Intégration de données hétérogènes et imprécises guidée par une ressource termino-ontologique

Patrice Buche, Stéphane Dervaux, Juliette Dibie-Barthelemy, Liliana Ibanescu, Lydie L. Soler, Rim Touhami

Intégration de données hétérogènes et imprécises guidée par une Ressource Termino-Ontologique

Application au domaine des sciences du vivant

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

1. INRA - UMR IATE, 2, place Pierre Viala, F-34060 Montpellier Cedex 2, France
Patrice.Buche@supagro.inra.fr

2. LIRMM - équipe GraphIK, 161 rue Ada, F-34095 Montpellier Cedex 5, France

3. INRA - Mét@risk, 16 rue Claude Bernard, F-75231 Paris Cedex 5, France
{Stephane.Dervaux,Lydie.Soler}@paris.inra.fr,Rim.Touhami@agroparistech.fr

4. AgroParisTech, 16 rue Claude Bernard, F-75231 Paris Cedex 5, France
{Juliette.Dibie,Liliana.Ibanescu}@agroparistech.fr

RÉSUMÉ. *Cet article présente les enjeux de l'ingénierie des connaissances dans le domaine des sciences du vivant et, à titre d'illustration, un système d'intégration de données thématiques ouvert sur le Web, appelé ONDINE (Ontology based Data INtEgration). Ce système propose un processus complet d'acquisition, d'annotation sémantique et d'interrogation de données à partir de tableaux trouvés dans des documents scientifiques issus du Web. L'élément central du système ONDINE est une Ressource Termino-Ontologique (RTO) qui permet la représentation de relations n-aires et dont les concepts sont utilisés pour annoter des tableaux de données. Nous présentons le modèle de la RTO, la méthode d'annotation semi-automatique de tableaux de données guidée par cette RTO, puis le logiciel @Web (Annotating Tables from the Web) d'annotation sémantique de tableaux.*

ABSTRACT. *In this paper we present some issues of knowledge engineering in the field of life sciences and, as an illustration, a data integration system opened on the Web, called ONDINE (Ontology based Data INtEgration), which proposes a complete workflow to extract, to semantically annotate and to query data from tables found in scientific documents from the Web. The*

core and key element of ONDINE is an Ontological and Terminological Resource (OTR) allowing the modeling of n-ary relations; concepts from this OTR are used to annotate tables. First we present the OTR model, then the semi-automatic method for semantic annotation of tables guided by this OTR, and finally our software system, @Web (Annotating Tables from the Web), designed to semantically annotate tables.

MOTS-CLÉS : *intégration de données, annotation sémantique, ontologie, linked data.*

KEYWORDS: *data integration, semantic annotation, ontology, Web de données.*

DOI:10.3166/RIA.x.1-30

1. Introduction

Depuis quelques années, tous les domaines scientifiques sont confrontés à un accroissement exponentiel des données provenant de l'adoption généralisée de nouvelles technologies et du développement des sciences et techniques de l'information d'une ampleur et à une échelle sans précédents [1]. Actuellement, la question est donc plus de savoir comment organiser et analyser ces données à l'aide de méthodes et outils informatiques et mathématiques que de les produire. Un des facteurs limitant du traitement informatique des données en sciences du vivant et de l'environnement est la difficulté à appréhender les données disponibles et à les exploiter conjointement. Cette difficulté s'explique par : (1) la grande dispersion des données scientifiques : les données sont éparses, stockées dans des bases de données de laboratoire, publiées dans des revues scientifiques ou sur des pages de sites Web, mais aussi dans des rapports de projet, des thèses ou des supports de cours ; (2) leur hétérogénéité aussi bien dans leurs formats (textes, tableaux, graphiques, images, . . .) que dans leurs vocabulaires ; (3) leur caractère multi-échelles dans le temps et l'espace; (4) leur imprécision notamment dûe à la prise en compte de la variabilité expérimentale; (5) leur fiabilité, qui concerne tout particulièrement les données du Web. Les questions de recherche liées au traitement informatique de ces données et connaissances intéressent depuis longtemps de nombreuses disciplines informatiques (e.g. bases de données, intégration de données, représentation des connaissances, acquisition de données), et en particulier l'Ingénierie des Connaissances (IC). D'après (Aussenac-Gilles *et al.*, 2012), "*l'Ingénierie des Connaissances (IC) propose des concepts, méthodes et techniques permettant de modéliser, de formaliser et d'acquérir des connaissances dans les organisations dans un but d'opérationnalisation, de structuration ou de gestion au sens large. L'IC trouve ainsi ses champs d'application dans les domaines où l'objectif est de modéliser les connaissances et les mettre à disposition comme support à une activité ou à un raisonnement. Les applications concernées sont celles liées à la gestion des connaissances, à la recherche d'information (sémantique), à l'aide à la navigation, à l'aide à la décision. Enfin, l'IC entretient des relations étroites avec le Web Sémantique, qui a un statut particulier en raison de forts recouvrements avec l'IC via*

1. "Data, data everywhere". The Economist. 25 February 2010.

le partage de nombreux outils et méthodes (ontologies, langages de représentation des connaissances, raisonnements, etc.).". L'IC étudie en particulier comment adapter et/ou développer des méthodes et des outils innovants de capitalisation et de modélisation de connaissances et de données afin de permettre leur exploitation (analyse, traitement, visualisation) et leur partage (pluridisciplinaire, national et international) par des systèmes informatiques. L'IC est ainsi devenue un enjeu stratégique majeur dans ce monde de données numériques, confronté à la gestion de données et de connaissances en forte croissance. Cet enjeu est particulièrement fort dans les domaines des sciences du vivant et de l'environnement où les systèmes informatiques développés devront, à terme, permettre d'aider les décideurs à répondre à des questions de portée mondiale comme la sécurité alimentaire, la santé humaine, l'impact du changement climatique ou encore la préservation de la biodiversité.

Pour pouvoir construire un système de capitalisation et de modélisation de connaissances et de données provenant de sources hétérogènes, une des multiples solutions possibles est d'utiliser des ontologies (N. F. Noy, 2004 ; Doan *et al.*, 2012) pour spécifier des vocabulaires conceptuels standardisés, l'indexation des sources de données avec le vocabulaire de l'ontologie permettant l'interopérabilité de ces sources. Une ontologie définit un ensemble de primitives de représentation pour modéliser un domaine; les primitives sont des classes (ou des ensembles), des attributs (ou des propriétés) et des relations entre les membres des classes (Guarino *et al.*, 2009). La construction d'une ontologie pour formaliser un domaine représente un des enjeux méthodologiques et applicatifs actuels dans le domaine de l'IC (Aussenac-Gilles *et al.*, 2012). Les ontologies sont, en effet, nées d'un besoin de standardisation des vocabulaires ressenti dans de nombreux domaines et connaissent, depuis quelques années, un incroyable succès comme en atteste le nombre croissant de sessions portant sur les ontologies dans les conférences nationales ou internationnales en Intelligence Artificielle, et donc de travaux de recherche qui travaillent sur ou utilisent des ontologies. D'une part, de nombreux travaux portent sur l'étude des ontologies à tous les stades de leur cycle de vie (Staab, Studer, 2009) allant de leur construction (Cimiano *et al.*, 2010) à leur évolution en passant par leur alignement (Shvaiko, Euzenat, 2013 ; Bernstein *et al.*, 2011). D'autre part, les ontologies sont de plus en plus utilisées pour répondre à différents besoins. Elles sont notamment utilisées en intégration de données pour garantir l'interropérabilité de sources de données comme nous l'illustrerons dans cet article ou pour guider l'interrogation de données hétérogènes (Corby *et al.*, 2006 ; Pan *et al.*, 2008 ; Amarger *et al.*, 2013). Elles peuvent en particulier permettre de caractériser la fiabilité de sources de données (Destercke *et al.*, 2013) ou tracer la provenance des données [2]. Enfin, l'explosion du Web de données [3] dont la promesse est le partage à grande échelle des données liées (Heath, Bizer, 2011), ne fait que renforcer cette montée en puissance des ontologies dans le domaine de l'Intelligence Artificielle.

2. http://www.w3.org/TR/prov-o/
3. http://www.w3.org/standards/semanticweb/data

Dans les domaines des sciences du vivant et de l'environnement, plusieurs initiatives ont été prises pour faire face au besoin de standardisation des vocabulaires. La première initiative a consisté à définir des thésaurus pour fixer le vocabulaire, ce qui est un pré-requis indispensable au croisement des sources de données. Plusieurs thésaurus ont été créés au niveau international, les deux plus importants étant actuellement AGROVOC et NALT. AGROVOC [4] a été créé dans les années 1980 par la FAO (Food and Agriculture Organization of the United Nations) comme un thésaurus structuré multilingues pour les domaines de l'agriculture, de la sylviculture, de la pêche, de l'alimentation et de domaines apparentés (comme l'environnement). Il est actuellement disponible en 19 langues, avec une moyenne d'environ 40 000 termes dans chaque langue (Caracciolo *et al.*, 2012). AGROVOC est mis à jour régulièrement et est utilisé pour indexer le contenu de nombreuses bibliothèques spécialisées. NALT [5] est un thésaurus bilingue, avec actuellement environ 91 000 termes en anglais et espagnol, comparable à AGROVOC en terme de domaine couvert et maintenu par la USDA (United States Department of Agriculture). 13 390 termes d'AGROVOC sont ainsi reliés aux termes de NALT (Caracciolo *et al.*, 2012). Ces deux thésaurus sont publiés sur le Web de données et s'imposent aujourd'hui comme thésaurus de référence. A titre d'exemple, le vocabulaire d'AGROVOC est actuellement relié à celui de 11 ressources internationales comme GeoNames [6], DBpedia [7] et GEMET [8]. Plus récemment, de nombreuses initiatives ont été entreprises pour créer des ontologies permettant de décrire des connaissances sur les objets étudiés. Dans le domaine de la culture des plantes une ontologie de référence, Ref-TO, a été proposée (Arnaud *et al.*, 2012). Cette ontologie intègre actuellement trois ontologies déjà utilisées par des experts de différentes cultures des plantes. Elle sera, à terme, utilisée pour répondre à des enjeux majeurs comme par exemple la sécurité alimentaire. De manière similaire, des ontologies ont été proposées pour représenter des connaissances sur les capteurs en agriculture et environnement (Bendadouche *et al.*, 2012 ; Compton *et al.*, 2012), les relations microorganismes-habitat (Bossy *et al.*, 2012) ou encore le risque alimentaire microbiologique étendu aux emballages (Touhami *et al.*, 2011). La dernière initiative que nous évoquerons ici, et qui est indispensable dans les domaines des sciences du vivant et de l'environnement, a été la création d'ontologies permettant de représenter de manière standardisée des données quantitatives. En effet, ces ontologies permettent la description des caractéristiques mesurées sur les objets étudiés. On pourra notamment citer les ontologies QUDT [9] et OM (Rijgersberg *et al.*, 2013) qui sont actuellement les ontologies les plus complètes dans ce domaine.

