Vers la publication d'une RTO dédiée à l'annotation de relations n-aires

Patrice Buche[1,2], Stéphane Dervaux[3], Juliette Dibie-Barthélemy[4], Liliana Ibănescu[4], Lydie Soler[3] and Rim Touhami[1,4]

[1] INRA - UMR IATE, 2, place Pierre Viala, F-34060 Montpellier Cedex 2, France
[2] LIRMM, Montpellier, France
[3] INRA - MÉT@RISK, 16 rue Claude Bernard, F-75231 Paris Cedex 5, France
[4] INRA - MÉT@RISK & AGROPARISTECH, 16 rue Claude Bernard, F-75231 Paris Cedex 5, France
Patrice.Buche@supagro.inra.fr,
{Stephane.Dervaux,Lydie.Soler}@paris.inra.fr,
{Juliette.Dibie,Liliana.Ibanescu,Rim.Touhami}@agroparistech.fr

Résumé : Nous présentons dans cet article la modélisation d'une RTO dédiée à la tâche d'annotation de relations n-aires, qui permettent de représenter des résultats expérimentaux, et un corpus de tableaux annotés avec cette RTO. Nous souhaiterions pouvoir publier cette RTO et le corpus de tableaux associés sur le Web de données, mais de nombreuses questions se posent. Nous espérons que notre participation à l'atelier Ontologies et Jeux de Données pour évaluer le web sémantique (OJD) nous aidera à y répondre.

Mots-clés : Ressource termino-ontologique, Annotation de relations n-aires

1 Introduction

La notion d'ontologie est une notion clé du Web sémantique et des recherches en ingénierie des connaissances. Les travaux récents de (Cimiano *et al.*, 2011; McCrae *et al.*, 2011; Roche *et al.*, 2009) ont proposé d'associer une partie terminologique et/ou linguistique aux ontologies afin d'établir une distinction claire entre la composante terminologique et la composante conceptuelle. (Roche *et al.*, 2009) ont en particulier défini la notion de ressource termino-ontologique (RTO).

Dans cet article, nous présentons notre modélisation d'une Ressource Termino-Ontologique (RTO) dédiée à la tâche d'annotation de relations n-aires (cf. (Touhami *et al.*, 2011)). Nous nous intéressons plus précisément aux relations n-aires permettant de représenter des résultats expérimentaux. Cette RTO a été utilisée pour annoter des relations n-aires dans des tableaux de données (Buche *et al.*, 2012) et dans des textes (Ghersedine *et al.*, 2012) issus de documents provenant de sources externes telles que des publications scientifiques ou des rapports internationaux. Notre travail sur l'annotation de tableaux guidée par une RTO a été expérimenté sur plusieurs corpus et en particulier sur un corpus dans le domaine de la microbiologie prévisionnelle, qui est actuellement composé d'une soixantaine de tableaux annotés. Notre RTO comme le corpus de tableaux annotés ont été définis dans les langages du Web sémantique : OWL pour la RTO et XML/RDF pour les tableaux et leurs annotations. Afin de les rendre accessibles à un plus grand nombre, afin de pouvoir bénéficier de retour d'expérience sur nos données, afin de pouvoir lier notre RTO à des ontologies existantes, nous souhaiterions publier, dans un proche avenir, sur le Web de données notre RTO et le corpus de tableaux annotés associé. Face à ce souhait de nombreuses questions se posent comme par exemple quel format de données choisir, comment établir des liens avec des ontologies existantes, comment ne pas se perdre dans le nuage de données, où publier... Nous espérons que l'atelier Ontologies et Jeux de Données pour évaluer le web sémantique (OJD), nous permettra d'obtenir des réponses et nous aider ainsi à atteindre notre objectif.

Dans cet article, nous commencons par quelques rappels sur la modélisation de notre RTO dédiée à la tâche d'annotation de relations n-aires. Nous présentons dans la deuxième section notre corpus composés de tableaux au format XML avec leurs annotations associées en RDF à travers un exemple. Enfin, nous concluons cet article et présentons nos attentes pour la publication de notre RTO et corpus sur le Web de données.

