` et

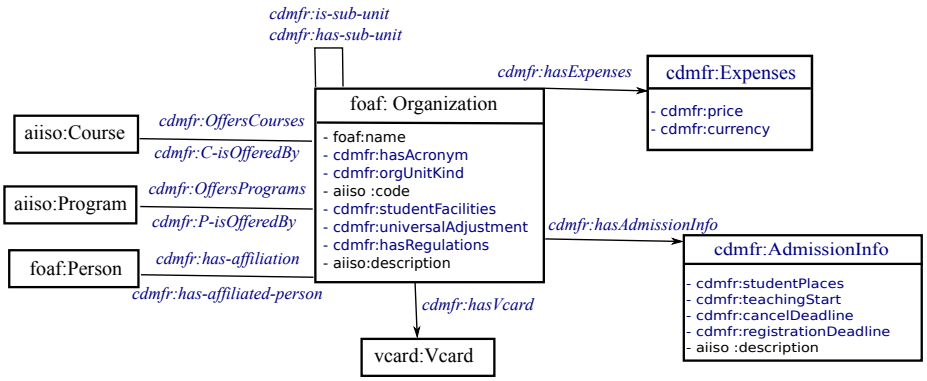

Figure 4. Modèle OWL de l'élément `OrgUnit`.

`has-sub-unit`. Elle a été conçue pour intégrer toutes les structures proposant des programmes d'enseignement (université, écoles d'ingénieur, etc). La figure 4 présente ce modèle. On constate que nous avons réutilisé ici le schéma `aiiso`, et qu'en comparaison de la partie précédente, nous avons ajouté beaucoup plus de classes et de propriétés.

3.3 Partie décrivant les parcours

Un parcours pédagogique est décrit par la classe `Programme`. Deux types d'informations sont associés à cet élément. Tout d'abord il y a des informations administratives, comme le lieu du parcours, la durée, la date de début, le nombre de crédits ECTS obtenus après validation, etc. Ensuite il y a des informations pédagogiques, par exemple les cours du parcours, les parcours ou cours prérequis, la forme pédagogique (en présentiel ou à distance), etc. Un parcours pédagogique peut être divisé en d'autres parcours pédagogiques. La figure 5 représente le schéma OWL de cette partie.

3.4 Partie décrivant les cours

Enfin la classe `Course` contient les informations relatives à un cours. Encore une fois, les informations le caractérisant sont des informations administratives ou pédagogiques. La figure 6 synthétise l'ensemble de ces informations.

3.5 Conclusion

L'ensemble de ces classes et propriétés OWL forment l'ontologie SemCDMfr. Toutes les informations présentes dans le schéma de métadonnées CDM-fr initial sont aussi présentes dans cette ontologie.

Un établissement d'enseignement supérieur exposant une instantiation de cette ontologie permettrait à toute personne ou organisation d'en tirer partie et