A titre d'illustration, nous présentons dans cet article un système complet de capitalisation et de modélisation de données et de connaissances s'appuyant sur une Res-

4. http://aims.fao.org/standards/AGROVOC/
5. http://agclass.nal.usda.gov/agt.shtml
6. http://www.geonames.org/
7. http://dbpedia.org/
8. http://www.eionet.europa.eu/gemet/
9. http://www.qudt.org

source Termino-Ontologique (RTO) qui permet d'enrichir des bases locales à partir de données extraites de documents scientifiques. La notion de RTO provient des travaux de (Reymonet *et al.*, 2007), (Roche *et al.*, 2009), (McCrae *et al.*, 2011) et de (Cimiano *et al.*, 2011) qui ont proposé d'associer une partie terminologique et/ou linguistique aux ontologies afin d'établir une distinction claire entre la manifestation linguistique (le terme) et la notion qu'elle dénote (le concept). Ce système, appelé ONDINE (Ontology based Data INtEgration), s'inscrit de manière originale dans les problématiques de l'IC à l'intersection de différentes thématiques : acquisition de connaissances, représentation des connaissances, annotation sémantique, interrogation flexible, sous-ensemble flous, intégration de données, ontologie. Le système ONDINE est au service des experts pour le traitement informatique de leurs données, qui peuvent être complexes, hétérogènes, éparses ou encore imprécises. Il leur permet d'accéder à de l'information pertinente pour leur domaine d'étude, ce qui représente un enjeu particulièrement crucial dans les domaines des sciences du vivant et de l'environnement. Ce système s'appuie sur une RTO qui assure, par la sémantisation des données extraites, leur intégration effective et leur exploitation avec les données locales. Le rôle central de la RTO dans ce système présente en outre un intérêt majeur pour sa généricité, puisqu'elle permet de stocker toute la connaissance du domaine d'application étudié et qu'il "suffit" de modifier la RTO pour appliquer le système à un autre domaine.

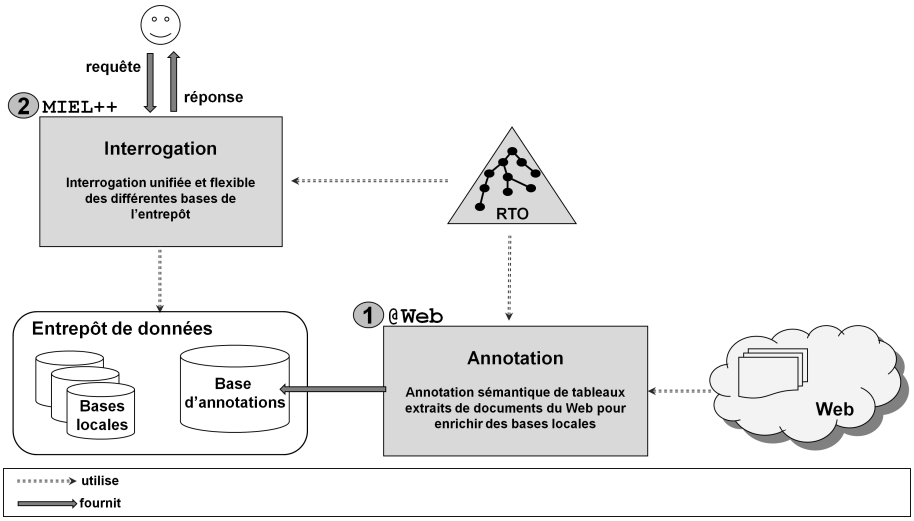


Figure 1. Architecture du système ONDINE

La RTO qui est au coeur du système ONDINE a été construite pour l'intégration de données quantitatives expérimentales. Le système ONDINE permet plus précisément d'acquérir, d'annoter et d'interroger des données extraites de tableaux trouvés dans des documents scientifiques (e.g. articles dans des revues, rapports), qui contiennent en général une synthèse des données quantitatives expérimentales publiées dans ces documents, ceci afin de pouvoir les exploiter et les traiter conjointement avec des données locales. Le système ONDINE est composé de deux sous-systèmes (cf. FI-

GURE 1): (1) un sous-système d'acquisition et d'annotation, appelé @Web (Annotating Tables from the Web), qui permet d'alimenter un entrepôt de données avec des données extraites de tableaux trouvés dans des documents scientifiques issus du Web qui ont été sémantiquement annotés avec des concepts de la RTO, ceci afin d'enrichir des bases locales; (2) un sous-système d'interrogation, appelé MIEL++ (Méthode d'Interrogation ELargie) (Buche *et al.*, 2013), qui propose un mécanisme d'interrogation unifiée et flexible des bases locales et de la base issue de l'annotation sémantique des données du Web. Le système ONDINE repose sur les standards du Web sémantique (XML, RDF, OWL, SPARQL) : la RTO est définie en OWL2-DL, la base de tableaux annotés en XML/RDF, les tableaux étant stockés en XML et leurs annotations en RDF, enfin, l'interrogation de la base d'annotations est effectuée en SPARQL.

Nous ne présenterons dans ce papier que le sous-système d'annotation @Web. Il est important de noter que ce sous-système n'a pas été construit pour annoter les tableaux de n'importe quels documents du Web, mais pour annoter de manière précise des tableaux ciblés extraits de documents identifiés comme pertinents pour un domaine donné, qui aura au préalable été décrit dans une RTO. Nous présenterons en section 5 l'originalité de ce sous-système par rapport à l'état de l'art sur l'annotation de tableaux. Le caractère semi-automatique de ce sous-système avec une validation manuelle à chaque étape est donc indispensable pour garantir un enrichissement des bases locales avec des données pertinentes et de qualité, qui sont ainsi directement exploitables avec les données locales. L'annotation précise de tableaux de données ciblés guidée par une RTO constitue la première originalité du sous-système @Web. Sa deuxième originalité repose sur la notion de relation n-aire définie dans la RTO qui permet d'annoter non pas des concepts indépendants mais des concepts reliés par une relation sémantique. Sa troisième originalité est de pouvoir instancier une relation n-aire dans un tableau à l'aide d'annotations floues avec une gestion distincte et adaptée des quantités imprécises et de leurs unités de mesure d'une part, et, des concepts symboliques d'autre part permettant d'apparier à l'aide de mesures de similarité les termes trouvés dans les cellules des tableaux et les termes définis dans la RTO.

Plusieurs articles ont déjà été publiés sur le système ONDINE ou ses sous-systèmes : (Hignette *et al.*, 2007), (Hignette *et al.*, 2009), (Buche *et al.*, 2009), (Touhami *et al.*, 2011) et (Buche *et al.*, 2013). Dans le dernier article en particulier, des résultats expérimentaux ont été présentés sur des corpus de tableaux à annoter dans trois domaines : le risque microbiologique, le risque chimique et l'aéronautique.

Nous présentons, dans la section 2, une nouvelle modélisation de la RTO permettant de représenter des relations n-aires sans arguments différenciés entre des données quantitatives expérimentales. Nous rappelons ensuite, dans la section 3, le sous-système @Web et son étape d'annotation sémantique semi-automatique de tableaux par des relations n-aires présentée dans (Touhami *et al.*, 2011) et (Buche *et al.*, 2013). Nous présentons, dans la section 4, un nouveau module du logiciel @Web permettant l'annotation sémantique manuelle de tableaux par des relations n-aires, annotation guidée par la RTO présentée dans la section 2. La section 5 positionne nos travaux par

rapport à l'état de l'art. Les exemples utilisés dans cet article concernent le domaine du risque alimentaire microbiologique étendu aux emballages [10].

2. La Ressource Termino-Ontologique

Le système ONDINE repose sur une Ressource Termino-Ontologique (RTO), qui est composée d'une ontologie à laquelle est associée une composante terminologique. Trois facteurs influencent la modélisation d'une RTO : la tâche à réaliser, le domaine d'intérêt et l'application (Reymonet *et al.*, 2007). Dans le système ONDINE, la tâche à réaliser est l'annotation et l'interrogation de tableaux de données par des relations n-aires. Un tableau peut être représenté par une ou plusieurs relation(s) n-aire(s), au sens des bases de données relationnelles. Les colonnes du tableau, correspondant aux différents arguments de la ou des relation(s), peuvent en effet être reliées par une ou plusieurs relations sémantiques. Ces relations sémantiques des tableaux de données nous intéressent particulièrement et la RTO a été définie pour les représenter. Dans le système ONDINE, le domaine d'intérêt est l'étude de données expérimentales quantitatives sachant que nous nous intéressons en particulier aux données dans le domaine des sciences du vivant. Il est à noter que les données quantitatives expérimentales contenues dans les tableaux requièrent la gestion des quantités avec leurs unités de mesure qui sont également pris en compte dans la RTO. Enfin, l'application du système ONDINE est la construction d'un entrepôt de données ouvert sur le Web.

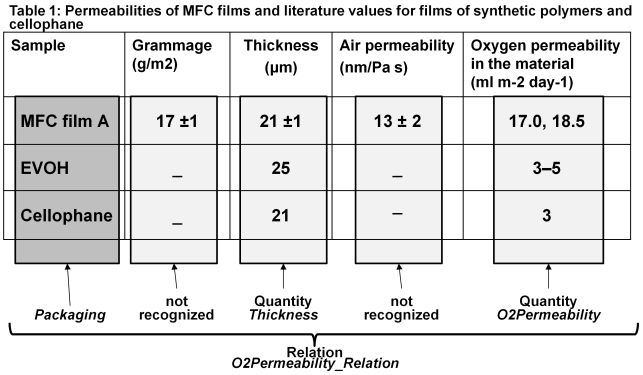


Figure 2. Un exemple de tableau annoté à partir de concepts définis dans une RTO dans le domaine du risque alimentaire microbiologique étendu aux emballages

EXEMPLE 1. — Le tableau présenté dans la FIGURE 2 est extrait d'un article scientifique dans le domaine du risque alimentaire microbiologique étendu aux emballages. Dans la dernière colonne du tableau se trouve la valeur de la perméabilité à l'oxygène d'un emballage (cf. première colonne) dans des conditions expérimentales données

10. L'impact des emballages et des transferts de gaz est pris en compte sur la croissance des microorganismes dans la matrice alimentaire.

par son épaisseur (cf. troisième colonne), son humidité relative, la différence de pression partielle à l'oxygène et la température ambiante. □

La RTO du système ONDINE, appelée dans la suite RTO naRyQ (n-ary Relations between Quantitative experimental data), a donc été définie pour représenter des relations n-aires entre des données quantitatives expérimentales. Nous avons choisi de représenter des relations n-aires sans arguments différenciés telles que recommandé par le W3C (N. Noy, Rector, 2006), ce qui correspond au cas 3 le plus général d'utilisation des relations n-aires. Nous avons de plus choisi d'utiliser le "patron 1" qui consiste à représenter une relation n-aire à l'aide d'un concept, relié à ses arguments par des propriétés. La relation n-aire *O2Permeability_Relation* utilisée pour annoter le tableau de la FIGURE 2 peut ainsi être représentée par le schéma de la FIGURE 3.

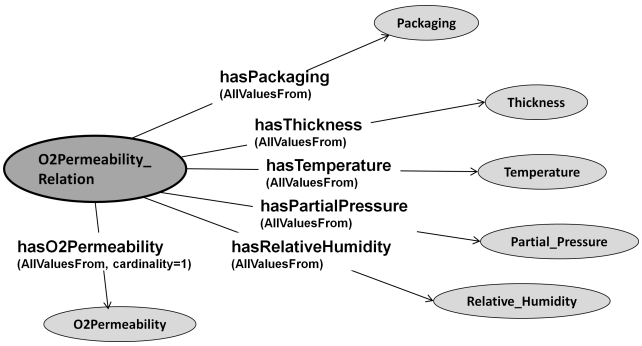


Figure 3. La relation n-aire O2Permeability_Relation

Nous présentons dans la suite la composante conceptuelle de la RTO naRyQ du système ONDINE, puis sa composante terminologique.

2.1. La composante conceptuelle

La composante conceptuelle de la RTO naRyQ du système ONDINE est composée de deux parties : une *ontologie noyau* qui permet de représenter des relations n-aires et une *ontologie de domaine* qui permet de représenter les concepts spécifiques à un domaine donné.

