État de l'art : Extraction d'information à partir de thésaurus pour générer une ontologie

Fabien Amarger[*, **] — Catherine Roussey[*] — Jean-Pierre Chanet[*] — Ollivier Haemmerlé[**] — Nathalie Hernandez[**]

[*] *IRSTEA/CEMAGREF - UR TSCF, 24 av. des Landais, BP 50085, Aubière, France - prenom.nom@irstea.fr*
[**] *IRIT, UMR 5505, Université de Toulouse le Mirail, Département de Mathématiques-Informatique, 5 allées Antonio Machado, F-31058 Toulouse Cedex, France - prenom.nom@univ-tlse2.fr*

RÉSUMÉ. *Afin de participer au Web de données pour l'agriculture, nous voulons réutiliser AGROVOC qui est un thésaurus multilingue maintenu par la FAO comportant plus de 40.000 termes. Nous présentons ici un état de l'art des techniques de transformation de thésaurus pour obtenir une ontologie de domaine. Pour cela, nous avons étudié dix approches suivant trois axes : l'extraction de classes, l'extraction de la hiérarchie et l'extraction de relations. Ainsi, nous avons mis en évidence certaines difficultés liées à la transformation de thésaurus comme la désambiguïsation des relations ou la validation des résultats. Nous constatons que les dernières approches mises en œuvre sont fondées sur des techniques manuelles pour répondre en partie à ces difficultés.*

ABSTRACT. *In order to participate to the Linked Data for agriculture, we want to use AGROVOC that is a multilingual thesaurus maintained by FAO with more than 40,000 terms. We present here a state of the art about techniques proposed to transform a domain ontology from thesaurus. For this we will study ten methodologies along three axes: the extraction of classes, the extraction of the hierarchy and other relation extraction. We were able to identify some complex aspects from methodologies such as the complexity of disambiguation or validation that led a return to manual techniques.*

MOTS-CLÉS : *ontologie, extraction d'informations, enrichissement, thesaurus*

KEYWORDS: *ontology, information extraction, enrichment, thesaurus*

1. Introduction

Les données disponibles sur le Web sont généralement de deux natures : (1) des données non structurées difficilement exploitables de manière automatique, comme un ensemble de pages HTML, ou (2) des données structurées destinées à une utilisation particulière, comme une base de données d'un site, difficilement réutilisables par d'autres applications. Le Web de données (Berners-Lee, 2006) est une application du Web sémantique (Berners-Lee *et al.*, 2001) facilitant l'accès, le partage et l'alignement des données. Le W3C[1] a proposé des standards de représentation des données et de leur schémas : les données sont représentées sous forme de triplets RDF[2], tandis que RDFS et OWL[3] définissent les schémas de données associés. Lorsque ces schémas sont suffisamment complexes ils portent le nom d'ontologies. La publication des données et de leurs schémas sur le Web facilitent leur réutilisation dans diverses applications.

Il existe actuellement de très nombreuses données disponibles sur le Web qui pourraient être transformées en ontologies pour enrichir le Web de données. Néanmoins, ces données sont souvent représentées dans des formalismes moins expressifs que les langages de représentation d'ontologies. C'est par exemple le cas des thésaurus, qui sont des réseaux terminologiques spécifiant des relations entre termes. La difficulté réside dans l'extraction et la découverte d'informations dans des sources non ontologiques. Les thésaurus représentent des sources pertinentes pour ce genre de transformations puisqu'ils contiennent un grand nombre d'informations dont la validité a généralement été assurée par de nombreuses années d'utilisation.

Dans cet article, nous proposons d'étudier les différentes méthodes d'extraction d'information à partir de thésaurus, afin de mettre en évidence les tendances actuelles et leurs limites. Dans la section suivante, nous détaillons les motivations de notre travail. Puis, nous définissons les thésaurus et les ontologies OWL. Nous nous intéressons ensuite à la transformation des différentes informations contenues dans un thésaurus : la terminologie, les relations hiérarchiques et les relations d'associations. Nous terminons par une analyse de ces méthodes puis traçons des perspectives de travail.

2. Motivations

Au fil des dernières décennies, les pratiques agricoles ont fortement évolué sous l'effet de diverses contraintes : enjeux sociétaux et environnementaux, cadre règlementaire, changement climatique… Parallèlement, le rôle des données en agriculture a également fortement évolué : d'abord utilisées à des fins de traçabilité et de sécurité alimentaire, les données, qui sont de plus en plus nombreuses, contribuent maintenant directement au changement des pratiques agricoles par leur aide à une meilleure compréhension de celles-ci. La multiplication des équipements embarqués,

1. World Wide Web Consortium (www.w3.org)
2. Ressource Description Framework (www.w3.org/RDF/)
3. OWL Web Ontology Language (www.w3.org/TR/owl-features/)

des smart-phones, des capteurs aux champs, etc., permet à l'heure actuelle de disposer de grands volumes de données spatiotemporelles (Ruiz-Garcia *et al.*, 2009 ; Xie *et al.*, 2008). Le prochain enjeu est de rendre ces données disponibles à l'ensemble des acteurs de la filière afin qu'ils puissent les mobiliser dans des outils d'aide à la décision et d'analyse (Xie *et al.*, 2008 ; Goumopoulos *et al.*, 2009). Le Web de données est une opportunité pour accélérer cette mutualisation des données et, par conséquent, l'évolution des pratiques agricoles.

