

Peuplement automatisé d'ontologies par analyse des programmes scolaires

Mahdi Gueffaz¹, Jirasri Deslis¹, Jean-Claude Moissinac¹

¹ Institut Mines-Télécom; Télécom ParisTech;

CNRS LTCI

46, rue Barrault

75634, Paris Cedex 13

{mahdi.gueffaz, jirasri.deslis, jean-claude.moissinac}@telecom-paristech.fr

Résumé. La construction d'ontologies et l'annotation documentaire sont des traitements très coûteux en temps et en ressources. Plusieurs travaux cherchent à mettre en place des solutions basées sur l'utilisation d'outils linguistiques pour extraire semi automatiquement ou automatiquement les informations pertinentes. Le volume d'information des programmes scolaires est important et une assistance à leur transformation en une ontologie apparaît nécessaire. Le but de nos travaux de recherche est le développement d'outils de création et d'enrichissement d'ontologies avec une assistance semi-automatique. Dans le cadre du projet ILOT¹ (Innovative Learning Object for Teaching) nous proposons une approche composée de trois phases. Une phase de peuplement d'ontologie, une phase d'enrichissement avec des ontologies propres et des ontologies extérieures et une dernière phase pour l'exploitation de données contenues dans l'ontologie dans le cadre des enseignements. Cet article porte principalement sur la première phase.

1 Introduction

Les enseignants sont amenés à manipuler une grande quantité d'informations numériques (texte, documents multimédias, documents composites). Le web sémantique peut faciliter les démarches d'apprentissage en aidant à faire face à la multiplicité et à la complexité des données à traiter. Il offre une extension au web actuel, pour que l'accès aux données pertinentes soit facilité par des automatismes. Il organise et structure lénorme quantité d'informations contenues dans le web.

Les ontologies sont utilisées pour formaliser la connaissance dans le Web sémantique. Elles sont représentées par un langage qui permet de spécifier une conceptualisation, définie comme une version simplifiée d'un domaine que nous voulons représenter de façon formelle en utilisant des concepts et leurs relations [1]. Un des objectifs des ontologies est la facilitation des échanges de connaissances entre les humains, entre humain et machine ou entre machines.

Dans le cadre du projet ILOT, nous avons décidé de travailler sur la création d'ontologies à partir de corpus de ressources pédagogiques. Dans un premier temps,

¹ <http://ilot.wp.mines-telecom.fr/>

nous avons élaboré une méthodologie en travaillant à la création d'ontologies pour la représentation des programmes scolaires français.

Notre corpus est composé d'une centaine de documents (d'une trentaine de pages par matière) et sa transformation en une ontologie manuellement serait une activité difficile qui nécessiterait des compétences à la fois dans la représentation des connaissances avec les ontologies et des connaissances sur les domaines couverts (pédagogie, histoire, mathématiques...etc.).

Dans cet article, nous proposons une méthode de création et d'enrichissement d'ontologie semi-automatique afin d'enrichir les possibilités de l'enseignant en s'appuyant sur le contenu des programmes scolaires. Dans la section 2, nous présentons les travaux relatifs à cette méthode. La section 3 donne une description du corpus utilisé pour la création de nos ontologies. Nous décrirons dans la section 4 la mise en place de notre solution. Ensuite, nous exposons les résultats de nos premières expérimentations. Enfin, nous commentons ces résultats afin d'apporter des améliorations dans les travaux futurs.

2 Travaux existants

Nous trouvons des travaux voisins dans les domaines de la formation, en particulier la formation en ligne. Plusieurs travaux de recherche ont été identifiés dans le domaine du e-learning exploitant des ontologies. Ils contribuent à confirmer l'intérêt de modéliser un champ de formation pour améliorer les outils de cette formation. Dans cette catégorie, citons : le projet Web Sémantique et E-Learning [2] [3] [4] qui propose des concepts, des méthodes et des outils sur les environnements informatiques pour l'apprentissage humain; le projet de développement de l'environnement de conception de curriculum et de cours présenté dans [5] utilise quatre ontologies constituant la base du curriculum qui exploite une ontologie de capacités, une ontologie d'objectifs, une ontologie de ressources, et une ontologie de liens.

Face à la masse croissante d'informations numériques exploitées, des méthodes automatiques de conception d'ontologie ont été proposées. Différents types d'approches sont distingués selon les types de données en entrée : à partir de texte, de dictionnaires, de base de connaissances, de schémas semi structurés et de schémas relationnels. [6].

L'application Text-To-Onto [7], développée à l'Institut AIFB de l'Université de Karlsruhe, sert à extraire à partir de corpus ou de documents Web des données pour la conception d'ontologie et permet également la réutilisation d'ontologies existantes [8]. OntoBuilder [9] permet de construire une ontologie à partir de ressources Web. L'extraction de l'ontologie à partir de fichiers XML est suivie d'une phase de raffinement guidée par l'utilisateur.