Nous nous intéressons dans cet article aux relations n-aires entre des données quantitatives expérimentales, ce qui suppose d'apporter une attention particulière à la gestion des arguments numériques (i.e. les quantités) et de leurs unités de mesure. Nous avons ainsi découpé l'**ontologie noyau** en 2 sous-parties (cf. FIGURE 4) : une partie supérieure, appelée *ontologie noyau supérieure*, qui permet de représenter des relations n-aires entre n'importe quels arguments, et une partie inférieure, appelée *ontologie noyau inférieure*, qui permet de représenter des relations n-aires entre des données quantitatives expérimentales. Ce découpage de l'ontologie noyau assure la généricité de la RTO naRyQ, qui peut non seulement être utilisée pour représenter des relations

n-aires entre des données quantitatives expérimentales, mais pourra aussi être réutilisée et étendue pour représenter des relations n-aires entre n'importe quels arguments.

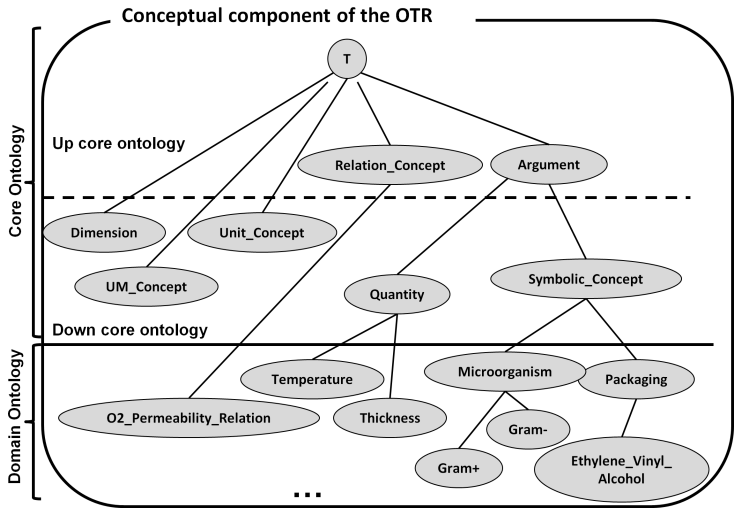


Figure 4. Un extrait de la RTO naRyQ_emb dans le domaine du risque alimentaire microbiologique étendu aux emballages

Dans l'ontologie noyau supérieure, les concepts génériques *Relation_Concept* et *Argument* permettent de représenter respectivement les relations n-aires et leurs arguments. Dans l'ontologie noyau inférieure, les concepts génériques *Dimension*, *UM_Concept*, *Unit_Concept* et *Quantity* permettent de gérer les quantités et leurs unités de mesure : le concept générique *Dimension* contient les dimensions qui permettent aux quantités et à leurs unités de mesures associées d'être classifiées (e.g. la quantité *Thickness* a pour dimension l'instance *length_dimension* du concept *Dimension*) et le concept générique *UM_Concept* contient les concepts qui permettent de gérer les conversions entre unités de mesure. Le concept générique *Symbolic_Concept* permet, quant à lui, de représenter les autres arguments (i.e. les arguments non numériques) des relations n-aires entre des données quantitatives expérimentales. Ces autres arguments permettent de représenter les objets d'étude (e.g. produits alimentaires, microorganismes, emballages) et les données exprimées de manière qualitative (e.g. croissance/non croissance/mort d'un microorganisme). On remarquera que pour définir des relations n-aires entre d'autres types de données, il faudra étendre la définition de l'ontologie noyau inférieure par l'ajout de nouveaux concepts, sous-concepts du concept générique *Argument*, permettant de prendre en compte les spécificités de leurs arguments.

L'**ontologie de domaine** contient les concepts spécifiques à un domaine d'application particulier. Ils apparaissent dans la RTO naRyQ comme des sous-concepts des concepts génériques de l'ontologie noyau. En OWL, tous les concepts sont représentés par des classes OWL, qui sont organisées hiérarchiquement à l'aide de la relation de subsomption *subClassOf* et sont deux à deux disjointes.

Nous présentons plus en détail dans la suite le concept générique *Relation_Concept* permettant de représenter les relations n-aires, le concept générique *Argument* permettant de représenter les arguments d'une relation n-aire, le concept générique *Unit_Concept* permettant de gérer les unités de mesure et le concept générique *UM_Concept* permettant de gérer les conversions entre unités de mesure. Nous illustrerons nos propos au travers de la présentation de la RTO naRyQ définie dans le domaine du risque alimentaire microbiologique étendu aux emballages, notée naRyQ_emb dans la suite pour plus de simplicité, et dont un extrait est présenté dans la FIGURE 4.

2.1.1. Le concept générique Relation_Concept

Le concept générique *Relation_Concept* permet de représenter une relation n-aire entre plusieurs arguments. Un concept relation est caractérisé par son label (i.e. un terme composé d'un ou plusieurs mots), défini dans la composante terminologique de la RTO naRyQ, et par sa signature qui permet de définir l'ensemble des concepts, sous-concepts du concept générique *Argument*, qui peuvent être reliés par la relation. Certains concepts de la signature d'un concept relation peuvent être obligatoires, c'est-à-dire qu'ils doivent nécessairement exister dans les instances du concept relation.

EXEMPLE 2. — Le concept relation *O2Permeability_Relation* présenté dans la FIGURE 3 est un des 16 concepts relations de la RTO naRyQ_emb. Ce concept relation permet de représenter la perméabilité à l'oxygène d'un emballage dans des conditions expérimentales données par son épaisseur, son humidité relative, la pression partielle à l'oxygène et la température ambiante. Sa signature est définie par : (*Packaging*, *Thickness*, *Temperature*, *Partial_Pressure*, *Relative_Humidity*, *O2Permeability*) où la quantité *O2Permeability* est obligatoire. □

2.1.2. Le concept générique Argument

Le concept générique *Argument* permet de représenter les arguments d'une relation n-aire. Sachant que nous ne nous intéressons ici qu'aux relations n-aires entre des données quantitatives expérimentales, il n'est le père que de deux sous-concepts : le concept générique *Quantity* et le concept générique *Symbolic_Concept*.

Un **concept symbolique**, sous-concept du concept générique *Symbolic_Concept*, est caractérisé par son label, défini dans la composante terminologique de la RTO naRyQ.

EXEMPLE 3. — La FIGURE 5 présente un extrait de la version actuelle de la hiérarchie des concepts symboliques de la RTO naRyQ_emb, qui contient 1087 concepts :

– 461 produits alimentaires (i.e. des sous-concepts de *Food_Product*),

– 185 micro-organismes (i.e. des sous-concepts de *Microorganism*),

– 150 emballages (i.e. des sous-concepts de *Packaging*),

– 288 facteurs impactant le comportement des micro-organismes (i.e. des sous-concepts de *Factor* non représenté sur la figure pour des raisons de place),

– et 3 réponses possibles (i.e. 3 sous-concepts de *Response*) : *growth*, *absence of growth* et *death*, qui représentent les comportements possibles d'un micro-organisme face à un traitement.

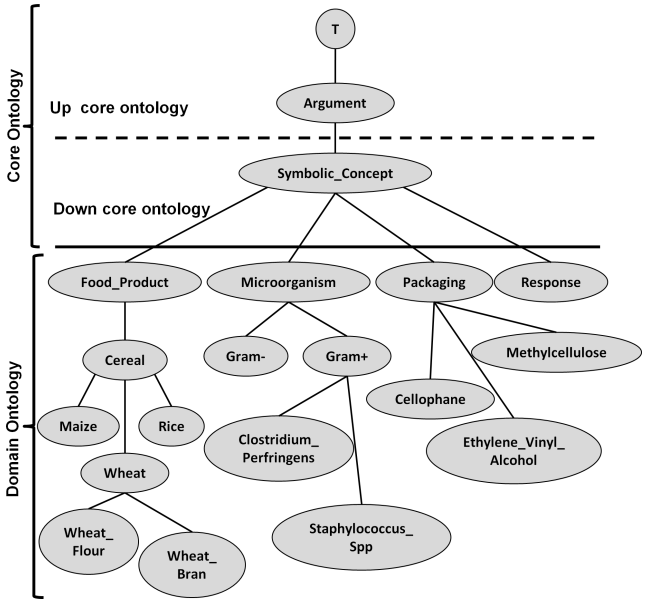


Figure 5. Un extrait de la hiérarchie des concepts symboliques de la RTO naRyQ_emb

□

On notera que nous n'avons pas pu réutiliser des terminologies existantes telles que AGROVOC [11] (de la FAO - Food and Agriculture Organisation des Nations Unies) ou NALT [12] (de la USDA - United States Department of Agriculture) pour représenter les produits alimentaires. En effet, ces terminologies ne sont pas assez spécifiques pour notre domaine d'intérêt. En utilisant la méthode d'alignement de termes proposée dans (Caracciolo *et al.*, 2012), nous obtenons seulement 30% (respectivement 30,26%) de termes en commun entre la RTO naRyQ_emb et AGROVOC (respectivement NALT).

Une **quantité**, sous-concept du concept générique *Quantity*, est caractérisée par son label, défini dans la composante terminologique de la RTO naRyQ, un ensemble d'unités de mesure, instances de sous-concepts du concept générique *Unit_Concept*, une dimension, instance du concept *Dimension*, et éventuellement un domaine de valeurs. Deux propriétés objets OWL ont été définies : (1) une propriété *hasUnitConcept* pour associer une quantité à l'ensemble de ses unités de mesure : elle a pour domaine

11. http://aims.fao.org/website/AGROVOC-Thesaurus
12. http://agclass.nal.usda.gov/agt.shtml

le concept générique *Quantity* et pour co-domaine le concept générique *Unit_Concept* et (2) une propriété *hasDimension* pour associer une quantité à sa dimension : elle a pour domaine le concept générique *Quantity* et pour co-domaine le concept générique *Dimension*. De plus, les restrictions de cardinalité OWL2 *minInclusive* et *maxInclusive* ont été utilisées pour représenter le domaine de valeurs d'une quantité.

EXEMPLE 4. — La FIGURE 6 présente un extrait de la hiérarchie des quantités de la RTO naRyQ_emb, qui contient 22 quantités. La quantité spécifique *Relative_Humidity*, par exemple, a pour unités les instances d'unités *Percent* et *One*, qui indique l'absence d'unité, pour dimension l'instance *Dimension_One*, et, pour domaine de valeurs l'intervalle [0, 100]. □

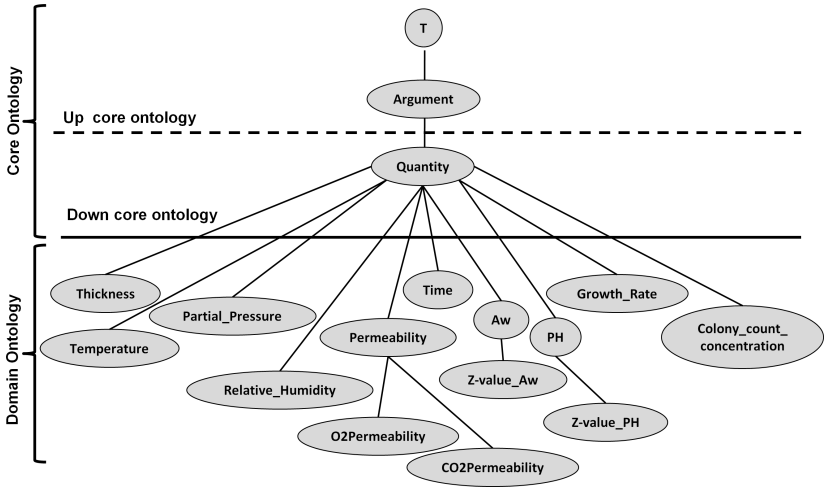


Figure 6. Un extrait de la hiérarchie des quantités de la RTO naRyQ_emb

2.1.3. Le concept générique Unit_concept

Les instances des quatre sous-concepts *Singular_Unit*, *Unit_Division_Or_Multiplication*, *Unit_Multiple_Or_SubMultiple* et *Unit_Exponentiation* du concept générique *Unit_Concept* permettent de représenter des unités de mesure. Un concept unité est caractérisé par son label, défini dans la composante terminologique de la RTO naRyQ, une dimension, instance du concept générique *Dimension*, et éventuellement des conversions.