2 La Ressource Termino-Ontologique (RTO)

Nous présentons dans cette première section la modélisation d'une Ressource Termino-Ontologique (RTO) dédiée à la tâche d'annotation de relations n-aires, qui permettent de représenter des résultats expérimentaux (Touhami *et al.*, 2011). Une RTO est une ressource comportant une composante conceptuelle (l'ontologie) et une composante terminologique, dans

laquelle la manifestation linguistique (le terme) se distingue de la notion qu'elle dénote (le concept). La composante conceptuelle de la RTO est composée de deux parties principales : une partie générique, appelée ontologie noyau (*core ontology*), qui permet de représenter la structure de l'ontologie dédiée à la tâche d'annotation de relations n-aires, et une partie spécifique, appelée ontologie de domaine, qui dépend du domaine étudié.

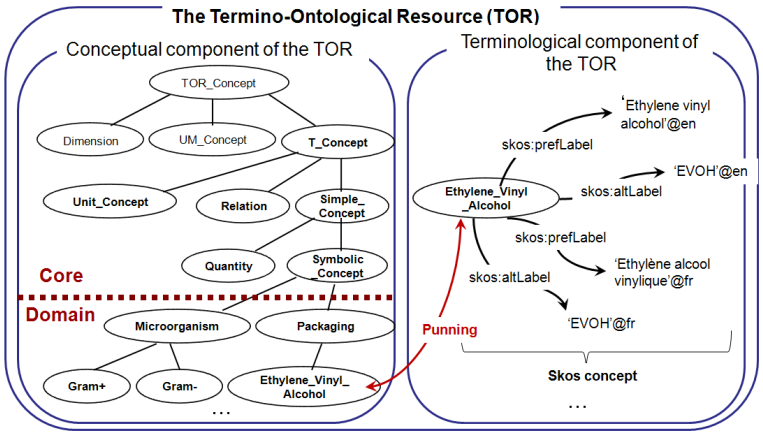

FIGURE 1 – La RTO dans le domaine de la microbiologie.

La figure 1 présente la partie générique de la RTO et sa spécialisation dans le domaine de la microbiologie prévisionnelle. Ils apparaissent dans la RTO comme sous concepts des concepts génériques de l'ontologie noyau. Considérons en particulier le concept générique *T_Concept*, pour *Terminological Concept*. Il est composé de trois sous-concepts : les concepts *Unit_Concept*, *Relation* et *Simple_Concept*, auxquels sont associés un ou plusieurs termes dans la composante terminologique. Le concept *Symbolic_Concept* est caractérisé par sa hiérarchie de spécialisation (par exemple *Gram+* et *Gram-* sont des spécialisations de *Microorganism*). Le concept *Quantity* est caractérisé par un ensemble d'unités de mesure, sous-concepts du concept générique *Unit_Concept*, et éventuellement un domaine de valeurs. Le concept *Relation* permet de représenter des relations n-aires entre des concepts simples, correspondant à la représentation de résultats expérimentaux. La signature d'une relation est définie par un domaine et un co-domaine. Le co-domaine est limité à un seul concept simple, appelé *concept résultat*, alors que le domaine peut être composé d'un ou

plusieurs concepts simples, appelés *concepts d'accès*. La RTO dans le domaine de la microbiologie prévisionnelle est composée de 689 concepts *Symbolic_Concept*, 16 concepts *Quantity* et de 18 concepts *Relation*.