Notre corpus des programmes scolaires français est composé de fichiers XML dont nous voulons obtenir une représentation sous forme d'une ou plusieurs ontologies. Les travaux précédents ont été une source d'inspiration pour notre méthodologie. Dans la littérature, nous avons aussi trouvé des travaux de recherche qui proposent des outils de transformation de fichiers XML en ontologies.

Les travaux de [10] proposent une approche de construction d'ontologie à partir d'un schéma XML et transforment le document XML en graphe RDF. Dans [11] est proposée une approche similaire pour la création d'ontologie OWL à partir de schémas XML. Dans cette approche, les classes OWL sont définies à partir des types complexes du schéma XSD, c'est-à-dire les éléments définis dans le XSD qui contiennent d'autres éléments ou ont au moins un attribut. Quand un élément contient un autre élément, une propriété d'objet (ObjectProperty) est créée entre les classes OWL correspondantes. Les propriétés de type de données (DataTypeProperty) sont définies à partir des attributs XML et à partir des éléments contenant seulement un littéral et pas d'attribut.

[12] proposent un outil X2OWL de transformation de documents XML en ontologie OWL. L'ontologie générée ne contient que les concepts et les liens entre concepts (ObjectProperty). Cette méthode est basée sur le schéma du document XML pour générer la structure de l'ontologie. Cette méthode inclut aussi une étape de raffinement permettant la restructuration de l'ontologie.

L'ensemble de ces travaux ont servis de base à la mise au point de la méthode que nous décrivons dans la suite de cet article.

3 Description du corpus

L'utilisation du programme scolaire dans notre démarche destinée à indexer des ressources éducatives s'est appuyée sur des expériences menées avec des utilisateurs et des résultats de recherche de projets européens dans le domaine de l'éducation, qui nous ont confortés dans l'idée d'utiliser des ontologies.

En premier lieu, citons le retour des expériences du portail Learning Resource Exchange for schools, dans le cadre du projet ASPECT. Cette expérimentation, dont un des objectifs est de vérifier comment les enseignants cherchent et trouvent des ressources, a été effectuée auprès des 44 enseignants venant de plusieurs pays européens [13]. Elle démontre en effet que « 85 % des enseignants définissent la qualité des ressources comme la correspondance entre le contenu et le programme éducatif qu'ils traitent. L'indexation par point de programme améliore la pertinence des résultats de recherche ». [14]. Cela conforte notre proposition de s'appuyer sur nos ontologies pour indexer, en phase de création, les ressources produites dans la plateforme ILOT.

Par ailleurs, nous pouvons également nous référer aux expérimentations du projet européen Intergeo, dédié à éliminer les freins à l'adoption de la géométrie dynamique par les enseignants et à l'utilisation des ressources existantes à travers l'Europe. L'ontologie des compétences GeoSkills [15] dont les informations proviennent de l'extraction des informations du programme scolaire dans le domaine de la géométrie a été ainsi mise en place afin de résoudre les barrières de la langue et obtenir un référencement adapté aux pratiques professionnelles des enseignants.

Notre corpus d'expérimentation est constitué des programmes scolaires de toutes les matières au niveau Collège et Lycée, mis à disposition par le Ministère de l'Education

nationale sur le portail *eduscol*². Le corpus contient des documents PDF et des documents HTML que nous avons transformés manuellement en fichiers XML. Ces programmes ont été conçus comme la trame de la pédagogie pour les enseignants. Ces programmes, consultés par tous les enseignants, deviennent ainsi un langage commun propre de la communauté.

Le corpus contient plusieurs matières à enseigner : Histoire-géographie-éducation civique, Arts plastiques, Education musicale, Education physique et sportive, Langues et cultures de l'Antiquité, Langues vivantes, Français, Histoire des Arts, Mathématiques, Physique-chimie, Sciences de la vie et de la Terre, Technologie. Quantitativement, le corpus est composé de 62 documents en format pdf et en html. Dans l'ensemble, les documents comprennent deux grandes parties. La partie introductory contenant les objectifs et la mise en œuvre des programmes et la partie des contenus de l'enseignement.

Cette dernière partie contient les entités pédagogiques (thème, démarche, capacité, description, etc.) qui seront extraits pour la construction des ontologies des programmes scolaires. Certains programmes comme Histoire, Mathématiques, Sciences de la vie et de la Terre, Technologie ont une représentation semi-structurée sous forme de tables contenant ces informations. Cette mise en forme nous aide à déterminer les relations et/ou définir la hiérarchisation des entités pédagogiques entre elles.