Notre classification des unités de mesure repose sur le Système Internationnal des Unités [13]. Pour définir nos concepts unités, nous nous sommes inspirés de la modélisation des unités de mesure définies dans des ontologies existantes (OM [14], OBOE [15],

13. http://www.bipm.org/en/si/
14. http://www.wurvoc.org/vocabularies/om-1.8/
15. http://marinemetadata.org/references/oboeontology

QUDT [16], QUOMOS, ...). Nous en avons également définies de nouveaux pour les besoins de notre domaine d'application. Nous avons par exemple défini les concepts unités ppm [17] et CFU\per\gram [18].

EXEMPLE 5. — La FIGURE 7 présente un extrait de la hiérarchie des concepts unités avec leurs instances dans la RTO naRyQ_emb, qui contient 83 unités. □

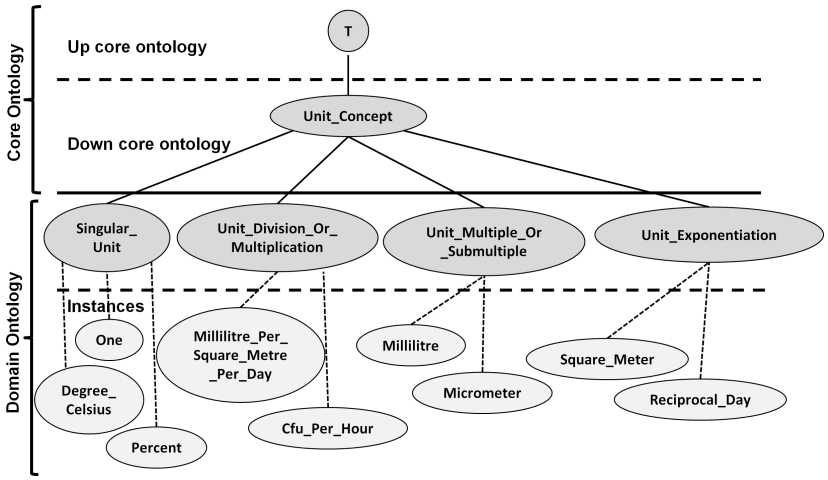


Figure 7. Un extrait de la hiérarchie des concepts unités avec leurs instances dans la RTO naRyQ_emb

2.1.4. Le concept générique UM_Concept

Le concept générique *Conversion*, sous-concept du concept générique *UM_Concept*, permet de gérer les conversions entre unités de mesure, qui sont des instances du concept générique *Unit_Concept*.

Les conversions entre unités de mesure modélisées dans la RTO naRyQ sont de la forme : $v_c = (v_s + o) * s$, où v_c est la valeur exprimée dans l'unité cible, v_s la valeur exprimée dans l'unité source, o un offset et s un facteur d'échelle. De nombreuses conversions entre unités de mesure peuvent être exprimées en utilisant un facteur d'échelle comme par exemple celles publiées par l'US National Institute of Standards and Technology [19]. Les conversions entre unités de mesure pour les températures requièrent l'introduction d'un offset additionnel [20].

16. http://www.qudt.org/
17. ppm, parts per million, est une unité de concentration souvent utilisée pour mesurer le niveau de polluants dans l'air, l'eau, les corps fluides, etc.
18. CFU\per\gram, Colony-Forming Units per gram, est utilisé pour mesurer le nombre de bactéries ou de fongiques viables en microbiologie.
19. http://ts.nist.gov/WeightsAndMeasures/Publications/appxc.cfm
20. e.g. http://en.wikipedia.org/wiki/Fahrenheit

Nous présentons la gestion des conversions entre unités de mesure à travers un exemple.

EXEMPLE 6. — Pour convertir une valeur de température exprimée en degrés Fahrenheit en degrés Celsius, nous utilisons la formule suivante :
$v_{Celsius} = (v_{\text{Faren}} - 32) \times \frac{5}{9}$. Dans la RTO naRyQ_emb, nous définissons l'instance *FahrenheitToCelsius* du concept *Conversion* (cf. FIGURE 8), où *Degree_Fahrenheit* et *Degree_Celsius* sont des instances du concept *Singular_unit*. □

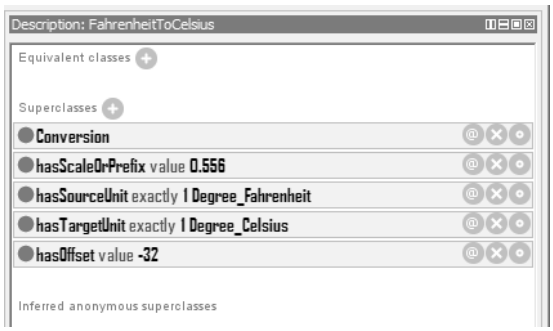


Figure 8. Un exemple de conversion entre degré Fahrenheit et degré Celsius

2.2. La composante terminologique

La composante terminologique de la RTO naRyQ contient l'ensemble des termes du domaine étudié. Comme nous l'avons précisé plus haut, les sous-concepts des concepts génériques *Relation_Concept*, *Symbolic_Concept* et *Quantity*, ainsi que les instances du concept générique *Unit_Concept*, sont chacun dénotés par au moins un terme de la composante terminologique. Chacun de ces sous-concepts ou instances est ainsi, dans une langue donnée, dénoté par un label préféré et éventuellement par un ensemble de labels alternatifs, qui correspondent à des synonymes ou des abréviations. Les labels sont associés à un concept ou une instance grâce aux propriétés SKOS de labellisation [21] (Simple Knowledge Organization Scheme), recommandées par le W3C. Par exemple, dans la FIGURE 9, les termes anglais *Ethylene vinyl alcohol* et *EVOH* dénotent le concept symbolique *Ethylene_Vinyl_Alcohol*.

3. L'annotation sémantique de tableaux par des relations n-aires

Nous avons présenté dans la section précédente la modélisation de la RTO naRyQ qui a été définie de manière générique pour représenter des relations n-aires entre des données quantitatives expérimentales. Cette RTO pourrait en effet facilement être étendue pour représenter des relations n-aires entre n'importe quels arguments. Cette

21. http://www.w3.org/TR/skos-reference/

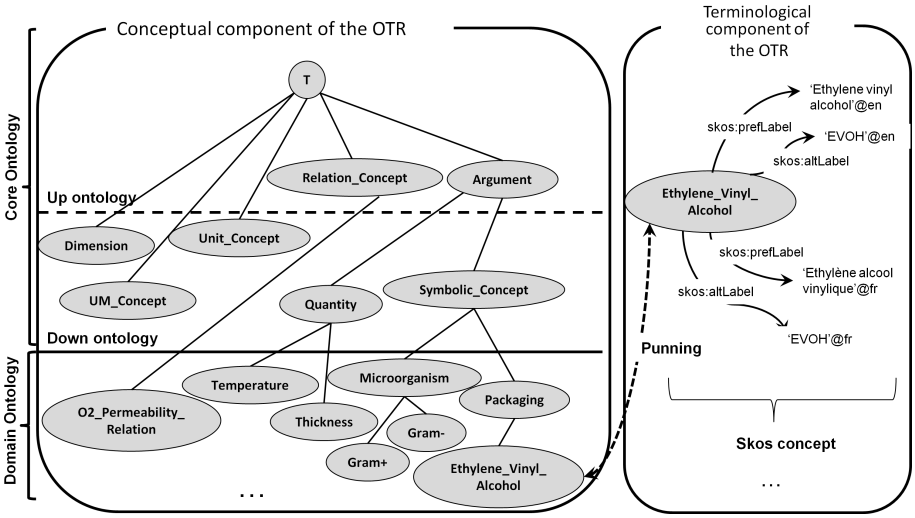


Figure 9. Un extrait de la RTO naRyQ_emb avec sa composante terminologique

RTO est utilisée par le système ONDINE notamment pour la tâche d'annotation de tableaux, contenant des données quantitatives expérimentales, par des relations n-aires, les tableaux étant issus de documents scientifiques (cf. FIGURE 1).

Nous nous concentrerons ici sur le sous-système d'annotation @Web du système ONDINE. La chaine de traitement de ce sous-système est composé de cinq grandes étapes (cf. FIGURE 10). Dans la première étape, des documents pertinents pour le domaine d'application étudié et décrit dans la RTO naRyQ sont trouvés sur le Web et filtrés manuellement par un expert du domaine. Dans la deuxième étape, les tableaux de données sont automatiquement extraits des documents et proposés à l'expert pour annotation. La troisième étape correspond à l'annotation sémantique des tableaux sélectionnés en utilisant la RTO naRyQ. Ces annotations sont ensuite, dans la quatrième étape, validées par un expert. Enfin, dans la cinquième étape, les tableaux et leurs annotations sont stockées dans une base d'annotations.

Le coeur du sous-système @Web repose sur sa troisième étape qui propose une méthode d'annotation semi-automatique de tableaux par des concepts relations définis dans la RTO naRyQ, la première étape qui concerne la sélection de documents n'étant qu'une entrée du sous-système et sa deuxième étape, une utilisation d'outils de détection de tableaux par la recherche de balises. La méthode d'annotation sémantique d'un tableau par des concepts relations consiste à associer de manière semi-automatique des annotations floues à chaque instance de concepts relations identifiés dans le tableau, une ligne du tableau pouvant être annotée par plusieurs concepts relations. Ces instances de concepts relations permettent d'annoter les données contenues dans les différentes lignes des tableaux avec le même vocabulaire que celui utilisé dans les bases locales et ainsi d'enrichir les bases locales avec de nouvelles données pertinentes, qui sont directement exploitables par les experts du domaine. Il est à noter

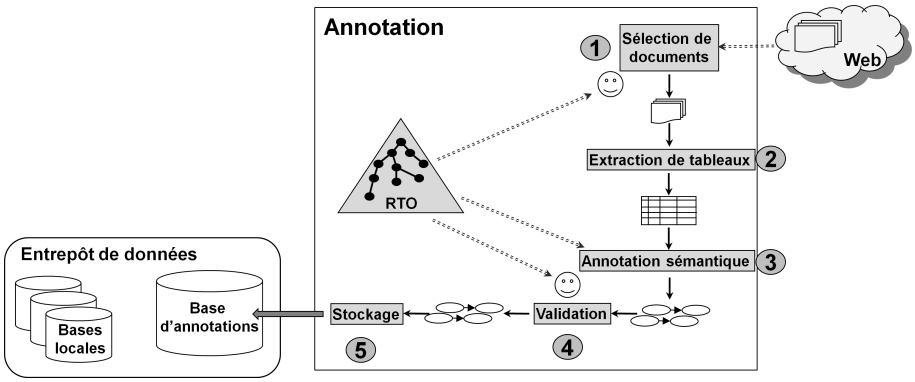


Figure 10. Sous-système @Web

que cette méthode permet d'annoter des tableaux de données ayant une structure similaire à celle du tableau présenté dans la FIGURE 2, à savoir la première ligne contient les entêtes de colonnes et les lignes suivantes les données correspondantes.

EXEMPLE 7. — Le sous-système @Web permet d'annoter trois des colonnes du tableau de la FIGURE 2 avec les concepts *Packaging*, *Thickness* et *O2Permeability* définis dans la RTO naRyQ_emb. Le tableau est alors partiellement annoté avec la relation n-aire *O2Permeability_Relation* qui a pour signature (*Packaging*, *Partial_Pressure*, *Relative_Humidity*, *Temperature*, *Thickness*, *O2Permeability*). Chaque ligne du tableau peut finalement être annotée par une instance de la relation n-aire *O2Permeability_Relation*. □

Nous présentons succinctement les différentes étapes de la méthode d'annotation de tableaux décrite dans la FIGURE 11 en illustrant nos propos sur l'annotation sémantique du tableau présenté en FIGURE 2 par la RTO naRyQ_emb. Une description détaillée de cette méthode est publiée dans (Buche *et al.*, 2013).