Afin de pouvoir publier ces données sur le Web de données, il convient de structurer les relations entre les différents concepts appartenant au domaine de l'agriculture : plantes, maladies, ravageurs, pesticides, rotation... Un certain nombre de ressources sont disponibles et mobilisables : taxinomies, thésaurus, bases de données, etc. Par exemple, sont déjà publiées sur le Web de données des données météorologiques issues des capteurs (Corcho et García-Castro, 2010) ou des statistiques agricoles par pays (Eurostat[4]). En revanche, lorsque l'on s'intéresse à un thème particulier comme, par exemple, la protection des cultures, il n'existe pas de ressource.

Nous souhaitons créer une ontologie permettant de décrire les données concernant l'observation des attaques des agresseurs sur les cultures, ainsi que les techniques de traitement des agresseurs. Cette ontologie permettra de publier les données disponibles ; elle permettra également d'annoter les nombreux documents mobilisables pour faire évoluer les pratiques de traitement des agresseurs. L'ensemble de ces données deviendront alors interrogeables par des requêtes exprimées en langage naturel en utilisant le système SWIP (Pradel *et al.*, 2012).

Pour construire notre ontologie, nous voulons utiliser des sources d'informations déjà existantes dans le domaine. Le thésaurus AGROVOC est une référence en agriculture et il est déjà publié sur le Web de données. C'est un thésaurus multilingue de 40.000 termes, développé et maintenu par la FAO[5]. L'existence de ce thésaurus est une des motivations de notre travail visant à transformer des thésaurus en ontologies.

3. Définitions

3.1. *Les thésaurus*

Nous définissons un thésaurus de la manière suivante : un thésaurus est un ensemble de labels (termes) du langage naturel, utilisés pour représenter, de manière sommaire, le sujet des documents. Comme le montre la figure 1, le thésaurus organise ce vocabulaire à l'aide de relations entre concepts. Cette définition est une adaptation de la norme (Isaac et Summers, 2009) en utilisant le vocabulaire SKOS.

SKOS[6] (Simple Knowledge Organisation System) est un standard du W3C permettant de décrire tout système d'organisation des connaissances comme les taxonomies,

4. http ://eurostat.linked-statistics.org/
5. Food and Agriculture Organization of the United Nations
6. http ://www.w3.org/2004/02/skos/

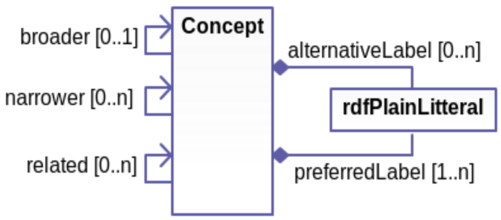

Figure 1. *Structure d'un thésaurus*

les thésaurus, etc. SKOS (Isaac et Summers, 2009) définit le concept comme une unité de pensée (idée, signification ou catégorie d'objets et d'événements)... Les concepts existent en tant qu'entités abstraites indépendamment des termes utilisés pour les identifier. Les labels (ou termes) sont des expressions du langage naturel qui référencent un concept. Dans la figure 1 un label d'un concept est de type "rdfPlainLitteral" c'est à dire. une chaîne de caractères. La figure 2 illustre le contenu d'un thésaurus.

Il existe plusieurs normes pour décrire les thésaurus, la dernière en date étant la norme ISO (25964-1, 2011). Au contraire des précédentes normes, celle-ci suit en partie le format SKOS et formalise explicitement les concepts et leurs liens avec les labels.

Tout thésaurus, quelle que soit la norme utilisée, dispose uniquement de trois relations pour expliciter les liens entre concepts :

– Narrower : relation hiérarchique de spécialisation ;

– Broader : relation hiérarchique de généralisation ;

– Related : relation d'association.

Chaque concept doit avoir un label préféré pour une langue donnée. Il peut également avoir d'autres labels, considérés alors comme des synonymes.

Il est à noter que nous ne souhaitons pas, dans nos travaux, nous limiter uniquement aux thésaurus qui suivent la dernière norme. Les techniques de transformation des thésaurus que nous souhaitons mettre en place ne peuvent donc pas se fonder sur la présence explicite de concepts dans les thésaurus : elles doivent être capable de découvrir les concepts en exploitant les relations entre labels.

Les thésaurus sont avant tout destinés à être utilisés par les documentalistes pour indexer les documents. Cette utilisation manuelle entraîne certaines libertés de conception. Par exemple, dans la figure 2, l'information selon laquelle les basiomycètes sont une maladie du sequoia est indiquée par deux relations "related" partant du concept "Basidiomycota" et allant vers les concepts "Disease" et "Sequoia". Ces formalisations ambiguës seront difficilement identifiables dans un traitement automatique de thésaurus, comme observé dans les travaux de (Soergel *et al.*, 2004).

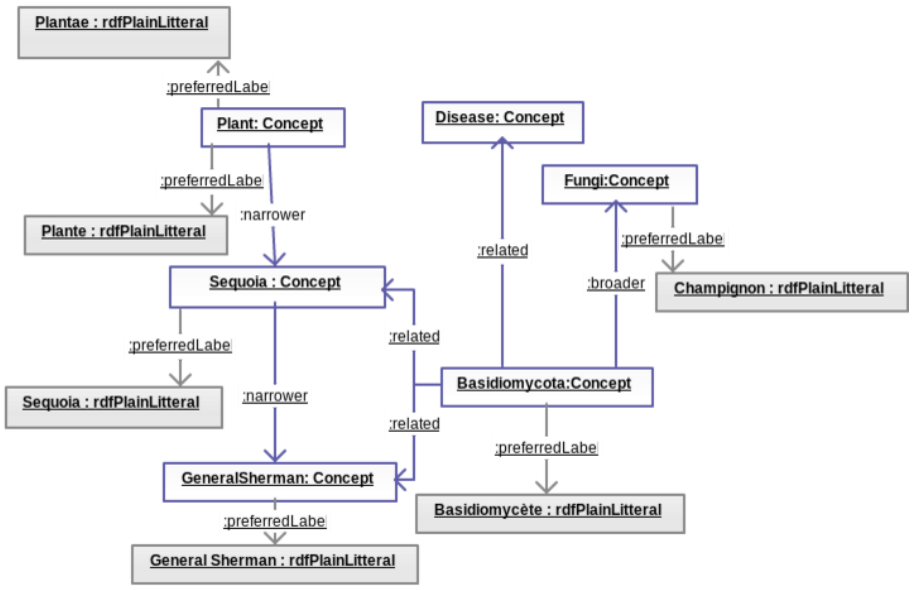

Figure 2. *Exemple du contenu d'un thésaurus sur les séquoias et une de leurs maladies*