4 Approche

[16] définit le cycle de vie de la génération automatique d'ontologie comme un processus composé de cinq étapes :

- Extraction : fournit les informations à partir du corpus pour constituer l'ontologie ; dans notre processus, une extraction semi-automatique des textes des fichiers PDF a permis une première structuration en XML ;
- Analyse : en partant des résultats de la première étape, cette étape utilise l'analyse morphologique ou lexicale, l'analyse sémantique pour détecter les synonymes, les homonymes et d'autres relations de ce type; de telles techniques nous permettent d'annoter le fichier XML pour en distinguer certaines parties ;
- Génération : cette étape porte sur la formalisation d'un modèle, par exemple avec OWL. C'est l'étape la plus importante du processus. Dans notre projet, nous construisons plusieurs ontologies pour les différents programmes scolaires regroupés par matière. Partant d'une ontologie de base, nous la peuplons par extraction d'information du fichier XML. L'ontologie de base est conçue manuellement pour tous les différents programmes scolaires. La conception de l'ontologie de base a été élaboré par une études détaillée des différents programmes scolaires de chaque matière de tout niveau.
- Validation : toutes les étapes précédentes peuvent introduire des concepts et des relations erronés, pour cela une phase de validation automatique et/ou

² <http://eduscol.education.fr/pid23391/programmes-de-l-ecole-et-du-college.html>

humaine est nécessaire. Nous avons procédé à une validation humaine, notamment avec les raisonneurs de l'outil Protégé ;

- Evolution : une ontologie ne doit pas être une représentation statique d'un domaine, mais doit évoluer avec lui ; cette phase n'a pas encore été abordée dans notre dispositif.

Une phase d'enrichissement avec des ontologies externes est en cours d'élaboration ; nous verrons ci-dessous qu'une première ébauche de cela est obtenue par l'utilisation de DBpedia Spotlight.

4.1. Phase de peuplement

Une classification sur la génération d'ontologie a été proposée dans [16]. Cette classification a regroupé les expériences et les outils en quatre catégories comme suit :

- Conversion ou traduction : il s'agit de logiciels qui assurent la transformation d'un format classique de représentation d'information (par exemple XML) vers une ontologie par des processus limités de traduction de format. C'est dans cet esprit que nous avons décomposé notre travail : la deuxième étape de notre traitement consiste à passer d'une représentation XML à une représentation RDF calquée sur la sémantique du XML initial et pas encore enrichie par des connaissances sur le domaine de connaissances visé.
- Basé sur les extractions : des techniques d'extraction ont été développées afin d'extraire d'informations pour générer une ontologie. La plupart des expériences sont axées sur des sources non structurées, comme des documents textuels ou des pages web et mettent en œuvre des techniques de traitement du langage naturel (TAL). Ces expériences nous disent que la récupération de concepts structurés de documents non structurés nécessite toujours une assistance humaine et que les techniques d'extraction de textes en langage naturel peuvent être utilisées en complément d'autres représentations existantes de connaissances structurées. Ce type d'outil inspire en partie la première étape de notre méthode : marquage sémantique local d'éléments dans un fichier source XML.
- Basé sur des connaissances externes : cette catégorie concerne les applications qui construisent ou enrichissent une ontologie de domaine par des ressources extérieures. Cette catégorie est classée avec les approches d'intégration des dictionnaires externes, des ontologies existantes ou des connaissances structurées (WordNet ou DBpedia). Ce type d'outil inspire une partie de la première étape de notre traitement (où nous cherchons des 'capacités' au sens des travaux de Bloom).
- Framework : regroupe les outils d'édition d'ontologie comme Protégé ; il s'agit là seulement d'une assistance à la création manuelle ou à la vérification d'ontologies. Protégé nous a servi à établir par nous-mêmes quelques ontologies de base pour notre travail.

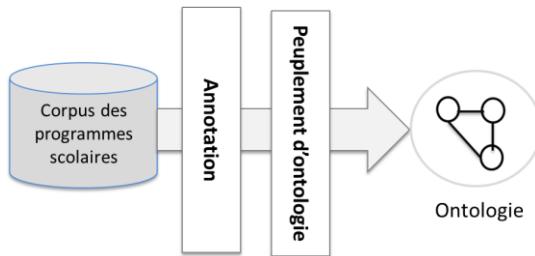


Figure 1. Architecture de la phase de peuplement d'ontologie.

Dans les programmes scolaires figurent des descriptions d'objectifs associés à chacune des sections des programmes. Comme c'est l'usage pour décrire ce type d'objectifs, les verbes d'action proposés par Bloom [17] est largement utilisées par la communauté des enseignants pour formuler les objectifs pédagogiques à atteindre. A la base, il s'agit d'une taxonomie composée de verbes à l'infinitif caractérisant des grandes catégories auxquels sont ensuite associés d'autres verbes caractérisant une hiérarchie de sous-catégories. Nous nous sommes inspirés de la taxonomie de Bloom proposée par l'European Schoolnet³.

Nous avons choisi une représentation dans le langage d'ontologie OWL. Plusieurs raisons ont guidé ce choix. D'abord une raison pragmatique de cohérence en terme d'outils techniques et conceptuels utilisés par notre équipe. Ensuite, parce que OWL nous a permis d'adoindre clairement la description de nombreux synonymes aux verbes de Bloom (nous verrons que cela nous permet d'améliorer le rappel dans nos traitements automatisés). Enfin, cette approche, bien intégrée aux technologies du web sémantique, nous paraît faciliter de futures évolutions de nos outils, par exemple pour la prise en compte de stratégies pédagogiques (par exemple, en associant un objectif pédagogique à des méthodes connues pour l'atteindre et qui seraient décrites dans cette ontologie ou dans d'autres).