Exemple 1
Le concept Relative_humidity *peut être exprimé dans l'unité '%' et a pour domaine de valeurs [0, 100].*
Le concept relation spécifique O2Permeability_Relation *(cf. figure 2) permet de représenter une relation n-aire. Ses concepts d'accès sont les concepts relatifs à l'emballage, à son épaisseur, à la température et à l'humidité relative de l'environnement. Le concept résultat représente la perméabilité de l'emballage à l'oxygène dans ces conditions expérimentales.*

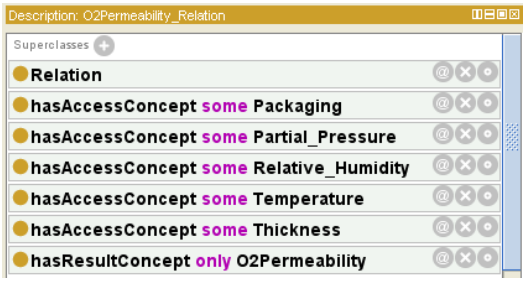

FIGURE 2 – Un exemple de relation dans le domaine de la microbiologie.

La composante terminologique de la RTO contient l'ensemble des termes du domaine étudié. Au moins un terme de cette composante est associé à chaque sous concept du concept générique *T_Concept*. Par exemple, dans la figure 1, les termes *Ethylene vinyl alcohol* et *EVOH* sont associés au concept *Ethylene_Vinyl_Alcohol*. Chaque sous concept du concept générique *T_Concept* est, dans une langue donnée, caractérisé par un label préféré et éventuellement par un ensemble de labels alternatifs, qui correspondent à des synonymes ou des abréviations. La RTO dans le domaine de la microbiologie prévisionnelle est composée de 1 492 termes.

La composante conceptuelle de la RTO est représentée en OWL2-DL et les labels sont associés aux concepts grâce aux propriétés SKOS de labellisation, recommandées par le W3C.

3 Le corpus de tableaux annotés

Nous avons proposé dans (Buche *et al.*, 2012) une méthode d'annotations de tableaux guidée par une RTO. L'annotation sémantique d'un tableau consiste à annoter le contenu de ses cellules, pour pouvoir ensuite identifier les concepts symboliques ou numériques représentées par ses colonnes, puis identifier la ou les relations sémantiques n-aires qui existent entre ses colonnes. La dernière étape de l'annotation consiste enfin à instancier les relations identifiées pour chaque ligne du tableau. Pour une ligne donnée, chaque valeur de cellule est représentée par un sous ensemble flou en fonction de la nature de la cellule. Deux cas de figure se présentent : (i) les cellules numériques sont instanciées par des sous-ensembles flous à support numérique, présentés sous la forme d'intervalles trapézoïdaux et qui représentent une distribution imprécise de valeurs et (ii) les cellules symboliques sont instanciées par des sous-ensembles flous à support symbolique contenant chacun une liste de termes de l'ontologie les plus similaires à la valeur d'origine avec leur score de similarité. Nous présentons cette méthode à travers l'exemple ci-dessous.

Exemple 2
La figure 3 présente une partie du graphe RDF flou correspondant à l'instanciation de la deuxième ligne du tableau. La deuxième ligne représentée par l'URI uriRow2 *est annotée par une instance du concept* O2Permeability_Relation *avec un degré de pertinence de 0.75, qui correspond à la confiance accordée à la reconnaissance de cette relation par la méthode d'annotation. La signature de l'instance de relation* O2Permeability_Relation *est composée d'un concept résultat, une instance du concept* O2Permeability *annotée par le sous-ensemble flou continu de support* sup *= [3,5] et de noyau* ker *= [3,5], et de 2 concepts d'accès, une instance du concept* Packaging, *annotée par le sous-ensemble flou discret {1.0/Ethylene-Vinyl Alcohol}, et une instance du concept* Thickness, *annotée par le sous-ensemble flou continu de support* sup *= [25,25] et de noyau* ker *= [25,25] (non présenté dans la figure 3).*

Notre corpus dans le domaine de la microbiologie prévisionnelle contient 60 tableaux annotés composés de 342 colonnes et 700 lignes. Les annotations associées aux tableaux sont représentés par 312 graphes RDF, soit 22 000 triplets.

