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

F. Amarger[1,2] **C. Roussey**[2] **N. Hernandez**[1] **J.P. Chanet**[2]
O. Haemmerlé[1]

[1] *IRIT*
UMR 5505, UTM, Département de Mathématiques-Informatique, 5 allées Antonio Machado, F-31058 Toulouse Cedex - prenom.nom@univ-tlse2.fr

[2] *IRSTEA*
Équipe COPAIN, 24 Av. des Landais CS 200 85, 63178, Aubiére, France - prenom.nom@irstea.fr

Pour mieux affirmer ses missions, le Cemagref devient Irstea

www.irstea.fr

29 mai 2013

Plan

I. Motivations

- Volonté d'aider les agriculteurs à diminuer l'usage des produits phytosanitaires
- Augmentation du nombre de données dans le domaine de l'agriculture
  - Bulletins de Santé du Végétal
  - thésaurus AGROVOC
  - base de données publique e-phy
  - etc ...
- Volonté de contribuer au Linked Open Data (LOD)
- Interrogation de l'ontologie par requête en langage naturel (projet SWIP)

II. Définitions
1. Formelles

Thésaurus ISO 25964-1 :2011

Un thésaurus est un ensemble de labels (termes) du langage naturel, utilisés pour représenter, de manière sommaire, le sujet des documents.

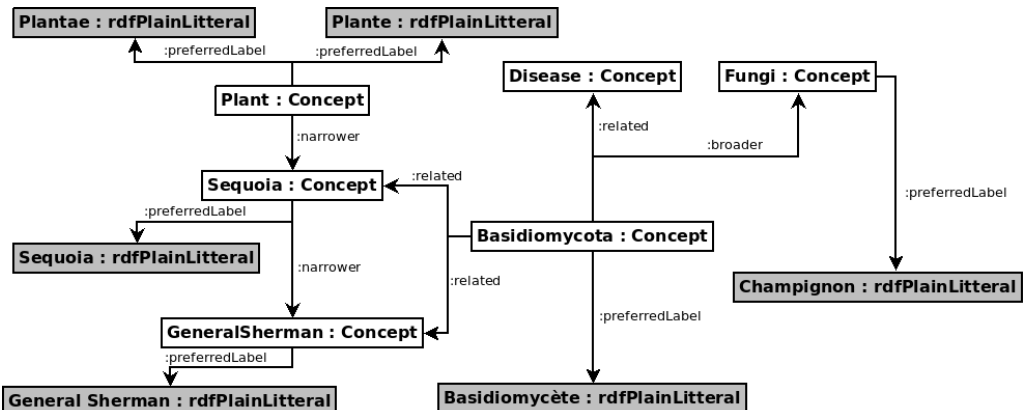
Ontologie traduction de (Gruber et al. 1995)

Une conceptualisation est un résumé, une vue simplifiée du monde que nous souhaitons représenter. (...) Une ontologie est une formalisation explicite d'une conceptualisation.