3.1. Distinction entre les colonnes symboliques et les colonnes numériques

La première étape de la méthode d'annotation sémantique d'un tableau consiste à distinguer les colonnes symboliques des colonnes numériques qui seront ensuite traitées différemment, en comptant, dans chaque colonne, le nombre d'occurences de valeurs numériques et de termes dénotant des concepts de la RTO naRyQ. Les termes dénotant des instances du concept générique *Unit_Concept* sont considérés comme devant apparaître dans des colonnes numériques. Dans cette étape, il est à noter que la première ligne du tableau, contenant le titre des colonnes, n'est pas prise en compte.

EXEMPLE 8. — Dans le tableau de la FIGURE 2, la première colonne est identifiée comme étant symbolique : elle ne contient que des termes qui ne dénotent pas d'instances du concept générique *Unit_Concept*. Les autres colonnes sont identifiées

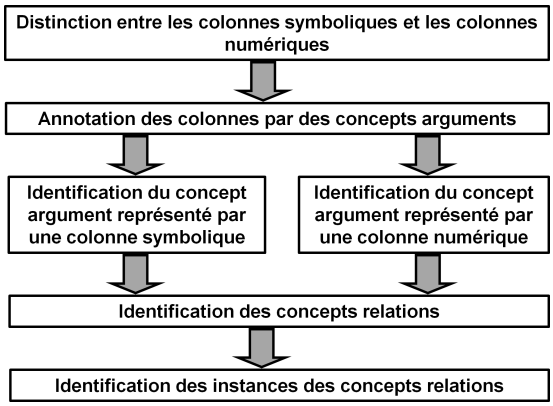


Figure 11. Les principales étapes de l'annotation sémantique d'un tableau guidée par la RTO naRyQ

comme étant numériques : elles ne contiennent que des valeurs numériques simples ou sous forme d'intervalles (e.g. 21 ± 1). □

3.2. Annotation des colonnes par des concepts arguments

Une fois que les colonnes ont été identifiées comme étant symboliques ou numériques, cette deuxième étape permet de déterminer quel concept argument de la RTO naRyQ annotera une colonne. Afin d'annoter une colonne avec un concept argument, deux scores sont calculés et combinés : le score du concept argument pour la colonne en fonction du titre de la colonne, et, le score du concept argument pour la colonne en fonction du contenu de la colonne. Etant donné que l'objectif de la méthode d'annotation est d'identifier quelles relations de la RTO naRyQ sont représentées dans un tableau, seuls les concepts arguments apparaisant dans les signatures des concepts relations de la RTO sont considérés ; ces concepts arguments sont appelés *concepts arguments cibles*.

EXEMPLE 9. — La RTO naRyQ_emb est composée de quatre concepts symboliques arguments cibles : *Food_Product*, *Microorganism*, *Packaging* and *Response* (cf. FIGURE 5) et de 22 quantités arguments cibles (cf. FIGURE 6). □

3.2.1. Identification du concept argument représenté par une colonne symbolique

L'annotation d'une colonne symbolique par un concept symbolique argument cible repose sur une comparaison entre les termes présents dans chaque cellule de la colonne et la liste des termes dénotant chaque concept symbolique argument cible ou l'un de ses sous-concepts de la RTO naRyQ (i.e. la liste de leurs labels préférés et alternatifs). Nous utilisons la mesure de similarité cosinus (Doan *et al.*, 2012) pour comparer deux termes t_1 et t_2, notée $sim(t_1, t_2)$, termes qui ont été au préalable transformés en vecteurs de mots lemmatisés à l'aide de WordNet.

EXEMPLE 10. — Considérons la première colonne du tableau *Table 1* de la FIGURE 2 qui a été identifiée comme étant symbolique. Cette colonne est annotée par le concept symbolique argument cible *Packaging* de la RTO naRyQ_emb avec le score suivant :

$$\begin{aligned} score_{Col}(col_1, Packaging) &= 1 - (1 - score_{TitreCol}(col_1, Packaging)) * \\ &\quad (1 - score_{ContenuCol}(col_1, Packaging)) \\ &= 1 - (1-0)(1-1) = 1. \end{aligned}$$

En effet,
$score_{TitreCol}(col_1, Packaging) = max_i sim(t_i, Sample) = 0,$
et
$score_{ContenuCol}(col_1, Packaging)$

$$= \frac{\#(\text{cellules de } col_1 \text{ annotees par } Packaging)}{\#(\text{cellules de } col_1)} = \frac{3}{3} = 1,$$

les trois cellules de la première colonne (i.e. *MFC film A*, *EVOH* et *Cellophane*) ayant pu être annotées par le concept *Packaging* par comparaison avec les termes de la RTO naRyQ_emb.

□

3.2.2. Identification du concept argument représenté par une colonne numérique

L'annotation d'une colonne numérique par une quantité argument cible repose sur les valeurs numériques et les unités de mesure présentes dans la colonne, les valeurs numériques devant respecter le domaine de valeurs définis dans la RTO naRyQ pour chaque quantité argument cible.

EXEMPLE 11. — Considérons la dernière colonne du tableau *Table 1* de la FIGURE 2 qui a été identifiée comme étant numérique. Cette colonne est annotée par le concept quantité argument cible *O2Permeability* avec le score suivant :

$$\begin{aligned} score_{Col}(col_5, O2Permeability) &= 1 - (1 - score_{TitreCol}(col_5, O2Perm)) * \\ &\quad (1 - score_{ContenuCol}(col_5, O2Perm)) \\ &= 1 - (1 - 0.816)(1 - 0.5) = 0.908. \end{aligned}$$

En effet,
$score_{TitreCol}(col_5, O2Permeability) = max_i sim(t_i, t_{titre}) = 0.816,$
et
$score_{ContenuCol}(col_5, O2Permeability)$

$$= \frac{1}{\#\{arg \mid MPSMPD \in hasUnitConcept(arg)\}} = \frac{1}{2} = 0.5,$$

où $\{arg \mid MPSMPD \in hasUnitConcept(arg)\}$ représente l'ensemble des quantités arguments cibles arg de la RTO naRyQ_emb ayant *Millilitre_Per_Square_Metre_Per_Day* (noté $MPSMPD$) comme unité de mesure. □

3.3. Identification des concepts relations

Une fois toutes les colonnes d'un tableau annotées à l'aide de concepts arguments cibles de la RTO naRyQ, la quatrième étape de la méthode d'annotation consiste à identifier quels concepts relations de la RTO sont représentés par le tableau. Afin d'annoter un tableau par un concept relation, deux scores sont calculés et combinés : le score du concept relation pour le tableau en fonction du titre du tableau et le score du concept relation pour le tableau en fonction du contenu du tableau. Ce deuxième score dépend de la proportion de concepts arguments cibles, apparaisant dans la signature du concept relation, qui sont représentés par des colonnes du tableau ; les concepts arguments cibles obligatoires dans la signature devant nécessairement être représentés par des colonnes du tableau. Notons de plus qu'un tableau peut être instancié par plusieurs concepts relations.

EXEMPLE 12. — D'après les exemples 10 et 11 la première colonne du tableau *Table 1* de la FIGURE 2 a été annotée par le concept symbolique argument cible *Packaging* et la troisième colonne par la quantité argument cible *O2Permeability*. Supposons de plus que la dernière colonne ait été annotée par la quantité argument cible *Thickness*, et, les deuxième et quatrième colonnes par le concept générique *Quantity*. Le tableau *Table 1* peut alors être annoté par le concept relation *O2Permeability_Relation* de la RTO naRyQ_emb, qui a pour concept argument cible obligatoire dans sa signature la quantité *O2Permeability* (cf. Exemple 2). Le score du concept relation *O2Permeability_Relation* pour le tableau *Table 1* en fonction de son contenu est :

$$score_{ContenuTab}(Table\ 1, O2Permeability\_Relation)$$
$$= \frac{\#(\text{concepts de la signature de Rel reconnus dans le tableau})}{\#(\text{concepts de la signature de Rel})} = \frac{3}{5} = 0.6.$$

Le score du concept relation *O2Permeability_Relation* pour le tableau en fonction de son titre *"Permeabilities of MFC films and literature values for films of synthetic polymers and cellophane"* est :

$$\begin{aligned} score_{TitreTab}(titre_{Table\ 1}, O2Permeability\_Relation) &= max_i sim(t_i, t_{titre}) \\ &= 0.35. \end{aligned}$$

Le score final du concept relation *O2Permeability_Relation* pour le tableau *Table 1* est donc :

$$score_{Tableau}(Table\ 1, O2Permeability\_Relation) = 1-(1-0.35)(1-0.6) = 0.74.$$

Comme il n'existe aucun autre concept relation dans la RTO naRyQ_emb dont les concepts arguments cibles obligatoires de la signature sont représentés dans les colonnes du tableau *Table 1*, alors le tableau *Table 1* est annoté par le concept relation *O2Permeability_Relation*. □

3.4. Identification des instances des concepts relations

La cinquième et dernière étape de la méthode d'annotation d'un tableau est l'identification, dans chaque ligne du tableau, des instances de chaque concept relation représenté par le tableau. L'identification des instances d'un concept relation repose sur l'identification des instances des concepts symboliques arguments cibles et des quantités arguments cibles apparaissant dans la signature du concept relation et représentées par des colonnes du tableau.

EXEMPLE 13. — L'instance du concept relation *O2Permeability_Relation* dans la deuxième ligne du tableau *Table 1* de la FIGURE 2 est représentée par l'ensemble de paires {(*valeur originale*, concepts arguments cibles reconnus : (valeurs d'annotation))} suivant :
{(*EVOH*, Packaging : (Ethylene Vinyl Alcohol)),
(*25 μm*, Thickness : (valeur : 25, unité de mesure : Micrometre)),
(*3-5 ml m-2 day-1*, O2Permeability : (intervalle de valeurs : [3, 5], unité de mesure : Millilitre_Per_Square_Metre_Per_Day)) }. □

4. Le logiciel @Web

Un premier prototype du sous-système d'annotation @Web (cf. FIGURE 10), dont l'étape d'annotation sémantique a été décrite dans la section 3, a été developpé et utilisé pour annoter de manière semi-automatique et en étant guidé par la RTO naRyQ des tableaux de données par des relations n-aires dans trois domaines : le risque microbiologique, le risque chimique et l'aéronautique (Buche *et al.*, 2013). Une nouvelle version ameliorée de ce premier prototype, appelé dans la suite logiciel @Web, est en cours de développement. Le logiciel @Web est actuellement composé de 2 modules : un module qui permet la gestion d'une RTO et un module qui permet d'annoter manuellement un tableau en étant guidé par une RTO, l'annotation étant faite à l'aide de concepts relations de la RTO. Nous avons fait le choix de commencer l'implémentation du logiciel @Web par un module d'annotation manuelle guidée par une RTO, et non semi-automatique en s'appuyant sur la méthode décrite dans la section 3, pour deux raisons essentielles. La première raison est que la méthode d'annotation semi-automatique guidée par une RTO s'applique aux tableaux ayant une structure similaire à celle du tableau présenté dans la FIGURE 2, à savoir la première ligne contient les entêtes de colonne et les lignes suivantes les données correspondantes. Cette hypothèse est respectée par de nombreux tableaux, mais le nouveau module d'@Web propose une solution plus générique permettant d'annoter n'importe quel tableau, quelqu'en soit sa structure, par des relations n-aires. Cette fonctionnalité est une demande forte des experts utilisateurs du système que nous n'avions pas prise en compte lors de la

réalisation du premier prototype dont l'objectif principal était d'évaluer notre méthode d'annotation semi-automatique de tableaux. La deuxième raison est que ce nouveau module, qui est actuellement utilisé et testé par des experts, nous permettra d'adapter et d'affiner l'implémentation de notre méthode d'annotation semi-automatique.