Sample	Grammage (g/m2)	Thickness (µm)	Air permeability (nm/Pa s)	Oxygen permeability in the material (ml m-2 day-1)
MFC film A	17 ±1	21 ±1	13 ± 2	17.0, 18.5
EVOH	_	25	_	3–5
Cellophane	_	21	_	3

Packaging | Thickness | O2Permeability | O2Permeability_ Relation

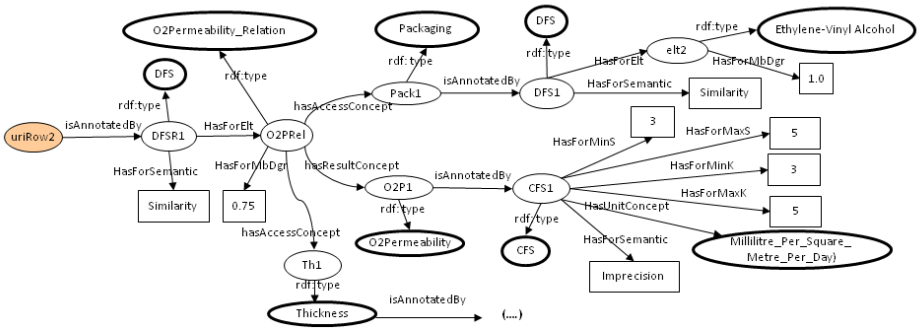

FIGURE 3 – Un exemple de tableau avec l'annotation de sa deuxième ligne.

4 Conclusion et perspectives

Nous avons présentés dans cet article la modélisation d'une RTO dédiée à la tâche d'annotation de relations n-aires, qui permettent de représenter des résultats expérimentaux, et un corpus de tableaux annotés avec cette RTO. Nous souhaiterions pouvoir publier cette RTO et le corpus de tableaux associés sur le Web de données, mais cet objectif soulève de nombreuses questions :

- Comment permettre l'accès à nos données (RTO et corpus) de telle sorte qu'elles soient visibles, facilement accessibles et utilisables par le plus grand nombre ? Comment les structurer, les formater, les annoter pour ce faire ?
- Où publier : faut-il se spécialiser et publier sur un portail tel que BioPortal [1] ou rester plus générique et publier sur un portail tel que DBpedia Ontology [2] ?
- Comment établir des liens avec d'autres ontologies et corpus existants ?

Nous espérons que notre participation à l'atelier Ontologies et Jeux de Données pour évaluer le web sémantique (OJD) nous aidera à trouver des réponses à nos questions et peut-être aussi à nous poser d'autres questions.

Références

BUCHE P., DIBIE-BARTHÉLEMY J., IBĂNESCU L. & SOLER L. (2012). Fuzzy Web Data Tables Integration Guided by an Ontological and Terminological Resource. *IEEE TKDE*, **99**(PrePrints).

CIMIANO P., BUITELAAR P., MCCRAE J. & SINTEK M. (2011). LexInfo : A declarative model for the lexicon-ontology interface. *J. Web Sem.*, **9**(1), 29–51.

GHERSEDINE A., BUCHE P., DIBIE-BARTHÉLEMY J., HERNANDEZ N. & KAMEL M. (2012). Extraction de relations n-aires interphrastiques guidée par une rto. In *actes de la 9ième édition de la COnférence en Recherche d'Information et Applications, CORIA*, p. 179–190.

MCCRAE J., SPOHR D. & CIMIANO P. (2011). Linking Lexical Resources and Ontologies on the Semantic Web with Lemon. In *ESWC (1)*, volume 6643 of *LNCS*, p. 245–259 : Springer.

1. http ://bioportal.bioontology.org/
2. http ://wiki.dbpedia.org/Ontology

ROCHE C., CALBERG-CHALLOT M., DAMAS L. & ROUARD P. (2009). Ontoterminology - a new paradigm for terminology. In *KEOD*, p. 321–326.

TOUHAMI R., BUCHE P., DIBIE-BARTHÉLEMY J. & IBĂNESCU L. (2011). An Ontological and Terminological Resource for n-ary Relation Annotation in Web Data Tables. In *OTM 2011 (2)*, volume 7045 of *LNCS*, p. 662–679 : Springer.