3.2. *Les ontologies*

Dans la littérature, la notion d'ontologie est souvent introduite en se fondant sur (Gruber, 1995) : *« A conceptualization is an abstract, simplified view of the world that we wish to represent for some purpose. (...) An ontology is an explicit specification of a conceptualization. »*

La Figure 3 présente une partie d'une ontologie sur les maladies du sequoia. Le sequoia géant de Californie appelé General Sherman est représenté par un individu, instance de la classe "Sequoia". Cette ontologie est définie à partir du profil UML associé au langage OWL défini par l'OMG (OMG, 2005).

Dans nos travaux, nous ne considérons qu'un sous-ensemble des ontologies : les ontologies exprimées en OWL, permettant la production d'inférences de nouvelles données à partir des données déjà présentes dans la base. Ces ontologies contiennent entre autre des axiomes permettant de spécifier des contraintes d'appartenances des individus à une classe. Reprenons l'exemple de la figure 3 et imaginons qu'il n'existe pas de lien "SubClassOf" entre les classes "Basidiomycota" et "Disease". Il serait possible en OWL de redéfinir la classe "Disease" de la figure 3 comme étant l'ensemble des individus ayant une relation "isPestOf" avec un individu de la classe "Plant". Dans ce cas, la classe "Basidiomycota" serait automatiquement identifiée comme une mala-

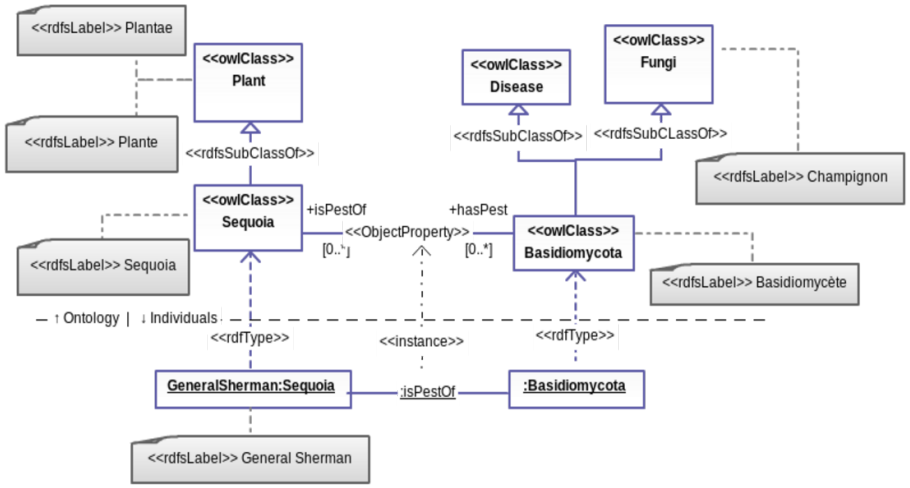

Figure 3. *Exemple d'une ontologie avec des individus représentants le séquoia General Sherman atteint par des basidiomycètes*

die (une classe fille de la classe "Disease"). Il est à noter que le profil UML de l'OMG ne nous permet pas de représenter ce type de contraintes.

Bien que les thésaurus et les ontologies puissent paraître proches par certains aspects, leurs fondements ne sont pas les mêmes : les thésaurus sont des réseaux terminologiques destinés à une utilisation humaine alors que les ontologies OWL sont des conceptualisations permettant de mener des inférences.

4. Travaux existants concernant l'extraction d'informations à partir de thésaurus en vue de générer une ontologie OWL

Nous allons nous attacher, dans la suite, à mettre en évidence un certain nombre de critères pour comparer les différents travaux étudiés :

– *Université/Laboratoire (Uni./Lab.)* indique les différentes universités ou laboratoires impliqués ;

– *Projet* précise le nom du projet ;

– *Usage (Usa.)* présente le but de l'ontologie générée, parmi trois possibilités : Recherche d'Information (RI), Intégration de Données (ID), Non défini (N) ;

– *Domaine* indique le domaine dans lequel est appliquée la méthode ;

– *Thésaurus utilisé(s) (Thés.)* précise le ou les thésaurus sur lesquels la méthode a été appliquée ;

– *Objectif (Obj.)* détermine l'objectif global de la méthode parmi :

- Construction d'Ontologies (CO) : création de classes et de relations à partir de rien,

- Enrichissement d'Ontologies (EO) : ajout de classes et de relations à une ontologie existante,

- Peuplement d'Ontologies (PO) : création d'individus.