Notre ontologie de Bloom est composée de 83 classes réparties dans une hiérarchie à partir de 6 classes principales (analyser, appliquer, créer, évaluer, mémoriser, comprendre). A ces classes ont été associés les labels des verbes correspondants, en français et en anglais. De plus, une référence aux mots anglais correspondants dans WordNet constitue une annotation qui peut aider à l'exploitation de ces verbes.

L'architecture de notre outil de peuplement d'ontologie est composée de deux parties principales (voir Figure 1) :

1. Le traitement des données (Annotation) : une étape de lemmatisation est appliquée pour la reconnaissance des mots de manière automatique sous différentes variations. Une autre étape de sélection de mots a été effectuée grâce à la mesure TF.IDF définie dans [18] [19]. Elle permet de proposer des 'topics' les plus porteurs de connaissances dans un domaine particulier, en mettant en évidence des mots singulièrement importants pour un programme. Enfin, nous utilisons notre ontologie 'de Bloom', qui sert à classifier les actions pédagogiques sous forme de verbes à la forme infinitive ; nous

³ <http://europeanschoolnet-vbe.lexaurus.net/lexaurus/browse>

l'avons créée d'après les travaux de Bloom. Ces étapes nous permettent d'annoter sémantiquement nos documents sources. (voir Figure 2)

2. Peuplement d'ontologie : la deuxième étape, nous permet de peupler l'ontologie de base automatiquement en s'appuyant sur les annotations effectuées sur le corpus. A chaque matière, nous associons une ontologie de base, créée manuellement à l'aide de l'éditeur d'ontologie Protégé après une analyse des concepts généraux utilisés par le programme (thèmes, capacités,...). Nos différentes ontologies de base sont très voisines et nous prévoyons de les unifier.

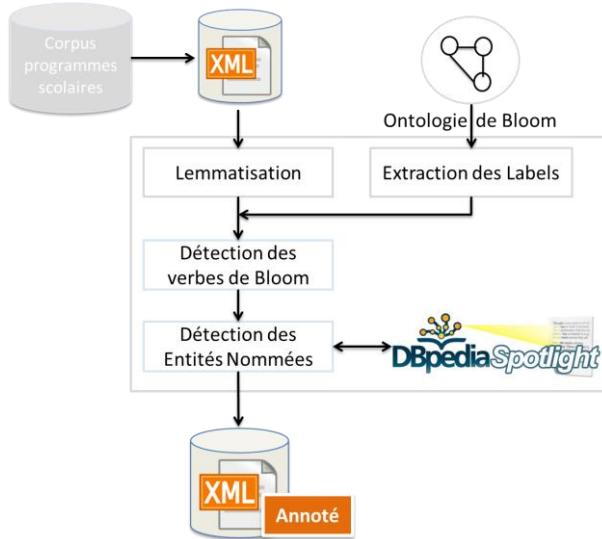


Figure 2. Etape d'annotation du corpus.

La lemmatisation est réalisée grâce à l'outil TreeTagger [20]. Cet étiqueteur grammatical permet de lemmatiser efficacement les phrases en français. Après lemmatisation, tous les mots sont représentés par leur forme générique. Cette étape nous permet de détecter les verbes de Bloom dans le corpus afin de localiser les capacités (un verbe+un objet). Les verbes de Bloom détectés dans le corpus sont marqués par des balises ajoutées au fichier XML d'entrée :

Tableau 1. La balise OntoClass.

<OntoClass uri="uri_vb_Bloom "> vb_Bloom </OntoClass>

La valeur de l'attribut URI de la balise OntoClass est la référence du concept détecté dans l'ontologie de Bloom.

Après la détection de tous les verbes de Bloom, une étape de reconnaissance de toutes les entités nommées est lancée. Cette étape utilise l'API DBPedia SpotLight dont les bons résultats sont démontrés [21]. Les entités nommées détectées sont marquées par des balises ajoutées au fichier XML d'entrée :

Tableau 2. La balise NamedEntity.

<NamedEntity type="type EN" uri="uri_EN"> EN </NamedEntity>

Nous récupérons grâce à DBpedia Spotlight le type (personne, lieu, homme politique, ...etc.) de l'entité nommée détectée et aussi son URI dans DBpedia.

La deuxième étape est l'étape de peuplement de l'ontologie de base. Une telle ontologie a été définie pour chaque matière du programme scolaire. La deuxième étape consistera à peupler l'ontologie de base avec des individus extraits du corpus XML annoté (voir Figure 3).

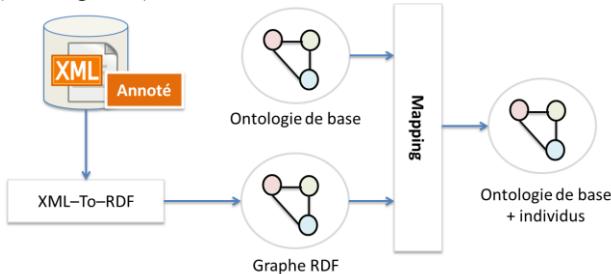


Figure 3. Etape de peuplement de l'ontologie de base.