Le logiciel @Web repose sur la modélisation de la RTO naRyQ du système ONDINE permettant de représenter des relations n-aires entre des données quantitatives expérimentales, telle que décrite dans la section 2. Il permet de gérer l'ontologie de domaine de cette RTO avec sa terminologie associée, l'ontologie noyau, n'ayant pas vocation à être modifiée, n'est pas rendue accessible aux ontologues. Plusieurs RTO sur des domaines d'applications différents peuvent être gérées simultanément dans @Web. Les mêmes unités de mesure pouvant être utilisées dans ces différentes RTO, nous avons fait le choix de les gérer de manière transversale en définissant une seule RTO d'unités de mesure, de telle manière que l'ajout de nouvelles unités puisse potentiellement profiter à toutes les RTO définies dans @Web.

Le logiciel @Web implémente les cinq grandes étapes d'annotation de tableau présentées dans la FIGURE 10. Une démonstration du logiciel est accessible en ligne [22].

Dans la première étape, des documents pertinents pour le domaine d'application étudié et décrit dans la RTO naRyQ sont fournis par un expert du domaine qui les aura au préalable sélectionnés. Cette sélection peut être réalisée en utilisant des outils de recherche bibliographiques classiques (par exemple FSTA [23] dans le domaine de la science des aliments) qui permettent de constituer une collection de documents scientifiques pertinents en paramétrant une requête combinant plusieurs mots-clés. @Web permet de charger ces documents depuis un poste de travail. Il permet également de charger une collection de documents gérés avec le logiciel de gestion bibliographique collaboratif Mendeley [24]. Après chargement, @Web stocke à la fois les références bibliographiques du document et le texte intégral au format HTML et PDF si celui-ci est disponible. Ces documents sont ensuite téléchargeables par l'expert (cf. FIGURE 12).

Dans la deuxième étape, les tableaux de données sont extraits des documents au format HTML par analyse automatique des balises. @Web présente à l'expert les tableaux qu'il a identifiés dans le document pour validation, ce dernier peut alors décider de les conserver ou non pour pouvoir ensuite les annoter sémantiquement.

La troisième étape correspond à l'annotation sémantique manuelle guidée par une RTO des tableaux sélectionnés en utilisant les concepts de la RTO. Au regard du contenu du tableau original extrait du document, l'expert choisit dans la RTO le ou les concept(s) relation(s) pertinent(s) pour annoter le tableau. Par exemple, dans la FIGURE 13, les deux concepts relations *O2Permeability_Relation* et *CO2Permeability_Relation* sont sélectionnés dans la liste déroulante qui présente l'ensemble des concepts relations définies dans la RTO naRyQ_emb, c'est-à-dire les concepts rela-

22. http://www.paris.inra.fr/metarisk/research_unit/knowledge_engineering/softwares/web__1
23. http://www.ifis.org/fsta/
24. http://fr.wikipedia.org/wiki/Mendeley

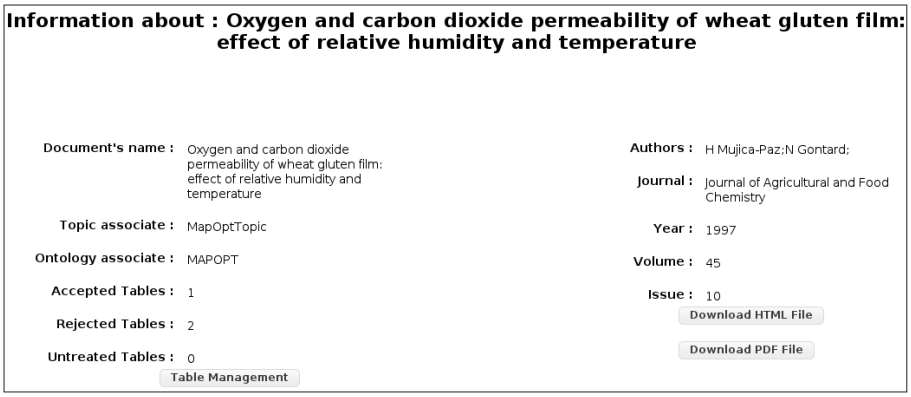


Figure 12. Les références bibliographiques d'un document chargé dans @Web. Cette page permet d'accéder au document intégral au format PDF ou HTML

tions qui ont été identifiés comme étant d'intérêt pour le domaine étudié. Les signatures de ces deux concepts relations sont visualisées dans un format tabulaire, une signature de chaque concept relation par ligne. La signature du concept relation va permettre de guider l'expert dans sa saisie. Elle lui permet notamment de ne pas oublier d'arguments, de renseigner les arguments obligatoires du concept relation (i.e. *CO2_Permeability* et *O2_Permeability*, identifiés par une coloration en rouge dans le logiciel et par des zones grisées dans la FIGURE 13) et de renseigner les arguments importants du concept relation (identifiés par une coloration en vert dans le logiciel et par des zones hachurées dans la FIGURE 13).

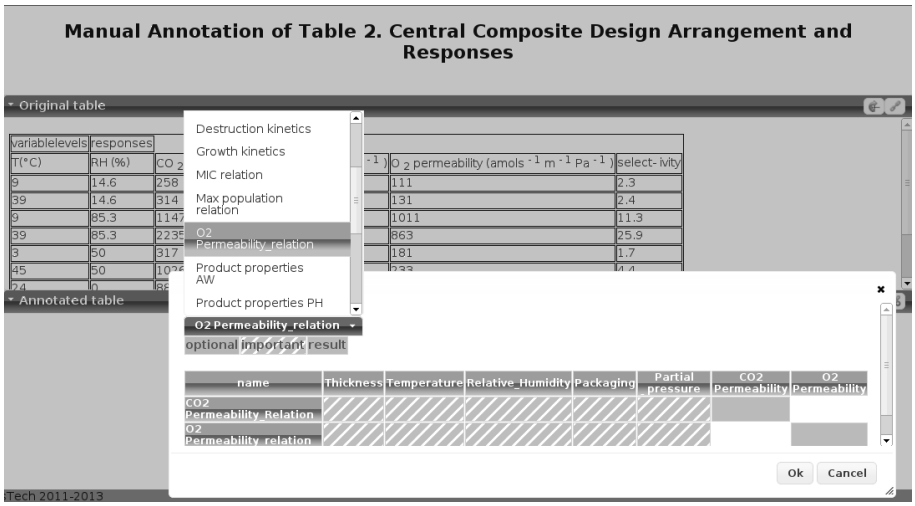


Figure 13. Sélection des concepts relations de la RTO naRyQ_emb pertinents pour annoter le tableau d'origine figurant dans la partie supérieure de la fenêtre

Lors de la saisie manuelle guidée par une RTO, le logiciel @Web propose une aide à la saisie des quantités et de leurs unités de mesure en permettant des conversions automatiques entre les unités. Il aide ainsi l'expert à déterminer dans quelles unités de mesure les quantités apparaissant dans la signature d'un concept relation doivent être enregistrées. L'expert peut en effet sélectionner une unité dans la liste des unités de mesure associée à la quantité définie dans la RTO. La FIGURE 14 présente la liste des unités de mesure associées à la quantité *CO2Permeability* dans la RTO naRyQ_emb. Une fois les unités de mesure associées aux quantités choisies, la saisie des instances de concepts relations peut commencer. En ce qui concerne les cellules correspondant à des quantités, l'expert peut cliquer/glisser les valeurs depuis le tableau d'origine dans le tableau annoté, ce qui permet de faciliter la saisie et de limiter les risques d'erreurs de saisie. L'unité de mesure choisie pour enregistrer le tableau annoté ne doit pas nécessairement correspondre à celle utilisée dans le tableau d'origine. Si elles sont différentes, le facteur de conversion entre les deux unités, défini dans la RTO sélectionnée, est automatiquement appliqué. Il est également possible de saisir une quantité sous la forme d'un intervalle de valeurs. Par exemple, dans la FIGURE 16, la quantité *Thickness* est définie dans l'intervalle [77, 83] μm.

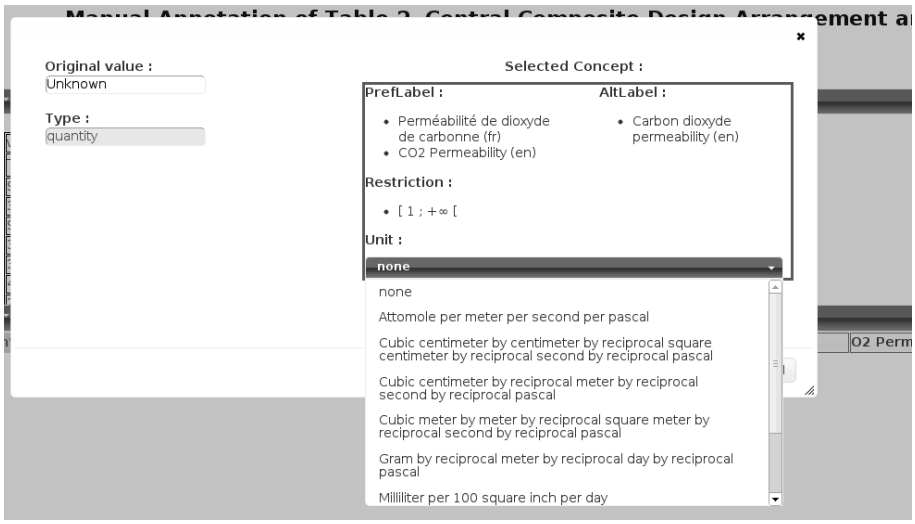


Figure 14. Sélection dans la RTO naRyQ_emb d'une unité de mesure pour l'associer à la quantité CO2Permeability

Le logiciel @Web propose également une aide à la saisie des concepts symboliques. En ce qui concerne les cellules correspondant à des concepts symboliques, @Web permet de naviguer dans la hiérarchie de concepts symboliques de la RTO sélectionnée pour aider l'expert à remplir le tableau annoté à l'aide d'un vocabulaire contrôlé et validé, qui permettra d'exploiter les données saisies conjointement avec les données locales. Par exemple, dans la FIGURE 15, la colonne *Packaging* est renseignée avec des noms de concepts sélectionnés dans la hiérarchie de spécialisation du concept *Packaging* extraite de la RTO naRyQ_emb (e.g. le concept *Proteins*).