Nous pouvons tout d'abord observer dans le Tableau 1 que la plupart des méthodes étudiées (Charlet *et al.*, 2012 ; Chrisment *et al.*, 2008 ; Hepp et De Bruijn, 2007 ; Soergel *et al.*, 2004 ; Van Assem *et al.*, 2004 ; Wielinga *et al.*, 2001) transforment les thésaurus en vue d'une utilisation dans un système de recherche d'information. Seuls (Hahn, 2003 ; Kless *et al.*, 2012 ; Li et Li, 2012) le font dans un but d'intégration de données. Les domaines sur lesquels sont appliquées les méthodes sont divers, ce qui montre l'intérêt d'outils de transformation aussi génériques que possibles. La plupart des méthodes étudiées ont été appliquées sur un seul thésaurus, voire sur une partie du thésaurus pour les méthodes fastidieuses. Certaines méthodes ont été appliquées sur plusieurs thésaurus, mais seul (Charlet *et al.*, 2012) utilise plusieurs sources pour enrichir une seule et même ontologie. La dernière colonne montre que (Charlet *et al.*, 2012) est la seule méthode d'enrichissement alors que les autres sont des méthodes de création d'ontologies. Nous pouvons remarquer que seules certaines méthodes vont jusqu'au peuplement de l'ontologie générée (Charlet *et al.*, 2012 ; Hahn, 2003 ; Li et Li, 2012 ; Soergel *et al.*, 2004 ; Van Assem *et al.*, 2004 ; Villazón-Terrazas *et al.*, 2010).

4.1. *Utilisation de la terminologie*

L'étape de transformation de la terminologie consiste à étudier les labels et les concepts d'un thésaurus pour générer les classes dans une ontologie. Pour cela, deux familles de méthodes se démarquent dans les travaux précédemment cités.

1) La plupart des travaux transforment uniquement les concepts du thésaurus en classes OWL.

2) Certains travaux (Hepp et De Bruijn, 2007 ; Villazón-Terrazas *et al.*, 2010) génèrent une classe de l'ontologie pour chaque label du thésaurus.

La génération des classes est le plus souvent validée manuellement par un expert du domaine.

Concernant nos travaux, la transformation des concepts semble être l'approche la plus intéressante. Néanmoins, une validation manuelle semble peu réaliste vu le nombre important de classes qui seront générées par la transformation d'AGROVOC.

	Uni./Lab.	Projet	Usa.	Domaine	Thés.	Obj.
(Wielinga et al., 2001)	Université d'Amsterdam	Non défini	RI	Art et architecture	AAT	CO
(Hahn, 2003)	Université de Freiburg	Non défini	ID	Médecine	UMLS	CO + PO
(Van Assem et al., 2004)	Université d'Amsterdam	IST-2002-507967 "HOPS", CHIME project	RI	Médecine, générique	MESH (en), WordNet (en)	CO + PO
(Soergel et al., 2004)	Université du Maryland et FAO	Non défini	RI	Agriculture	AGROVOC (multi)	CO + PO
(Hepp et De Bruijn, 2007)	Université d'Innsbruck et Digital Enterprise Research Institute	Non défini	RI	Produits et services	UNPCS (en), eCl@ss (multi)	CO
(Chrisment et al., 2008)	IRIT	Masse de données en astronomie	RI	Astronomie	IAU (multi)	CO
(Villazón-Terrazas et al., 2010)	Université polytechnique de Madrid	NeOn (Suarez-Figueroa *et al.*, 2012)	N	Générique	ASFA (multi)/ ETT (multi)	CO + PO
(Li et Li, 2012)	Université de Qujing	Non défini	ID	Agriculture	MESH	CO + PO
(Kless et al., 2012)	Universités de Melbourne, de Rostock, RWTH Aachen et Information Consultant Lübeck	Non défini	ID	Agriculture	AGROVOC	CO
(Charlet et al., 2012)	INSERM, Hôpitaux de Paris, Armand Trousseau et Compiègne et MONDECA	LERUDI	RI	Urgences médicales	SNOMED (fr), UMLS (multi)	EO + PO

Tableau 1. *Présentations des travaux*

4.2. *Utilisation de la hiérarchie*

Une fois les classes de l'ontologie extraites, il faut traiter les relations hiérarchiques du thésaurus. Les relations narrower et broader témoignent d'une spécialisation ou d'une généralisation thématique et non d'une relation de subsomption entre classes. Cette distinction induit certaines confusions puisque, comme le montre (Soergel *et al.*, 2004), les relations hiérarchiques d'un thésaurus sont parfois utilisées pour représenter des relations de composition (on trouve par exemple dans AGROVOC "Cow, narrower, Cow Milk"[7]). La transformation d'une relation hiérarchique en relation "subClassOf" ne peut pas être systématique.

Pour étudier le traitement de ces relations hiérarchiques, nous proposons différents critères. Le résultat de notre analyse est présenté dans le tableau 2. Les critères liés au traitement de la hiérarchie du thésaurus sont :

– *Sélection (Sél.)* indique si la méthode filtre les branches de la hiérarchie ;

– *Traitement (Trai.)* précise le niveau d'automatisation du traitement (automatique ou manuel) ;

– *Types* énumère les types de relations possibles dans l'ontologie (Hiérarchie "subClassOf" -> HsCO, Hiérarchie Compositionnelle -> HC, Hiérarchie Fonctionnelle -> HF) ;

– *Désambiguïsation* précise si la méthode applique un processus de désambiguïsation pour déterminer la nature de la relation dans l'ontologie et, si oui, quelle technique est utilisée ;

– *Validation (Val.)* indique si la méthode applique un processus de validation du résultat et, si oui, s'il s'agit d'une validation automatique ou manuelle.

Il apparaît dans le tableau 2 que seulement (Charlet *et al.*, 2012 ; Hahn, 2003 ; Kless *et al.*, 2012) effectuent une sélection de la partie du thésaurus à traiter. Les méthodes de traitement de la hiérarchie se décomposent en deux grandes familles :

1) celles qui transforment toutes les relations hiérarchiques narrower/broader en un seul type de relation ontologique ;

2) celles qui transforment les relations hiérarchiques en plusieurs types de relations ontologiques.