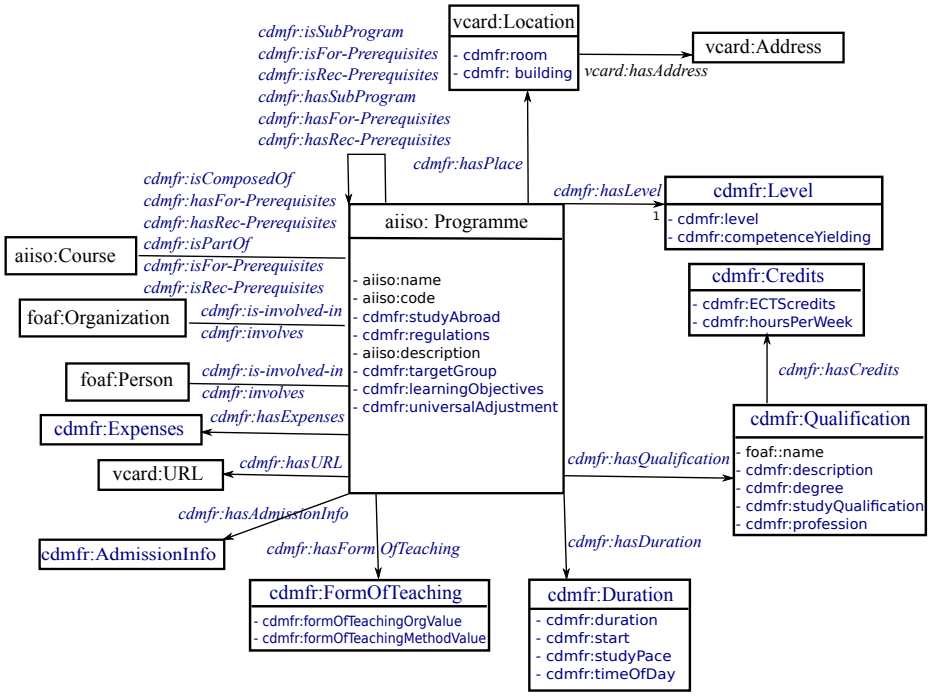

Figure 5. Modèle OWL de l'élément `Program`.

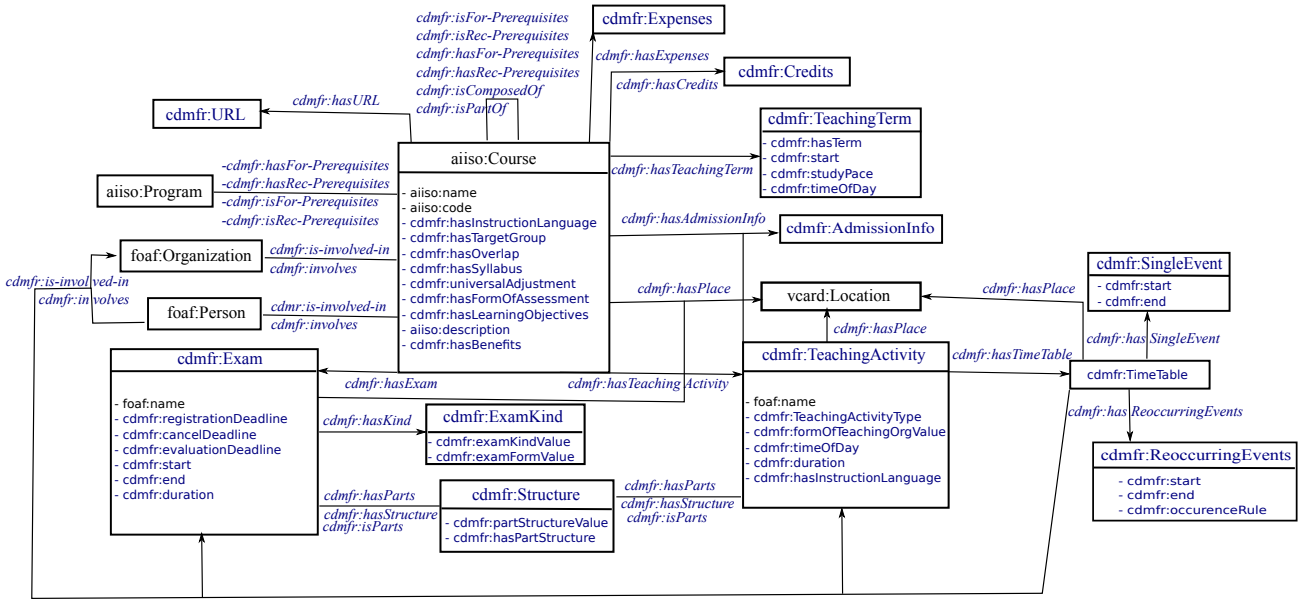

Figure 6. Modèle OWL de l'élément `Course`.

de proposer des services innovants utilisant ces informations. C'est ce que nous nous allons présenter maintenant à travers un exemple de service : le serveur LOFAR.

4 Liaison entre offres et ressources pédagogiques

On se propose dans cette section de concevoir une application, nommée LOFAR, pour Liaison entre OFfres de formAtion et Ressources pédagogiques, qui utiliserait une instance de l'ontologie SemUNT et une instance de SemCDMfr.

4.1 Cas d'utilisation de l'application LOFAR

Avant de s'intéresser à la conception de cette application, identifions quelques cas d'utilisation.

Le premier cas est celui d'un étudiant qui recherche des ressources pour un parcours auquel il vient de s'inscrire. Il se connecte sur LOFAR, il saisit ou choisit l'URL de l'établissement exposant ses informations au format SemCDMfr. Le serveur LOFAR cherche alors les ressources UNIT compatibles avec les cours de ce parcours.