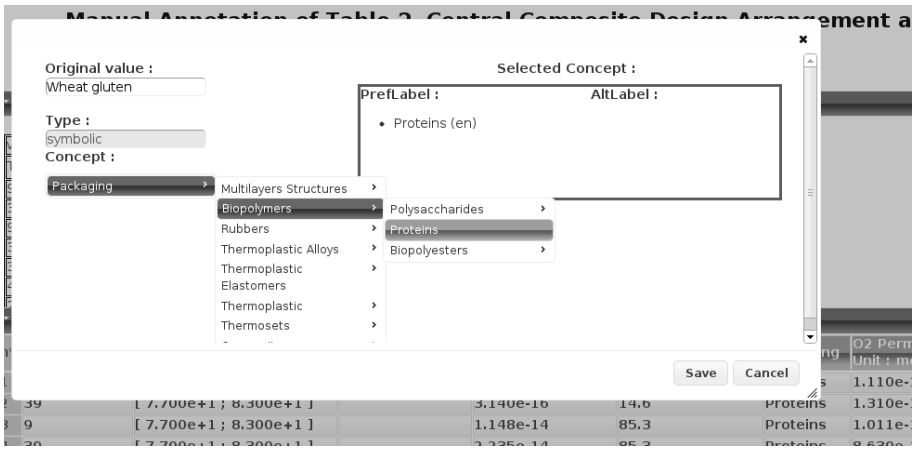
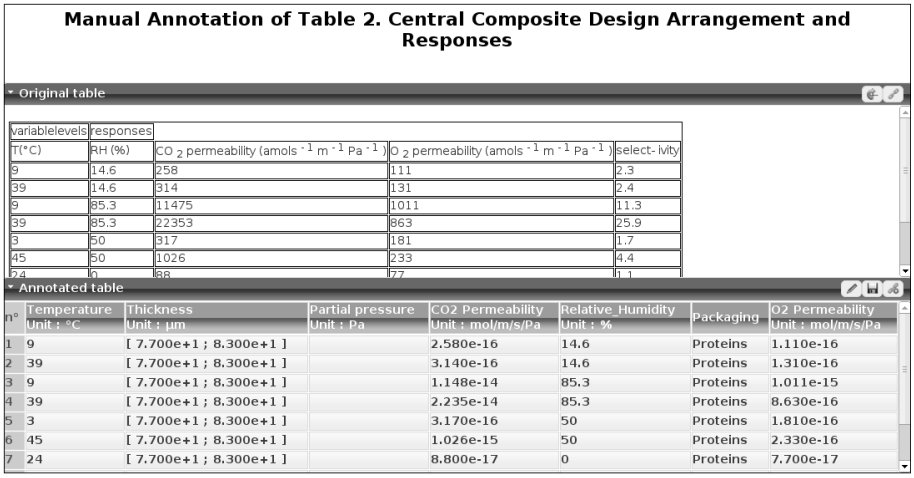


Figure 15. Annotation de la première ligne du tableau d'origine avec la RTO naRyQ_emb. La fenêtre blanche, au premier plan, permet de naviguer dans la hiérarchie de spécialisation du concept Packaging.

Il est à noter que certains arguments du ou des concept(s) relation(s), retenus pour annoter un tableau, ne sont pas forcément renseignés dans ce tableau. Considérons le tableau de la FIGURE 16. La colonne *Thickness* est renseignée à partir d'une information trouvée dans la section *Material & Method* du document. La colonne *Partial Pressure* n'est pas renseignée car cette information n'est pas présente dans le document. En effet, seule la méthode employée pour mesurer la perméabilité au CO2 est décrite dans la section *Material & Method*. L'expert devra donc retrouver, par lui-même, la description de cette méthode dans un autre document pour pouvoir compléter le tableau annoté.



Manual Annotation of Table 2. Central Composite Design Arrangement and Responses

Original table

variablelevels	responses			
T(°C)	RH (%)	CO_2 permeability (amols $^{-1}$ m $^{-1}$ Pa $^{-1}$)	O_2 permeability (amols $^{-1}$ m $^{-1}$ Pa $^{-1}$)	select- ivity
9	14.6	258	111	2.3
39	14.6	314	131	2.4
9	85.3	11475	1011	11.3
39	85.3	22353	863	25.9
3	50	317	181	1.7
45	50	1026	233	4.4
24	0	88	77	1.1

Annotated table

n°	Temperature Unit : °C	Thickness Unit : µm	Partial pressure Unit : Pa	CO2 Permeability Unit : mol/m/s/Pa	Relative_Humidity Unit : %	Packaging	O2 Permeability Unit : mol/m/s/Pa
1	9	[7.700e+1 ; 8.300e+1]		2.580e-16	14.6	Proteins	1.110e-16
2	39	[7.700e+1 ; 8.300e+1]		3.140e-16	14.6	Proteins	1.310e-16
3	9	[7.700e+1 ; 8.300e+1]		1.148e-14	85.3	Proteins	1.011e-15
4	39	[7.700e+1 ; 8.300e+1]		2.235e-14	85.3	Proteins	8.630e-16
5	3	[7.700e+1 ; 8.300e+1]		3.170e-16	50	Proteins	1.810e-16
6	45	[7.700e+1 ; 8.300e+1]		1.026e-15	50	Proteins	2.330e-16
7	24	[7.700e+1 ; 8.300e+1]		8.800e-17	0	Proteins	7.700e-17

Figure 16. Le tableau annoté avec la RTO naRyQ_emb

La quatrième étape de validation des annotations par l'expert est effectuée au fur et à mesure de l'étape d'annotation sémantique manuelle guidée par une RTO.

Enfin, la cinquième et dernière étape correspond au stockage des tableaux et de leurs annotations dans une base d'annotations au format RDF. Les tableaux et leurs annotations peuvent également être exportés au format CSV pour pouvoir être exploités par d'autres logiciels.

Le logiciel @Web propose donc, dans sa version actuelle, un processus complet d'acquisition de données issus de tableaux trouvés dans des documents scientifiques, allant de l'extraction des tableaux dans les documents à l'annotation sémantique de chacune de leurs lignes par des concepts relations définis dans la RTO naRyQ afin de pouvoir les intégrer avec les données locales.

5. Comparaison avec l'état de l'art

Le besoin d'extraire de l'information dans les tableaux du Web ont conduit à des travaux qui proposent des méthodes pour annoter sémantiquement ces tableaux, afin de pouvoir ensuite les interroger (Liu *et al.*, 2007 ; Cafarella, Halevy, Zhang *et al.*, 2008 ; Cafarella, Halevy, Wang *et al.*, 2008 ; Assem *et al.*, 2010 ; Limaye *et al.*, 2010 ; Venetis *et al.*, 2011).

TableSeer (Liu *et al.*, 2007) permet l'extraction d'un ensemble prédéfini de métadonnées (légende, position du tableau dans la page HTML, ...) à partir de tableaux de données du Web, mais ne permet pas la comparaison du schéma des tableaux avec des schémas définis dans une ontologie. Le système WebTables, présenté dans (Cafarella, Halevy, Wang *et al.*, 2008) et (Cafarella, Halevy, Zhang *et al.*, 2008), permet d'identifier des tableaux relationnels dans un corpus de documents HTML et de les indexer, ceci afin de pouvoir les interroger. Cependant, le langage d'interrogation de WebTables permet uniquement une interrogation par combinaison de mots-clés qui sont comparés avec les titres de colonnes des tableaux. Le contenu des lignes n'est pas exploité dans l'interrogation qui repose uniquement sur des statistiques de fréquences des cooccurrences des noms de colonnes.

Dans (Tenier *et al.*, 2006), les relations de l'ontologie sont instanciées avec différents types de structures HTML incluant les tableaux. Cependant, ce système permet uniquement l'identification de relations binaires entre des instances de concepts en faisant l'hypothèse que ces instances de concepts ont été préalablement annotées (soit manuellement, soit par un autre système d'annotation). Dans (Limaye *et al.*, 2010) et (Venetis *et al.*, 2011), les auteurs proposent des méthodes pour annoter les colonnes de tableaux extraits du Web et les tableaux eux-mêmes avec des relations binaires. Notre travail prolonge ces approches puisque nous proposons une méthode de reconnaissance d'une relation n-aire comprenant la reconnaissance et l'instanciation des concepts apparaissant dans les tableaux.

L'approche de (Embley *et al.*, 2002) est voisine de la nôtre dans la mesure où les auteurs transforment des tableaux de données de structures différentes dans un

même schéma relationnel incluant des relations n-aires. Cependant notre méthode étend celle de (Embley *et al.*, 2002) de plusieurs manières : meilleure distinction dans l'ontologie entre concepts et terminologie, annotation des cellules avec des ensembles de termes similaires ou avec des valeurs numériques imprécises, interrogation flexible de relations n-aires gérant la comparaison avec des annotations floues, cette dernière fonctionnalité étant présentée dans (Buche *et al.*, 2013).

Le travail de (Assem *et al.*, 2010) concerne l'identification des quantités dans les tableaux du Web, en utilisant une ontologie d'unités de mesures. Des heuristiques définissent des stratégies pour désambigüer les termes qui font références à des quantités différentes. Ce travail est complémentaire du nôtre et pourrait être intégré comme une extension de notre méthode de reconnaissance de quantités.

Le seul outil aujourd'hui comparable au logiciel @Web est, à notre connaissance, Rosanne (Rijgersberg *et al.*, 2011), une application de l'ontologie OM, ontologie de quantités et d'unités de mesure, développée comme un "add-in" Excel. Cette application permet d'annoter les quantités et les unités de mesure associées aux colonnes d'un tableau Excel avec l'ontologie OM. Elle permet également de faire des conversions entre unités de mesure. Par contre, à la différence du logiciel @Web, elle ne gère ni la notion de concept symbolique qui permet de représenter les données non numériques, comme par exemple les objets d'étude, ni la notion de concept relation permettant de relier entre elles des données quantitatives expérimentales. Les auteurs de Rosanne ont fait le choix de développer un "add-in" Excel qui a l'avantage de proposer une sémantisation des données à l'aide d'Excel, un outil connu et largement utilisé par les experts dans les domaines à caractère scientifique. Par contre, Rosanne ne permet pas de gérer le partage de données, ni le travail collaboratif, contrairement au logiciel @Web, développé sous la forme d'une application Web. Ces deux outils sont donc complémentaires et reposent sur une représentation ontologique en partie commune, la composante quantités-unités de mesure de la RTO naRyQ du système ONDINE étant très proche de celle d'OM.

6. Conclusion

Nous avons présenté dans cet article le système ONDINE, un système complet de capitalisation et de modélisation de données et de connaissances s'appuyant sur une RTO naRyQ qui permet d'enrichir des bases locales à partir de données extraites de documents scientifiques. La RTO naRyQ est définie pour modéliser des relations n-aires entre des données quantitatives expérimentales et est exprimée en OWL2-DL et SKOS. La RTO est structurée en une partie noyau et une partie domaine. La partie noyau est générique pour la modélisation de relations n-aires entre des données quantitatives expérimentales. Pour construire une RTO du domaine du risque alimentaire microbiologique étendu aux emballages, nous avons ajouté à la partie noyau environ 1 100 concepts du domaine, ainsi que des concepts pour la gestion des unités de mesure. La composante terminologique de la RTO permet actuellement de gérer les synonymes et les abréviations, en anglais et en français. Le sous-système @Web du

système ONDINE repose sur la RTO naRyQ et permet d'extraire, à partir des tableaux trouvés dans des documents, de nouvelles données quantitatives expérimentales. Ces nouvelles données sont annotées avec des concepts de la RTO et enregistrées dans une base d'annotations au format RDF. La version d'annotation semi-automatique de @Web, développée sous la forme d'un premier prototype, a été utilisée pour annoter trois corpus de tableaux de trois domaines: le risque microbiologique, le risque chimique et l'aéronautique (Buche *et al.*, 2013). La version actuelle du logiciel @Web, décrite dans la section 4, est un assistant à la saisie manuelle, qui est en cours de développement. Ce logiciel est actuellement utilisé dans le cadre du projet ANR ALIA Map'OPT pour saisir et sémantiser des données de perméabilité d'emballages à l'oxygène et au dioxide de carbone, ainsi que des données de solubilité et de diffusivité de l'oxygène et du dioxide de carbone dans les aliments.

Le sous-système @Web permet de lever plusieurs verrous méthodologiques en intégration et capitalisation des données et des connaissances. Il propose des solutions qui permettent de prendre en compte l'hétérogénéité du vocabulaire utilisé et de la structure des sources de données. Il permet de représenter l'imprécision des données quantitatives. De plus, la notion de relation n-aire définie dans la RTO naRyQ, sur laquelle repose le sous-système @Web, est une contribution originale par rapport à l'état de l'art. Elle permet d'annoter des tableaux de données en représentant la relation sémantique reliant les concepts associés à chacune des colonnes des tableaux. Cette notion est indispensable pour assurer une réelle capitalisation des données expérimentales. En effet, la notion de relation n-aire permet de définir explicitement l'ensemble des facteurs contrôlés qui doivent être renseignés pour que les données expérimentales soient réutilisables et exploitables, notamment dans des outils d'aide à la décision (Destercke *et al.*, 2011). Enfin, le sous-système @Web est une contribution à l'initiative du Web de données dont l'objectif est de permettre la réutilisation des données en leur associant une couche sémantique. Dans @Web, cette couche sémantique est définie grâce à la RTO naRyQ en utilisant les standards du Web sémantique : le langage OWL pour représenter les concepts et le langage SKOS pour représenter la terminologie associée aux concepts.