Dans la première famille, nous trouvons les travaux de (Chrisment *et al.*, 2008 ; Hepp et De Bruijn, 2007 ; Wielinga *et al.*, 2001 ; Van Assem *et al.*, 2004). Ils proposent un traitement automatique de la hiérarchie en la transformant directement en hiérarchie "subClassOf". (Chrisment *et al.*, 2008) précise que le thésaurus (IAU) est bien organisé par la relation "subClassOf". (Van Assem *et al.*, 2004) propose de choisir la transformation de la hiérarchie et de déterminer si c'est une hiérarchie "subClassOf", compositionnelle ou autre (fonctionnelle). Mais toutes les relations de hiérarchie du thésaurus seront traduites de la même façon. Dans la deuxième famille nous trouvons

7. Vache, plus spécifique, Lait de Vache

	Sél.	Trait.	Types	Désambiguïsation	Val.
(Wielinga et al., 2001)	non	Auto.	HsCO	non	non
(Hahn, 2003)	oui	Auto.	HsCO - HC - HF	non	Auto.
(Van Assem et al., 2004)	non	Auto.	HsCO - HC - HF	non	non
(Soergel et al., 2004)	non	Auto.	HsCO - HC - HF	Patrons de transformations	non
(Hepp et De Bruijn, 2007)	non	Auto.	HsCO	non	non
(Chrisment et al., 2008)	non	Auto.	HsCO	non	Man.
(Villazón-Terrazas et al., 2010)	non	Auto.	HsCO - HC - HF	Ressource externe (non définie)	non
(Li et Li, 2012)	non	Auto.	HsCO - HC - HF	non	non
(Kless et al., 2012)	oui	Man.	HsCO - HC - HF	Man.	Man.
(Charlet et al., 2012)	oui	Man.	HsCO - HC - HF	Man.	Auto.

Tableau 2. *Utilisation de la hiérarchie*

les travaux de (Soergel *et al.*, 2004 ; Villazón-Terrazas *et al.*, 2010 ; Kless *et al.*, 2012 ; Li et Li, 2012 ; Charlet *et al.*, 2012). Ils proposent un traitement automatique de la hiérarchie tout en prenant en compte les éventuelles ambiguïtés qu'elle peut receler. (Soergel *et al.*, 2004) propose une utilisation de patrons de transformations qui permettent la désambiguïsation des relations hiérarchiques du thésaurus, mais aucune autre source n'est utilisée. (Villazón-Terrazas *et al.*, 2010) indique dans sa méthode que toute relation hiérarchique du thésaurus doit être désambiguïsée, sans proposer de ressource particulière. Les travaux les plus récents (Charlet *et al.*, 2012 ; Kless *et al.*, 2012 ; Li et Li, 2012) transforment manuellement la hiérarchie des thésaurus. Il est intéressant de noter que (Hahn, 2003 ; Kless *et al.*, 2012) sont les seuls auteurs qui utilisent des thésaurus ayant plusieurs types de relations hiérarchiques, ce qui facilite leur transformation. Par rapport à la validation de la hiérarchie de l'ontologie, nous pouvons remarquer que (Charlet *et al.*, 2012) applique toujours une validation par patrons structurels et (Chrisment *et al.*, 2008) une validation manuelle. (Hahn, 2003) utilise un moteur d'inférences pour détecter les incohérences à corriger. Tous les autres auteurs ne mentionnent pas de validation particulière.

Nous pouvons déduire de cette étude que l'utilisation des relations hiérarchiques reste une problématique majeure. En effet, la plupart des travaux utilisent une transformation simple (la hiérarchie transformée directement en "subClassOf"). Les seuls travaux mettant en place une méthode complexe sont (Soergel *et al.*, 2004) qui propose une mise en place de patrons de transformations ou (Villazón-Terrazas *et al.*, 2010) qui propose une désambiguïsation par ressources externes. Pour nos travaux, une transformation directe n'est pas envisageable puisque nous souhaitons une représentation correcte du domaine. L'utilisation de patrons de transformations peut permettre de pallier ce problème. Néanmoins, l'écriture de patrons pour la transformation de thésaurus tels que AGROVOC peut se révéler fastidieuse en raison du nombre de patrons nécessaires. L'utilisation de ressources externes pour définir la nature de la

	Util.	Trait.	Désamb.	Val.
(Wielinga et al., 2001)	non	non	non	non
(Hahn, 2003)	oui	Auto.	Man.	non
(Van Assem et al., 2004)	non	non	non	non
(Soergel et al., 2004)	oui	Auto.	Patrons de transformations	non
(Hepp et De Bruijn, 2007)	non	non	non	non
(Chrisment et al., 2008)	oui	Semi-auto.	Corpus	Man.
(Villazón-Terrazas et al., 2010)	oui	Auto.	Ressource externe	non
(Li et Li, 2012)	oui	Man.	Man.	non
(Kless et al., 2012)	oui	Man.	Man.	non
(Charlet et al., 2012)	oui	Man.	Man.	Auto.

Tableau 3. *Utilisation des associations*

relation est ce qui semble le plus viable car le plus automatique. Cette méthode peut pourtant introduire un certain nombre d'erreurs. Il faut donc envisager une validation des désambiguïsations effectuées, comme nous l'évoquons dans les perspectives de ce travail.