Un deuxième cas d'utilisation envisageable est celui d'un enseignant reprenant le cours d'un collègue. Si l'enseignant n'est pas spécialiste du dit cours, il devra étudier les supports que lui donnera son collègue, mais afin d'avoir un point de vue plus large, il devra aussi étudier d'autres ressources sur ce même sujet. Il pourra alors utiliser LOFAR pour les obtenir.

4.2 LOFAR : un moteur de recherche d'information

Ces deux cas d'utilisation montrent que le serveur LOFAR est un moteur de recherche d'information (RI) qui utilise conjointement deux ontologies, l'une décrivant des ressources pédagogiques (SemUNT) et l'autre décrivant des parcours pédagogiques (SemCDMfr). Ainsi pour un cours donné, décrit à l'aide de SemCDMfr, le système doit proposer les meilleures ressources, décrites à l'aide de SemUNT.

La constitution d'un tel classement nécessite tout d'abord d'identifier les métadonnées des cours et ressources qui sont utiles à cet ordonnancement. Pour chaque couple de métadonnées retenu, il est nécessaire d'établir une mesure de similarité (MS_{m_i}) permettant d'estimer la proximité sémantique de leurs instances. Enfin par combinaison de ces similarités entre métadonnées, on veut définir une mesure de similarité (MS_{CR}) entre cours et ressource. Ainsi, une ressource R_i sera mieux classée qu'une ressource R_j pour un cours C donné, si et seulement si $MS_{CR}(C, R_i) > MS_{CR}(C, R_j)$.

La mesure de similarité MS_{CR} entre un cours C et une ressource R utilisant k couples de métadonnées (m_C, m_R) peut alors être décrite par :

$$MS_{CR}(C, R) = \sum_{i=1}^{k} \alpha_i MS_{m_i}(m_{C_i}, m_{R_i})$$

telles que toutes les mesures de similarité sont comprises entre 0 (pour des données totalement orthogonales) et 1 (pour des données totalement similaires) et la somme des α_i vaut 1.

Après l'étude de notre ontologie SemCDMfr et de l'ontologie SemUNT, nous avons pour l'instant retenu deux couples de métadonnées :

1. le titre des cours et des ressources (`aiiso:Name` et `dc:title`) ;
2. leurs descriptions (`aiiso:Description` et `dc:description`).

De ce fait, en notant m_{Ct} et m_{Rt} les valeurs des métadonnées pour les titres (cours et ressources), et m_{Cd} et m_{Rd} les valeurs des métadonnées pour les description, MS_{CR} peut maintenant s'écrire :

$$MS_{CR}(C,R) = \alpha_t MS_t(m_{Ct}(C), m_{Rt}(R)) + \alpha_d MS_d(m_{Cd}(C), m_{Rd}(R))$$

Les métadonnées décrivant les cours et les ressources peuvent être de différents types. Si l'on se situait dans un web sémantique idéal, on disposerait des réseaux sémantiques décrivant les valeurs de chaque métadonnées d'un cours et d'une ressource. Dans ce cas, il serait nécessaire de calculer une similarité entre deux graphes. En réalité aujourd'hui, nous ne disposons pas de représentations aussi fines. Par exemple, dans notre cas, les métadonnées « description » d'un cours ou d'une ressource sont purement textuelles. Nous allons donc devoir utiliser un algorithme de mesure de similarité entre des métadonnées textuelles afin de calculer MS_t et MS_d.

4.3 Mesures de similarité entre textes

Nous avons besoin théoriquement de deux mesures de similarité, l'une entre titres et l'autre entre descriptions. Toutefois les valeurs de ces métadonnées étant de même type (textuelles), bien que de natures différentes (les titres sont plus concis que les descriptions) nous faisons le choix d'une unique mesure de similarité.

Deux grandes catégories d'algorithmes existent : ceux représentant les textes par des ensembles de mots et ceux représentant les textes par des vecteurs numériques. Une fois la représentation choisie, on peut choisir l'algorithme de calcul de similarité comme par exemple la distance cosinus, la distance de Jaccard, le coefficient de Dice, etc. (Cf. [5], p299).