Dans la phase de peuplement, une étape de transformation de notre corpus XML annoté en triplet RDF est accomplie par la méthode décrite dans les travaux de [22]. Cette transformation utilise un fichier XSLT. Les éléments XML sans élément fils sont représentés en sujet avec un prédicat « rdf:value ». Un exemple de transformation d'une portion d'un document XML en triplet RDF (turtle) de la balise section et de la balise titre est présentée dans le Tableau 3 ci-dessous,

Tableau 3. Transformation XML vers RDF.

...
<section>
<titre>Thème 4 - LE MONDE DEPUIS LE DEBUT DES ANNÉES 1990</titre>
<connaissances>CONNAISSANCES Les principales lignes de force de la géopolitique mondiale depuis le début des années 1990.
</connaissances>
...
</section>
...
...
<http://www.ilot.org/#programme/matiere_3/section_7/section_6>
mp:titre
<http://www.ilot.org/#programme/matiere_3/section_7/section_6/titre> . ;
<http://www.ilot.org/#programme/matiere_3/section_7/section_6/titre>
rdf:value
"Thème 4 - LE MONDE DEPUIS LE DEBUT DES ANNÉES 1990" .
...

La balise *section* a pour prédicat *mp:titre* car la balise *titre* est un fils de type texte (balise sans fils) de cette balise. La balise *titre* est de type texte dans le fichier XML et dans ce cas elle aura le prédicat *rdf:value* et un objet avec comme valeur le texte de la balise *titre*. Le document RDF généré est pauvre sémantiquement et peut être enrichi

par un mapping vers une ontologie OWL. L'étape de mapping est nécessaire pour peupler l'ontologie de base avec les données du graphe RDF ou pour ajouter aux triplets RDF générés des liens vers l'ontologie de base. L'ontologie de base nous permet d'avoir un modèle cohérent de la structure des programmes scolaires.

Sur la base des travaux de [23] définissant un mapping entre l'ontologie DBpedia et Wikipédia, nous proposons le mapping de nos documents aux formats RDF vers l'ontologie de base. Dans les triplets RDF, nous cherchons ceux avec le prédictat « mp :nom_propriété ». mp est le namespace associé à notre document XML annoté. Le « nom_propriété » désigne le nom de la propriété qu'on trouve dans le RDF. On aura, par exemple, mp:titre, mp:theme...

Tableau 4. Le mapping vers la classe Theme de la propriété titre.

```
 {{ TemplateMapping
mapToClass = ops:theme;
mappings = {{ PropertyMapping
templateProperty = mp:titre; ontologyProperty = ops :titre;
}}
}}
```

La ligne mapToClass spécifie l'URI de la classe de l'ontologie de base à instancier. Dans la partie propertyMapping on définit l'attribut templateProperty spécifie l'URI de la propriété qui nous intéresse et qui correspond à un prédictat du document RDF et l'attribut ontologyProperty spécifie le nom de la propriété de la classe qui recevra les données. Ces deux lignes, nous permettent de faire correspondre les valeurs de nos triplets RDF avec la propriété d'une classe.

Tableau 5. La requête SPARQL générée automatiquement.

```
SELECT ?vti
WHERE {
?th mp:titre ?ti .
?ti rdf:value ?vti .
}
```

A partir du mapping défini ci-dessus, une requête SPARQL est générée automatiquement pour extraire les données des documents RDF. La requête SPARQL du Tableau 5 est générée à partir du mapping du Tableau 4.

Deux types de propriétés sont distingués dans notre construction d'ontologie : les propriétés d'objet (ObjectProperty) permettent de relier des instances à d'autres instances ; les propriétés de type de donnée (DataTypeProperty) permettent de relier des instances à des types de données (entier, chaîne de caractère...etc.).

Dans les travaux de [24], le mapping des propriétés de type ObjectProperty n'est pas présenté pour le peuplement de l'ontologie DBpedia. Le mapping entre les informations des articles Wikipédia et l'ontologie DBpedia n'y est décrit que pour une ou plusieurs propriétés de type DataTypeProperty pour une classe donnée.

Nous proposons dans notre approche, une étape supplémentaire afin de remédier aux limites du mapping précédent. Pour obtenir les propriétés reliant les instances de classe dans notre ontologie de base, après chaque instantiation de chaque classe x de

notre ontologie par des individus, nous récupérons toutes les propriétés de type ObjectProperty de la classe x. Ensuite, avec une requête SPARQL, nous établissons le lien avec les autres instances des classes liées avec la classe x.

Tableau 5. Algorithme de la première étape de peuplement.