4.3. *Utilisation des associations*

Considérons maintenant le traitement des relations d'associations présentes dans un thésaurus. Ces relations permettent de définir une relation entre deux concepts du thésaurus sans en définir la nature exacte. Pour étudier leur traitement, nous allons, là-encore, définir plusieurs critères qui seront ensuite utilisés dans le tableau 3 :

– *Utilisation (Util.)* précise si la méthode prend en compte les relations d'associations (relations "related" dans les thésaurus) ;

– *Traitement (Trait.)* détermine le niveau d'automatisation du traitement des relations d'association (aucun traitement, manuel, semi-automatique, automatique) ;

– *Désambiguïsation (Désamb.)* détermine si ces relations sont désambiguïsées lors de la transformation, et précise la technique utilisée ;

– *Validation (Val.)* précise si la méthode applique un processus de validation (manuel ou automatique) concernant ces relations.

Les relations d'associations sont les plus difficiles à transformer car leur sens n'est pas précisé. Elles peuvent en effet représenter tout type de relations ontologiques. C'est pour cela que, comme nous l'observons dans le tableau 3, certains travaux (Hepp et De Bruijn, 2007 ; Van Assem *et al.*, 2004 ; Wielinga *et al.*, 2001) ne prennent pas en compte ces relations du thésaurus. (Charlet *et al.*, 2012 ; Kless *et al.*, 2012 ; Li et Li, 2012) traitent ces relations de manière entièrement manuelle, ce qui permet

par la même occasion de les désambiguïser. La méthode proposée par (Hahn, 2003) extrait automatiquement les relations ambiguës et un expert les désambiguïse manuellement. Nous retrouvons (Chrisment *et al.*, 2008 ; Soergel *et al.*, 2004 ; Villazón-Terrazas *et al.*, 2010) qui traitent ces relations de manière automatique. (Soergel *et al.*, 2004) utilise le même procédé que pour les relations hiérarchiques, c'est-à-dire une utilisation des patrons de transformations lui permettant d'identifier des relations entre les éléments. (Villazón-Terrazas *et al.*, 2010) utilise, lui aussi, le même procédé que pour les relations hiérarchiques, autrement dit l'utilisation d'une ressource externe pour désambiguïser cette relation (sans préciser la méthode exacte). La méthode qui effectue un traitement intéressant de ces relations est celle présentée par (Chrisment *et al.*, 2008). Elle extrait toutes les relations d'association du thésaurus pour poser des hypothèses de relations candidates entre les classes de l'ontologie. Ces relations sont ensuite désambiguïsées par un traitement automatique d'un corpus de texte du domaine. Si plusieurs relations existent, le choix est laissé à l'utilisateur.

L'objectif que nous nous sommes fixé dans notre travail fait qu'il nous est impossible d'ignorer les relations d'associations d'un thésaurus. En effet, celles-ci représentent (le plus souvent) des relations entre les classes de l'ontologie générée. Malheureusement, leur extraction reste complexe. Pour les mêmes raisons que pour l'extraction de relations hiérarchiques, l'utilisation de patrons de transformations n'est pas raisonnable. Le traitement d'un corpus de textes, comme le préconise (Chrisment *et al.*, 2008), est une méthode très intéressante de par l'utilisation de relations candidates. Néanmoins, il nous faudrait disposer d'un corpus suffisamment important pour qu'il nous permette d'identifier tous les noms de relations entre classes de l'ontologie. Ici encore, l'utilisation de ressources externes est la plus envisageable, à condition que ces ressources nomment explicitement des relations du domaine. WordNet et DBpedia sont des ressources de qualité pour extraire des relations hiérarchiques ou compositionnelles. Nous n'avons pas encore trouvé de ressources de qualité pour extraire des relations du domaine (comme "isPestOf"). Nous retrouvons aussi le problème de la validation des informations extraites, car la désambiguïsation à l'aide de ressources externes n'est pas totalement fiable, ces ressources pouvant également être affectées par la présence d'erreurs.

5. Analyse

Terminologie : Plusieurs aspects ressortent de l'étude de ces méthodes de transformation de thésaurus en ontologies. Tout d'abord, concernant l'identification des classes de l'ontologie à partir des termes du thésaurus, nous avons observé qu'il n'y a que très peu de validation des classes générées. Les seules validations qui existent sont manuelles (Chrisment *et al.*, 2008 ; Charlet *et al.*, 2012). Les autres méthodes étudiées considèrent donc les classes générées comme valides. Le manque de validation peut causer certains problèmes puisque, comme montré dans (Soergel *et al.*, 2004), certaines relations de synonymies ne sont pas valides et peuvent donc apporter des confusions lors du traitement automatique de thésaurus. Dans AGROVOC, par

exemple, nous trouvons que « Hydrophilicity» et «Hydrophobicity » sont des labels du même concept. Une solution pouvant permettre une validation de ces classes serait l'alignement des ontologies générées avec une ontologie déjà existante. Par exemple, s'il y a un alignement avec DBpedia, nous pouvons déduire que les classes alignées sont valides. Sinon, nous pouvons imaginer la suppression des classes qui semblent erronées ou la fusion de classes qui partagent le même concept du thésaurus.