Avant chaque algorithme, les textes peuvent subir un pré traitement, comme celui de la lématisation ou de la racinisation. La lématisation consiste à remplacer les pluriels par leurs singuliers, les verbes conjugués par leurs infinitifs, etc. La racinisation quant à elle, remplace les mots par leurs racines, soit par troncature fixe (indépendemment de la langue), soit par l'utilisation d'algorithmes plus complexes, fonction de la langue, comme par exemple *PORTER* pour l'anglais ou *CARRY* [6] pour le français.

Il est à noter qu'il existe aussi un troisième type de prétraitement, nommé $n-$grammes qui consiste à mettre le texte en minuscule, puis à enlever tous les caractères de ponctuation, espace compris, pour enfin le découper en éléments

successifs de n caractères, chacun décalé du précédent d'un caractère. Ainsi le texte « Un exemple » sera transformé en l'ensemble de 4-grammes suivants $\{unex, nexe, exem, xemp, empl, mple\}$. n est habituellement choisi en essayant plusieurs valeurs (généralement comprises entre 2 et 5) et en retenant celle qui retourne le meilleur indicateur de performance sur un corpus contrôlé.

4.4 Choix de la mesure pour MS_t et MS_d

Afin de choisir une bonne mesure de similarité pour nos deux couples de métadonnées nous avons testé trois algorithmes différents :

— distance de Jaccard avec prétraitement n-grammes ;
— distance de Jaccard avec racinisation fixe ;
— variante de la distance cosinus, basée sur une représentation $tf.idf$ avec racinisation utilisant l'algorithme de *CARRY* [6](proposé par l'API Lucène).

Il existe plusieurs critères permettant de qualifier un algorithme de RI. Mais la plupart, comme le rappel, la précision ou l'AUC (*Area Under Curve*) nécessite d'étiqueter, pour une requête donnée, l'ensemble des documents, ce qui n'est pas envisageable lorsque la taille de la base est grande. Toutefois, il existe entre autres le critère « précision moyenne au rang k » qui permet de qualifier un algorithme de RI uniquement en fonction des k documents retournés.

La précision moyenne au rang k nécessite d'étiqueter, au regard d'une requête q, les k documents retournés avec leur valeur V_i qui vaut 0 lorsque le document i est non pertinent et 1 lorsqu'il est pertinent. Cet indicateur est alors la moyenne de ces k valeurs :

$$P_k(q) = \frac{1}{k} \sum_{i=1}^{k} V_i$$

Nos trois algorithmes ont été testés sur un corpus de treize cours scientifiques dispensés par le département Architecture des Systèmes d'Information de l'INSA de Rouen. Pour chaque cours les dix premières ressources proposées par chaque algorithme ont été étiquetées par deux enseignants comme étant « pertinentes » ou « non pertinentes ». La figure 7 présente les valeurs de précision moyenne au rang 10 pour les treize cours, telles que valeurs pour les α_t et α_d soient de $0,5$ (on attribue autant d'importance aux titres qu'aux descriptions).

Nous constatons que c'est l'algorithme proposé par l'API Lucène qui retourne en général les meilleurs classements. Il est à noter que cet algorithme ne fonctionne pas (seulement une ressource jugée pertinente) pour le seul cours « UML et design pattern » (n° 11). Après étude de la description de ce cours, c'est semble-t-il l'utilisation fréquente du mot « diagramme » qui pose problème. Au final en moyenne pour une requête donnée, 65% des documents retournés sont considérés comme pertinents avec une médiane à 70%.

5 Architecture du système

La figure 8 présente l'architecture logicielle du prototype. Nous proposons deux nouveaux services : SemINSA qui fournit un accès SPARQL à la base

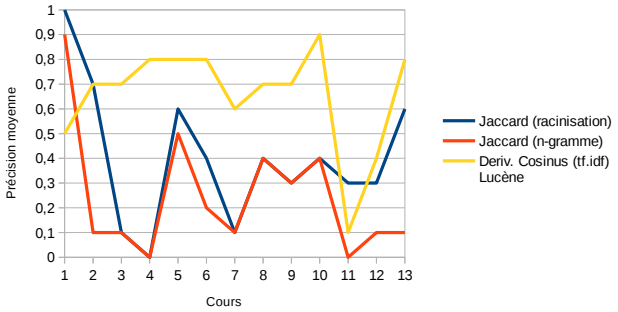

Figure 7. Précision moyenne des trois algorithmes pour les treize cours

ontologique CDM-fr du département ASI de l'INSA de Rouen, LOFAR qui propose les dix meilleures ressources UNIT pour un cours donné. Le service LOFAR (ensemble de pages JSP et de classes métiers Java) interroge part défaut (paramétrable) les services SemUNIT et SemINSA en SPARQL. Ce dernier service a été instancié à partir des données XML du Système d'Information du département ASI.