Plusieurs perspectives sont actuellement en cours d'exploration. Dans la nouvelle version du logiciel @Web, la méthode d'annotation de tableaux présentée dans la section 3 pourra être combinée avec le module d'annotation manuelle pour aider l'expert à annoter ses tableaux. Par exemple, @Web pourrait permettre de proposer des concepts relations candidats à l'annotation d'un tableau, si la structure de ce dernier est conforme à ce qui est exigé par la méthode d'annotation, à savoir la première ligne contient les entêtes de colonnes et les lignes suivantes les données correspondantes. Une fois ces relations validées par l'expert, @Web pourrait également proposer une instanciation des lignes du tableau en s'appuyant sur la méthode d'annotation décrite dans la section 3. Par ailleurs, d'un point de vue méthodologique, nous travaillons actuellement sur deux axes de recherche. D'une part, nous concevons actuellement une méthode (et un outil) qui permet de gérer les évolutions de la RTO du système ONDINE. Nous nous intéressons plus particulièrement au problème de la gestion de l'évolution de concepts interdépendants (Touhami *et al.*, 2013) (e.g. un concept rela-

tion avec ses arguments ou une quantité avec ses unités). D'autre part, comme nous l'avons évoqué dans la présentation du logiciel @Web (cf. section 4), les arguments d'une instance de concept relation ne sont pas toujours renseignés dans le tableau à annoter. Nous étudions la possibilité d'étendre la méthode d'annotation de tableaux présentée dans la section 3 pour pouvoir rechercher les informations manquantes dans le texte du document et ne plus se limiter aux seuls tableaux (Ghersedine *et al.*, 2012).

Remerciements

Le travail de recherche ayant mené aux résultats présentés dans cet article a reçu le soutien de l'Agence Nationale de la Recherche (ANR) dans la cadre du projet ALIA MAP'OPT.

Bibliographie

Amarger F., Haemmerlé O., Hernandez N., Pradel C. (2013). Taking SPARQL 1.1 Extensions into Account in the SWIP System. In H. D. Pfeiffer, D. I. Ignatov, J. Poelmans, N. Gadiraju (Eds.), *ICCS*, vol. 7735, p. 75-89. Springer.

Arnaud E., Cooper L., Shrestha R., Menda N., Nelson R. T., Matteis L. *et al. (2012). Towards a Reference Plant Trait Ontology for Modeling Knowledge of Plant Traits and Phenotypes. In J. Filipe, J. L. G. Dietz (Eds.),* KEOD*, p. 220-225. SciTePress.*

Assem M. van, Rijgersberg H., Wigham M., Top J. L. (2010). Converting and Annotating Quantitative Data Tables. In P. F. Patel-Schneider et al. (Eds.), *International Semantic Web Conference (1)*, vol. 6496, p. 16-31. Springer.

Aussenac-Gilles N., Charlet J., Reynaud-Delaître C. (2012). Chapitre 7 - les enjeux de l'ingénierie des connaissances. In F. Sèdes, J.-M. Ogier, P. Marquis (Eds.), *Information-Interaction-Intelligence : le point sur le I3*, p. 244-266. Toulouse, Cépaduès.

Bendadouche R., Roussey C., Sousa G. D., Chanet J.-P., Hou K. M. (2012). Extension of the Semantic Sensor Network Ontology for Wireless Sensor Networks: The Stimulus-WSNnode-Communication Pattern. In C. A. Henson, K. Taylor, Ó. Corcho (Eds.), *SSN*, vol. 904, p. 49-64. CEUR-WS.org.

Bernstein P. A., Madhavan J., Rahm E. (2011). Generic schema matching, ten years later. *PVLDB*, vol. 4, n° 11, p. 695-701.

Bossy R., Jourde J., Manine A.-P., Veber P., Alphonse É., Guchte M. van de *et al. (2012). Bionlp shared task - the bacteria track.* BMC Bioinformatics*, vol. 13, n° S-11, p. S3.*

Buche P., Dibie-Barthélemy J., Chebil H. (2009). Flexible SPARQL Querying of Web Data Tables Driven by an Ontology. In T. Andreasen, R. R. Yager, H. Bulskov, H. Christiansen, H. L. Larsen (Eds.), FQAS*, vol. 5822, p. 345-357. Springer.*

Buche P., Dibie-Barthélemy J., Ibanescu L., Soler L. (2013). Fuzzy Web Data Tables Integration Guided by an Ontological and Terminological Resource. IEEE Trans. Knowl. Data Eng.*, vol. 25, n° 4, p. 805-819.*

Cafarella M. J., Halevy A. Y., Wang D. Z., Wu E., Zhang Y. (2008). WebTables: exploring the power of tables on the web. PVLDB*, vol. 1, n° 1, p. 538-549.*

Cafarella M. J., Halevy A. Y., Zhang Y., Wang D. Z., Wu E. (2008). Uncovering the relational web. In WebDB.

Caracciolo C., Stellato A., Rajbhandari S., Morshed A., Johannsen G., Keizer J. et al. (2012). Thesaurus maintenance, alignment and publication as linked data: the AGROVOC use case. *IJMSO*, vol. 7, nº 1, p. 65-75.

Cimiano P., Buitelaar P., McCrae J., Sintek M. (2011). LexInfo: A declarative model for the lexicon-ontology interface. *J. Web Sem.*, vol. 9, nº 1, p. 29-51.

Cimiano P., Buitelaar P., Völker J. (2010). Ontology construction. In N. Indurkhya, F. J. Damerau (Eds.), *Handbook of natural language processing, second edition*, p. 577-604. Boca Raton, FL, CRC Press, Taylor and Francis Group. (ISBN 978-1420085921)

Compton M., Barnaghi P. M., Bermudez L., Garcia-Castro R., Corcho Ó., Cox S. *et al. (2012). The SSN ontology of the W3C semantic sensor network incubator group.* J. Web Sem.*, vol. 17, p. 25-32.*

Corby O., Dieng-Kuntz R., Faron-Zucker C., Gandon F. L. (2006). Searching the Semantic Web: Approximate Query Processing Based on Ontologies. IEEE Intelligent Systems*, vol. 21, nº 1, p. 20-27.*

Destercke S., Buche P., Charnomordic B. (2013). Evaluating Data Reliability: An Evidential Answer with Application to a Web-Enabled Data Warehouse. IEEE Trans. Knowl. Data Eng.*, vol. 25, nº 1, p. 92-105.*

Destercke S., Buche P., Guillard V. (2011). A flexible bipolar querying approach with imprecise data and guaranteed results. Fuzzy Sets and Systems*, vol. 169, nº 1, p. 51-64.*

Doan A., Halevy A. Y., Ives Z. G. (2012). Principles of data integration*. Morgan Kaufmann.*

Embley D. W., Tao C., Liddle S. W. (2002). Automatically extracting ontologically specified data from html tables of unknown structure. In S. Spaccapietra, S. T. March, Y. Kambayashi (Eds.), Er*, vol. 2503, p. 322-337. Springer.*

Ghersedine A., Buche P., Dibie-Barthélemy J., Hermandez N., Kamel M. (2012). Extraction de relations n-aires interphrastiques guidée par une RTO. In M. Beigbeder, V. Eglin, N. Ragot, M. Géry (Eds.), CORIA*, p. 179-190.*

Guarino N., Oberle D., Staab S. (2009). What is an ontology? In S. Staab, R. Studer (Eds.), Handbook on ontologies*, p. 1-17. Springer Berlin Heidelberg.*

Heath T., Bizer C. (2011). Linked Data: Evolving the Web into a Global Data Space *(vol. 1) nº 1. Morgan & Claypool.*

Hignette G., Buche P., Dibie-Barthélemy J., Haemmerlé O. (2007). An ontology-driven annotation of data tables. In M. Weske, M.-S. Hacid, C. Godart (Eds.), Wise workshops*, vol. 4832, p. 29-40. Springer.*

Hignette G., Buche P., Dibie-Barthélemy J., Haemmerlé O. (2009). Fuzzy Annotation of Web Data Tables Driven by a Domain Ontology. In L. Aroyo et al. (Eds.), *ESWC*, vol. 5554, p. 638-653. Springer.

Limaye G., Sarawagi S., Chakrabarti S. (2010). Annotating and Searching Web Tables Using Entities, Types and Relationships. *PVLDB*, vol. 3, nº 1, p. 1338-1347.

Liu Y., Bai K., Mitra P., Giles C. L. (2007). TableSeer: automatic table metadata extraction and searching in digital libraries. In *Proceedings of the 7th ACM/IEEE-CS joint conference on digital libraries*, p. 91–100. New York, NY, USA, ACM.

McCrae J., Spohr D., Cimiano P. (2011). Linking Lexical Resources and Ontologies on the Semantic Web with Lemon. In G. Antoniou *et al. (Eds.),* ESWC (1)*, vol. 6643, p. 245-259. Springer.*

Noy N., Rector A. (2006). Defining N-ary Relations on the Semantic Web. *Consulté sur http://www.w3.org/TR/swbp-n-aryRelations/*

Noy N. F. (2004). Semantic integration: A survey of ontology-based approaches. SIGMOD Record*, vol. 33, n° 4, p. 65-70.*

Pan J. Z., Stamou G. B., Stoilos G., Taylor S., Thomas E. (2008). Scalable querying services over fuzzy ontologies. In J. Huai et al. (Eds.), *Www*, p. 575-584. ACM.

Reymonet A., Thomas J., Aussenac-Gilles N. (2007). Modelling ontological and terminological resources in OWL DL. In *OntoLex 2007 - workshop at ISWC07.* Busan, South-Korea.

Rijgersberg H., Assem M. van, Top J. L. (2013). Ontology of units of measure and related concepts. *Semantic Web*, vol. 4, n° 1, p. 3-13.

Rijgersberg H., Wigham M., Top J. L. (2011). How semantics can improve engineering processes: A case of units of measure and quantities. *Advanced Engineering Informatics*, vol. 25, n° 2, p. 276-287.

Roche C., Calberg-Challot M., Damas L., Rouard P. (2009). Ontoterminology - a new paradigm for terminology. In J. L. G. Dietz (Ed.), *KEOD*, p. 321-326. INSTICC Press.

Shvaiko P., Euzenat J. (2013). Ontology matching: State of the art and future challenges. *IEEE Trans. Knowl. Data Eng.*, vol. 25, n° 1, p. 158-176.

Staab S., Studer R. (Eds.). (2009). *Handbook on ontologies*. Springer Berlin Heidelberg.

Tenier S., Toussaint Y., Napoli A., Polanco X. (2006). Instantiation of relations for semantic annotation. In *Web intelligence*, p. 463-472. IEEE Computer Society.

Touhami R., Buche P., Dibie-Barthélemy J., Ibanescu L. (2011). An Ontological and Terminological Resource for n-ary Relation Annotation in Web Data Tables. In R. Meersman *et al. (Eds.),* OTM Conferences (2)*, vol. 7045, p. 662-679. Springer.*

Touhami R., Buche P., Dibie-Barthélemy J., Ibanescu L. (2013). Évolution d'une ontologie dédiée à la représentation de relations n-aires. In C. Vrain, A. Péninou, F. Sèdes (Eds.), EGC*, vol. RNTI-E-24, p. 413-418. Hermann-Éditions.*

Venetis P., Halevy A. Y., Madhavan J., Pasca M., Shen W., Wu F. et al. (2011). Recovering semantics of tables on the web. *PVLDB*, vol. 4, n° 9, p. 528-538.