Relations hiérarchiques : Le traitement de la hiérarchie du thésaurus est aussi, dans la plupart des méthodes, effectué de façon naïve. Nous avons vu que les auteurs interprètent généralement la hiérarchie du thésaurus comme une hiérarchie "subClassOf", ce qui n'est pas toujours pertinent (Soergel *et al.*, 2004). Les auteurs prenant en compte cette problématique sont (Soergel *et al.*, 2004 ; Villazón-Terrazas *et al.*, 2010), bien que ce dernier indique uniquement le fait que les relations hiérarchiques du thésaurus doivent être désambiguïsées par une ressource externe, sans préciser la manière de le faire. (Soergel *et al.*, 2004) propose une mise en place de patrons de transformations permettant l'identification de la nature d'une relation du thésaurus, quelle qu'elle soit, relations hiérarchiques comprises. Néanmoins, ce système nécessite d'identifier spécifiquement dans ces patrons toutes les relations du thésaurus qui représentent une relation de "subClassOf" dans l'ontologie. Ceci peut devenir contraignant dans le cas du traitement d'un gros thésaurus tel qu'AGROVOC. (Villazón-Terrazas *et al.*, 2010) propose une désambiguïsation de ces relations en utilisant une ressource externe. En effet, il existe des ressources telles que WordNet ou DBpedia qui définissent plusieurs relations hiérarchiques (« subClassOf », compositionnelle). En utilisant ces ressources, nous pouvons désambiguïser la relation hiérarchique du thésaurus. Néanmoins, ces ressources n'étant pas totalement fiables, il est peu judicieux de s'appuyer entièrement sur leur contenu. (Gangemi *et al.*, 2003) montre d'ailleurs que WordNet ne contient pas une modélisation qui peut être reprise totalement pour en générer une ontologie.

Relations d'associations : Concernant le traitement des relations d'associations, trois possibilités nous semblent envisageables.

1) L'utilisation de patrons de transformations, comme le présente (Soergel *et al.*, 2004), qui nécessite une définition de toutes les relations existantes et demande donc un travail important avant la transformation.

2) La proposition d'hypothèses d'identification des relations candidates à partir de l'analyse d'un corpus de textes, comme le propose (Chrisment *et al.*, 2008). Une limitation réside dans le fait que le traitement du corpus peut lui aussi être erroné, ce qui peut amener à des erreurs dans l'ontologie finale.

3) La désambiguïsation des relations par ressources externes, proposée par (Villazón-Terrazas *et al.*, 2010), est aussi possible, bien que les ressources externes ne définissent généralement pas de relations spécifiques de domaine et peuvent également entraîner certaines erreurs d'interprétation.

Il apparaît dans l'ensemble de ces travaux une nécessité de désambiguïser toutes les relations des thésaurus. Nous pouvons remarquer que les travaux les plus récents (Charlet *et al.*, 2012 ; Kless *et al.*, 2012 ; Li et Li, 2012) s'orientent plutôt vers une transformation manuelle du thésaurus.

Un seul auteur s'est intéressé à la génération des axiomes pour définir les classes (Kless *et al.*, 2012). Il enrichit tout d'abord l'ontologie par des définitions en langage naturel issues des dictionnaires spécialisés. Il en déduit ensuite manuellement la définition formelle des classes. L'extraction d'axiomes pour enrichir l'ontologie représente à elle seule un sujet de recherche à part entière.

6. Perspectives et conclusion

Une solution envisageable pour résoudre les problèmes énumérés ci-dessus pourrait consister à étendre les travaux de (Chrisment *et al.*, 2008) en utilisant le procédé explicité dans (Villazón-Terrazas *et al.*, 2010). Cela consisterait à utiliser le principe de relations candidates extraites du thésaurus, tout en appliquant cette méthode aux relations hiérarchiques et plus aux seules relations d'associations. La désambiguïsation de ces relations ne se ferait plus par utilisation d'un corpus de texte mais par l'utilisation de ressources externes, comme le préconise (Villazón-Terrazas *et al.*, 2010). La validation des informations extraites pourrait être en partie résolue en générant une ontologie "floue", associant un score de pertinence à ses composants. Les relations et classes candidates extraites du thésaurus initial pourraient être désambiguïsées grâce à l'utilisation de plusieurs ressources externes (et non plus une seule comme présenté dans les travaux de (Villazón-Terrazas *et al.*, 2010)). Cette pluralité permettrait d'augmenter le score de pertinence des éléments de notre ontologie floue dans le cas où les éléments seraient présents dans plusieurs ressources. Il serait aussi intéressant de pouvoir recouper les informations provenant de plusieurs sources[8] (un autre thésaurus ou une source d'un autre type) en augmentant ce score de pertinence si l'information est présente dans une autre source. Nous pourrons par exemple croiser les informations d'AGROVOC avec celles présentes dans e-phy[9]. Les informations extraites de ces différentes sources seraient par conséquent plus fiables, car validées plusieurs fois, même si cela ne remplacera jamais la validation d'un expert du domaine. Il existe aussi des méthodes de validation communautaire telles que (Lafourcade et Joubert, 2008) qui propose une validation par l'intermédiaire d'un jeu qui demande d'identifier des relations ou encore (Jupp *et al.*, 2010) qui propose un outil pour faciliter le travail de l'expert pour la validation de patrons de transformations. Ce système serait difficilement applicable dans notre cas puisque les utilisateurs capables de valider sémantiquement une information sont des agronomes qui ne sont pas suffisamment disponibles pour

8. Nous faisons la distinction entre une "source" qui est l'entrée de notre processus et que nous souhaitons transformer en ontologie, et une "ressource" qui est une donnée externe aidant à la transformation de la (ou des) source(s)

9. une base de données de produits phytosanitaires contenant un grand nombre d'informations concernant le domaine agricole

que cette validation soit efficace. Le problème de la validation reste donc une question ouverte.

Dans ce document nous avons présenté un état de l'art des méthodes d'extraction d'informations à partir de thésaurus dans un but de génération d'ontologie. Nous avons pu observer que ce travail est plus complexe qu'il n'y paraît, ce qui explique que la tendance actuelle s'oriente plutôt vers des méthodes manuelles, le problème principal étant la désambiguïsation des relations du thésaurus. Les méthodes automatiques de désambiguïsation utilisent une ressource externe, ou des patrons de transformations ou encore une analyse d'un corpus. La méthode que nous allons développer devra reposer sur plusieurs sources à transformer, ainsi que sur des ressources externes de désambiguïsation, afin de générer une ontologie pondérée par des scores de vraisemblance. Une telle ontologie sera particulièrement bien adaptée à un mécanisme d'interrogation intuitif tel que le système SWIP (Pradel *et al.*, 2012).