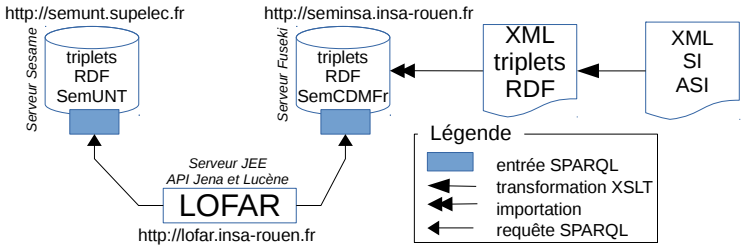

Figure 8. Architecture du prototype

6 Conclusion

Dans cet article nous avons d'une part proposé une ontologie permettant de décrire des parcours pédagogiques SemCDMfr et d'autre part proposé le service LOFAR qui utilise une instance de cette dernière ainsi que *SemUNIT* pour proposer des ressources au regard d'un cours. Les résultats de notre expérimentation sont de nature à valider notre approche même s'ils pourraient être encore améliorés.

Une première amélioration serait de fixer automatiquement l'importance des deux appariements de métadonnées utilisées pour construire la mesure de similarité MS_{CR}. On pourrait ainsi faire varier les coefficients α_i entre 0 et 1 (tel que $\sum \alpha_i = 1$) et choisir le couple de valeur qui donne un meilleur indicateur.

Une deuxième perspective serait de tester d'autres similarités entre textes. En effet dans certains cas, la mesure que nous utilisons n'est pas performante (par exemple lorsque les métadonnées textuelles contiennent des mots fortement polysémiques). Nous devrions tester et comparer d'autres représentations de texte, comme par exemple *lsa* [2], voire tester des mesures de similarité conçues spécialement pour des textes courts, comme par exemple [7].

Une troisième amélioration serait d'utiliser d'autres métadonnées, lors de la recherche (par exemple celles décrivant le niveau d'étude), cela aurait toutefois comme inconvénient d'augmenter le nombre de paramètres α_i de MS_{CR}. Ou encore lors de la restitution, il serait possible de structurer les résultats.

Enfin nous prévoyons d'étudier plus en détail les schémas des bases de données utilisés dans les deux logiciels les plus utilisés par les établissements de l'enseignement supérieur français (les suites de logiciels *Cocktail* et de l'*AMUE*) ainsi que les usages qui en sont faits, de façon à concevoir des traducteurs automatiques ou semi-automatiques des données issues de ces systèmes d'information vers notre ontologie.

Nous remercions la fondation UNIT pour avoir en partie financé ce travail.

Références

1. Jean Charlet, Bruno Bachimont, and Raphaël Troncy. Ontologies pour le web sémantique. *Revue I3*, page 31p, 2004.
2. Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6) :391–407, 1990.
3. Fabien Gandon, Catherine Faron-Zucker, and Olivier Corby. *Le WEB sémantique, comment lier les données et les schémas sur le web.* Dunod, 2012.
4. Yoann Isaac, Yolaine Bourda, and Monique Grandbastien. SemUNIT - French UNT and Linked Data. In *Proceedings of the 2nd International Workshop on Learning and Education with the Web of Data*, volume 840, page 6 pages, Lyon, France, 2012. CEUR workshop proceedings.
5. Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing.* MIT Press, Cambridge, MA, USA, 1999.
6. M. Paternostre, P. Francq, J. Lamoral, D. Wartel, and M. Saerens. Carry, un algorithme de désuffixation pour le français. Technical report, Paul Otlet Institute, 2002.
7. Mehran Sahami and Timothy D. Heilman. A web-based kernel function for measuring the similarity of short text snippets. In *Proceedings of the 15th International Conference on World Wide Web*, WWW '06, pages 377–386. ACM, 2006.
8. SDTICE. Spécification des métadonnées de description de cours (cdm). Technical report, MENESR, 2004.