Algorithme de peuplement input: Ontology O1, RDF graph G, Mapping M; output: Ontology O1+individus Pour chaque classe C cible d'un mapping défini dans M Pour chaque mapping PM de propriété défini dans M pour cette classe Ajouter dans l'ontologie un individu I de classe C R est la liste de résultats de la requête SPARQL créée à partir de PM Pour chaque résultat r de la liste R Créer une propriété de I de type défini par le champ ontologyProperty de PM et de valeur r

La figure 4 montre un exemple de peuplement de l'ontologie de base « histoire et géographie » avec des données récupérées depuis le document du programme scolaires de 5eme. Le concept *capacités* est défini par les concepts *Bloom* et *Topic*. Le concept *topic* peut avoir un ou plusieurs *éléments*. Ces éléments peuvent être des entités nommées. La capacité récupérée du document « *Raconter une journée de Louis XIV* » est composée d'un verbe de Bloom et d'un Topic. Le topic « une journée de Louis XIV » est constitué d'une entité nommée « *Louis XIV* ». On voit ainsi que nous pourrons relier l'acquisition de cette capacité à des descripteurs liés à notre ontologie de Bloom et à des descripteurs externes liés à *Louis XIV*.

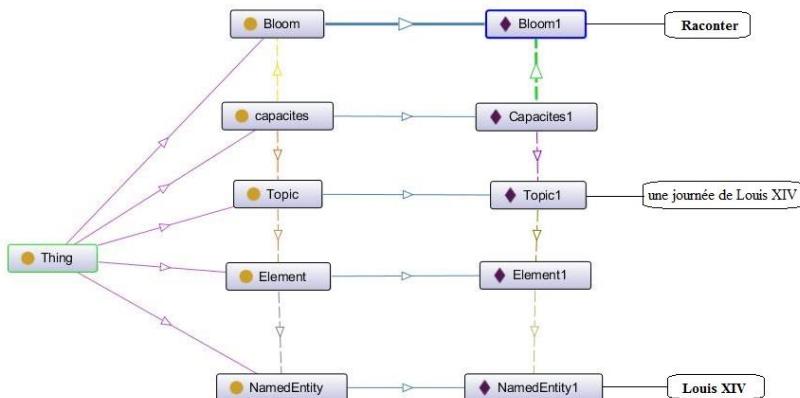


Figure 4. Exemple de peuplement d'ontologie avec une capacité et un topic.

4.2. Phase d'enrichissement

Pour enrichir les ontologies de base du programme scolaire, nous utilisons des collections de données ouvertes (open data). Nous pouvons distinguer deux types de telles collections : des données générales et des données spécifiques.

4.2.1. Enrichissement à l'aide des ‘Catégories’ de Wikipédia

La première collection concerne les différents types de données extraits de DBpedia, particulièrement Wikipédia Catégorie, destiné à classer les articles de Wikipédia. Ce type de donnée utilise les vocabulaires [24]. Depuis sa création, le graphe des ‘Wikipédia catégories’, est exploité comme objet de recherche à part entière. Nous pouvons citer les travaux récents comme la constitution d’une ressource sémantique issue du treillis des catégories de Wikipédia [25] et l’usage de catégorie pour la conception d’un système de recommandation inter-domaines [26].

Dans notre étude de cas, nous utilisons la classification hiérarchique de ‘Wikipédia catégorie’ pour enrichir les ontologies de programme scolaire de base. Par exemple, dans l’ontologie de l’Histoire des arts, une des classes primaires est la classe « Domaine artistique » contenant les informations concernant les six domaines artistiques : Arts de l'espace, Arts du langage, Arts du quotidien, Arts du son et Arts du spectacle vivant. Le programme officiel propose des listes d'exemples concrets pour chaque domaine. Les exemples des Arts de l'espace sont architecture, urbanisme, arts des jardins, paysages aménagé. Dans l’ontologie Histoire des Arts de base, ces derniers deviennent des concepts faisant partie des sous-classes de la classe «Arts de l'espace».

Un des scénarios d’enrichissement de l’ontologie grâce à ‘Wikipédia catégorie’ est la proposition semi-automatique de nouvelles entités associées. Concrètement, le système propose l’arborescence de « Catégorie : architecture » à l’utilisateur qui souhaite personnaliser son ontologie ‘Histoire des Arts’. La figure 1 représente l’arborescence de « Catégorie : architecture », catégorie mère, et les chemins à valider par l’enseignant jusqu’à la sous-catégorie souhaitée. Le chemin contient ainsi : Architecture > Style_architectural > Architecture_gothique. A l’étape finale, l’enseignant peut valider des entités souhaitées et ajouter la « Catégorie : Architecture_gothique » en tant que nouvelles sous-classes du domaine Architecture dans son ontologie dérivée de l’ontologie de base d’Histoire des Arts.

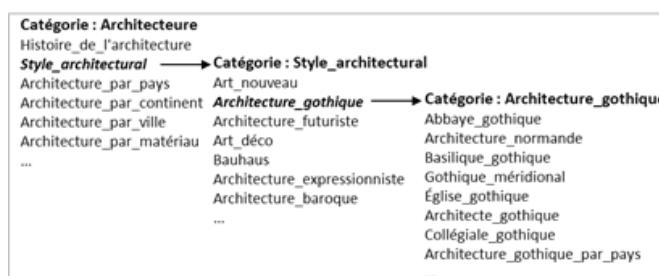


Figure 5. Arborescence de « Catégorie : Architecture » dans DBpedia.