7. Bibliographie

25964-1 I., « Information and documentation – Thesauri and interoperability with other vocabularies – Part 1 : Thesauri for information retrieval », *Thesauri and interoperability with other vocabularies. Part*, vol. 1, 2011.

Berners-Lee T., Hendler J., Lassila O., « The semantic web », *Scientific american*, vol. 284, n° 5, 2001, page 28–37.

Berners-Lee T., « Linked Data - Design Issues », http ://www.w3.org/DesignIssues/LinkedData.html, 2006.

Charlet J., Declerck G., Dhombres F., Gayet P., Miroux P., Vandenbussche P., others, « Construire une ontologie médicale pour la recherche d'information : problématiques terminologiques et de modélisation », *Actes des 23es journées francophones d'Ingénierie des connaissances*, vol. 1, 2012, page 33–48.

Chrisment C., Haemmerlé O., Hernandez N., Mothe J., « Méthodologie de transformation d'un thesaurus en une ontologie de domaine », *Revue d'Intelligence Artificielle*, vol. 22, n° 1, 2008, p. 7–37.

Corcho O., García-Castro R., « Five challenges for the semantic sensor web », *Semantic Web*, vol. 1, n° 1, 2010, page 121–125.

Gangemi A., Navigli R., Velardi P., « The OntoWordNet Project : extension and axiomatization of conceptual relations in WordNet », *On The Move to Meaningful Internet Systems 2003 : CoopIS, DOA, and ODBASE*, vol. 1, 2003, page 820–838.

Goumopoulos C., Kameas A. D., Cassells A., « An ontology-driven system architecture for precision agriculture applications », *International Journal of Metadata, Semantics and Ontologies*, vol. 4, n° 1, 2009, page 72–84.

Gruber T., « Toward principles for the design of ontologies used for knowledge sharing », *International journal of human computer studies*, vol. 43, n° 5, 1995, page 907–928.

Hahn U., « Turning informal thesauri into formal ontologies : a feasibility study on biomedical knowledge re-use », *Comparative and functional genomics*, vol. 4, n° 1, 2003, page 94–97.

Hepp M., De Bruijn J., « GenTax : A generic methodology for deriving OWL and RDF-S on-

tologies from hierarchical classifications, thesauri, and inconsistent taxonomies », *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 4519 LNCS, Innsbruck, 2007, p. 129–144, cited By (since 1996) 5 ; Conference of 4th European Semantic Web Conference, ESWC 2007 ; Conference Date : 3 June 2007 through 7 June 2007 ; Conference Code : 70113.

Isaac A., Summers E., « SKOS Simple Knowledge Organization System Primer. W3C Working Group Note », *World Wide Web Consortium*, vol. 1, 2009.

Jupp S., Horridge M., Iannone L., Klein J., Owen S., Schanstra J., Stevens R., Wolstencroft K., « Populous : A tool for populating ontology templates », *arXiv preprint arXiv :1012.1745*, vol. 1, 2010.

Kless D., Jansen L., Lindenthal J., Wiebensohn J., « A method for re-engineering a thesaurus into an ontology », Ios PressInc, 2012, page 133.

Lafourcade M., Joubert A., « JeuxDeMots : un prototype ludique pour l'émergence de relations entre termes », *JADT'08 : Journées internationales d'Analyse statistiques des Données Textuelles*, 2008, page 657–666.

Li P., Li Y., « On Transformation from The Thesaurus into Domain Ontology », vol. 1, 2012.

OMG, « Ontology Definition Metamodel », , 2005.

Pradel C., Haemmerlé O., Hernandez N., « A semantic web interface using patterns : the SWIP system », *Graph Structures for Knowledge Representation and Reasoning*, vol. 1, 2012, page 172–187.

Ruiz-Garcia L., Lunadei L., Barreiro P., Robla I., « A review of wireless sensor technologies and applications in agriculture and food industry : state of the art and current trends », *Sensors*, vol. 9, n° 6, 2009, page 4728–4750.

Soergel D., Lauser B., Liang A., Fisseha F., Keizer J., Katz S., « Reengineering thesauri for new applications : The AGROVOC example », *Journal of Digital Information*, vol. 4, n° 4, 2004, cited By (since 1996) 51.

Suarez-Figueroa M.-C., Gomez-Perez A., Motta E., Gangemi A., *Ontology Engineering in a Networked World*, 2012.

Van Assem M., Menken M., Schreiber G., Wielemaker J., Wielinga B., « A method for converting thesauri to RDF/OWL », *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 3298, 2004, p. p17–31, cited By (since 1996) 10.

Villazón-Terrazas B. C., Suárez-Figueroa M., Gómez-Pérez A., « A Pattern-Based Method for Re-Engineering Non-Ontological Resources into Ontologies », *International Journal on Semantic Web and Information Systems*, vol. 6, n° 4, 2010, p. 27–63.

Wielinga B. J., Schreiber A. T., Wielemaker J., Sandberg J. A. C., « From thesaurus to ontology », Victoria, British Columbia, Canada, 2001, ACM, p. 194–201.

Xie N., Wang W., Yang Y., « Ontology-based agricultural knowledge acquisition and application », *Computer And Computing Technologies In Agriculture, Volume I*, vol. 1, 2008, page 349–357.