L'usage de 'Wikipédia catégorie' illustré dans le scénario précédent pourra effectivement s'appliquer à l'enrichissement de l'ensemble des ontologies du programme scolaire de façon plus générique.

Partant d'une étude manuelle humaine des possibilités, nous avons entrepris une systématisation de la démarche avec des possibilités de propositions et de validation des enrichissements.

4.2.2. Enrichissement avec des données spécifiques

Le deuxième type de données que nous pouvons exploiter pour la phase d'enrichissement concerne les données ouvertes mises à disposition par les institutions publiques. Nous avons identifiés trois jeux de données ayant à voir avec nos ontologies de programmes scolaires sur la plateforme *data.gouv.fr*. Nous analysons les possibilités de connexion entre ces données et les concepts de nos ontologies. Il s'agit de « Histoire des arts-Notices textes » du Ministère de la Culture et de la Communication, d'une liste de 156 dossiers pédagogiques produits par le Centre national d'art et de culture Georges Pompidou et enfin de la collection des données sur la plateforme *data.bnf.fr*. Ces données pourront être exploitées pour l'enrichissement des ontologies du programme scolaire telles qu'Histoire des Arts, Histoire et Français.

A titre d'exemple, « Histoire des arts-Notices textes » est un document CSV peu structuré qui contient environ 4777 références utiles pour le programme d'Histoire des Arts. Nous l'avons transformé en XML, puis par un traitement XSLT, nous avons structuré le document. A partir du schéma XML du document, nous avons tiré une ontologie. En calquant la méthode sur ce qui proposé dans les sections précédentes, nous avons produit un ensemble de triplets RDF représentant les références disponibles. Puis nous avons pu établir des liens entre les thèmes indiqués dans ce document et les thèmes présents dans notre ontologie du programme scolaire d'Histoire des Arts.

A l'aide des outils Datalift [27], nous avons publié notre fichier XML, puis exporté une représentation RDF. Nous pouvons ainsi faire des interrogations SPARQL de ces connaissances sur l'Histoire des Arts ; voici un exemple de requête SPARQL :

```
SELECT DISTINCT ?titre WHERE {  
?tags ha:tag "Normandie" .  
?row ha:tags ?tags .  
?peinture ha:tag "Peinture" .  
?row ha:tags ?peinture .  
?row ha:titre ?titre  
} LIMIT 100
```

qui nous donne le titre de toutes les références qui ont pour tags les mots "Peinture" et "Normandie".

Le fait de mettre en correspondance les ontologies du programme scolaire avec d'autres données structurées, traitées et préparés par les professionnels du domaine a un double avantage. Il facilite, en premier lieu, la tâche de l'enseignant pour trouver des ressources éducatives de bonne qualité et fiables correspondants aux programmes

officiels. Il est, en outre, bénéfique pour les producteurs des données culturelles, la réutilisation de leurs ressources étant considérée comme un moyen de la valorisation du patrimoine par l'éducation. Nous sommes également en train d'identifier et d'analyser d'autres données culturelles produites par les collectivités locales. L'objectif du repérage de telles données est d'exploiter ces données dans le contexte de l'adaptation du programme scolaire pour la valorisation des patrimoines régionaux. Cette action fait partie des démarches pédagogiques fortement recommandées par le programme de l'Histoire des Arts.

4.3. Perspective: phase d'exploitation

C'est la dernière phase de notre méthode qui servira d'interface entre l'enseignant et les programmes scolaires. Notre formalisation est actuellement constituée de plusieurs ontologies de chaque matière. Nous comptons rapidement créer une ontologie intégrant l'ensemble, ce qui facilitera l'établissement de liens entre programmes. Nous avons déjà créé des interfaces qui exploitent nos ontologies pour présenter les programmes en mettant en évidence les capacités à acquérir et les liens avec des ressources tirées de DBpedia. La perspective est de rendre d'autres ressources externes exploitables pour un enrichissement tant du travail du professeur en phase de préparation de cours que du parcours des cours par les élèves.

5 Conclusion

Dans cet article, nous avons proposé une méthode de création d'ontologie à partir d'un corpus de programmes scolaire français. Pour cela, nous avons définis une ontologie de base pour chaque matière afin d'éviter des problèmes d'incohérence sur la génération automatique de l'ontologie.

Dans cette approche, nous avons utilisé l'ontologie de Bloom pour la détection des capacités dans notre corpus et l'API DBpedia SpotLight [21] pour la détection des entités nommées. Les documents XML ont été ensuite transformés en RDF qui seront mappé vers nos ontologies de base. Ces ontologies vont être enrichie et exploité semi-automatique afin d'enrichir les possibilités de l'enseignant en s'appuyant sur le contenu des programmes scolaires.

Dans nos travaux futurs, nous enrichissons l'ontologie de Bloom avec des synonymes de verbes (167 synonymes) pour chaque classe de verbe afin de permettre la détection d'autres capacités dans notre corpus ce qui augmentera la précision dans notre approche.

Références

1. Gruber, T., Toward principles for the design of ontologies used for knowledge sharing. International Journal of Human-Computer Studies, 43, pp. 907-928. 1993.

2. Hérin, D., Sala, M., & Pompidor, P., Evaluating and Revising Courses from Learning Web Resources. ITS'2002. 2002.
3. Pompidor, P., Sala, M., & Hérin, D., An incremental method for extraction of pedagogical knowledges on the web. SW-WL'2003 & EIAH'2003. 2003.
4. Sala, M., Pompidor, P., & Hérin, D., Aid to the Semantic Maintenance of the Web Site. IADIS WWW/Internet'03. 2003.
5. Nkambou, R., Frasson, C., & Gauthier, G., CREAM-Tools: An Authoring Environment for Knowledge Engineering in Intelligent Tutoring Systems. In Authoring Tools for Advanced Technology Learning Environments: Toward coste effective, adaptative, interactive, and intelligent educational software, pp. 93-138. 2003.
6. Hernandez, N., & Mothe, J., TtoO: une méthodologie de construction d'ontologie de domaine à partir d'un thésaurus et d'un corpus de référence. Rapport de recherche, IRIT/RR 2006-04--FR, IRIT. 2006.
7. Cimiano, P., & Völker, J., Text2Onto - A Framework for Ontology Learning and Data-driven Change Discovery. (L. N. Springer, Éd.) Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB), pp. 227-238. 2005.
8. Maedche, A., & Staab, S., Ontology Learning for the Semantic Web. IEEE Intelligent Systems, Special. 2001.
9. Roitman, H., & Gal, A., OntoBuilder: fully automatic extraction and consolidation of ontologies from web sources using sequence semantics. EDBT'06 Proceedings of the 2006 international conference on Current Trends in Database Technology, pp. 573-576. 2006.
10. Ferdinand, M., Zirpins, C., & Trastour, D., Lifting XML Schema to OWL. Web Engineering Lecture Notes in Computer Science Volume 3140, pp. 354-358. 2004.
11. Bohring, H., & Auer, S., Mapping XML to OWL Ontologies. In Leipziger Informatik-Tage, vol. 72, pp. 147-156. 2005.
12. Ghawi, R., & Cullot, N., Building Ontologies from XML Data Sources. DEXA '09. 20th International Workshop on Database and Expert Systems Application, pp. 480-484. 2009.
13. Gras-Velazquez, A., Teachers and content packaging standards. Initial conclusions from the ASPECT evaluation. Récupéré sur Adopting Standards and Specifications for Educational Content: <http://aspect-project.org/node/84>. 2010.
14. Gómez de Regil, R., Retour d'expérience sur le pilote ASPECT. Récupéré sur <http://www.lom-fr.fr/scolomfr/communication/conferences.html>. 2011.
15. Desmoulins, C. Construction avec des enseignants d'une ontologie des compétences en géométrie, Geoskills. Ingénierie des connaissances (IC 2010). <http://www-limbio.smbh.univ-paris13.fr/GBPOnto/data/documents/2010/5desmoulins.pdf>
16. Bedini, I., & Nguyen, B., Automatic Ontology Generation: State of the Art. University of Versailles Technical report. 2007.
17. Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. Taxonomy of educational objectives: Handbook I: Cognitive domain. New York: David McKay, 1956.
18. LUHN, H. P. (1958). The automatic creation of literature abstracts. IBM Journal on Research and Development, 2(2).
19. SPÄRCK JONES, K. (1972). A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation, 28
20. Schmid, H., Probabilistic part-of-speech tagging using decision trees. In International Conference on New Methods in Language Processing. 1994.
21. Mendes P.N., Jakob M., García-Silva A., Bizer C. D., DBpedia Spotlight: Shedding Light on the Web of Documents. In the Proceedings of the 7th International Conference on Semantic Systems (I-Semantics 2011). Graz, Austria. 2011.
22. Breitling, F., A standard transformation from XML to RDF via XSLT. Astronomical Notes, pp. 755-760. 2009.

23. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P., & Bizer, C., Dbpedia-a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*. 2013.
24. Torres, D., Molli, P., Skaf-Molli, H., & Diaz, A., Improving wikipedia with DBpedia. In *Proceedings of the 21st international conference companion on World Wide Web*, pp. 1107-1112. 2012.
25. Collin, O., Gaillard, B., & Bouraoui, J.-L., Constitution d'une ressource sémantique issue du treillis des catégories de Wikipedia. *TALN 2010-Session Posters*. 2010.
26. Fernández-Tobías, I., Kaminskas, M., Cantador, I., & Ricci, F., A generic semantic-based framework for cross-domain recommendation. *HetRec '11 Proceedings of the 2nd International Workshop on Information Heterogeneity and Fusion in Recommender Systems*, pp. 25-32. 2011.
27. Scharffé, F., Atemezing, G., Troncy, R., Gandon, F., Villata, S., Bucher, B., & Vatant, B., Enabling linked-data publication with the datalift platform. In *Proc. AAAI workshop on semantic cities*. 2012.