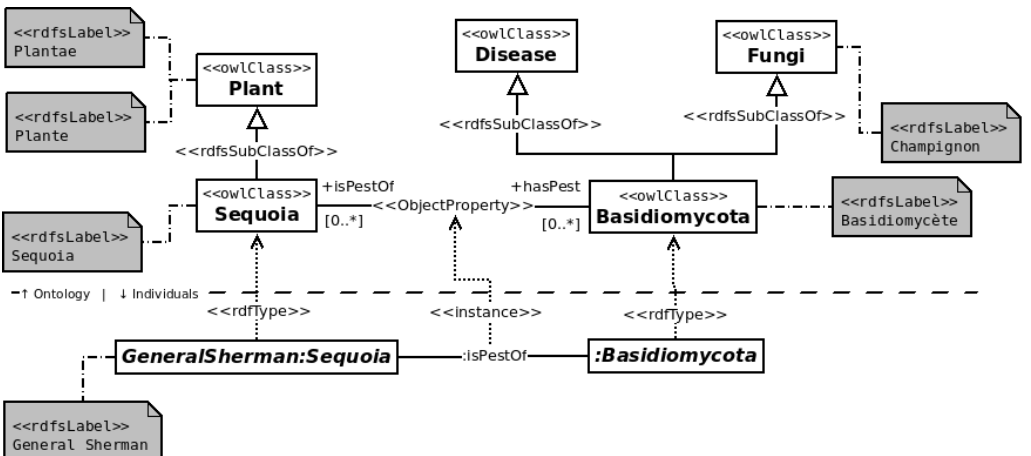
Plus spécifiquement

Formalisation en OWL, ontologie avec inférences valides

II. Définitions
2. Exemples
Thésaurus



II. Définitions
2. Exemples
Ontologie



29 mai 2013

III. Etat de l'art
1. Travaux étudiés

Rérérence	Domaine	Objectif	Usage
Charlet et al. 2012	Urgences médicales	EO+PO	RI
Kless et al. 2012	Agriculture	CO	ID
Li et al. 2012	Agriculture	CO+PO	ID
Villazon-Terrazas et al. 2010	Générique	CO + PO	N
Chrisment et al. 2008	Astronomie	CO	RI
Hepp et al. 2007	Produits et services	CO	RI
Soergel et al. 2004	Agriculture	CO + PO	RI
Van Assem et al. 2004	Médecine	CO+PO	RI
Hahn et al. 2003	Médecine	CO+PO	ID
Wielinga et al. 2001	Art et architecture	CO	RI

Légende

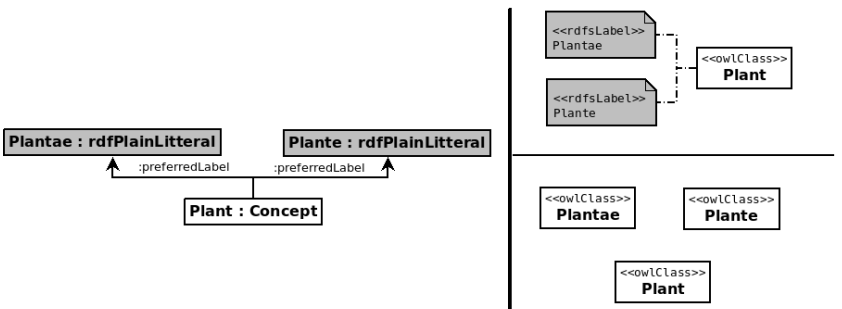
EO = Enrichissement d'ontologie
CO = Construction d'ontologie
PO = Peuplement d'ontologie
RI = Recherche d'information
ID = Intégration de données
N = non renseigné

III. Etat de l'art

2. Utilisation de la terminologie

Analyse

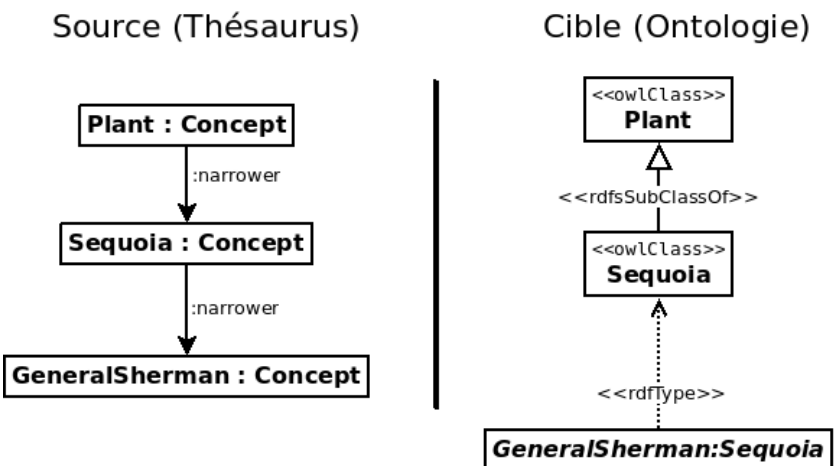
- Deux familles de méthodes :
  - un concept du thésaurus = une classe OWL
  - un label du thésaurus = une classe OWL
- Validation manuelle



III. Etat de l'art

3. Utilisation de la hiérarchie

Exemple



III. Etat de l'art

3. Utilisation de la hiérarchie

Critères

- **Sélection** : Filtrage des branches de la hiérarchie à traiter
- **Traitement** : Méthode d'extraction automatique ou manuelle
- **Types** : Liste les types de relations possibles lors de l'extraction :
  - fonctionnelle
  - hiérarchique (subClassOf)
  - compositionnelle (partOf)
- **Désambiguïsation** : Technique de désambiguïsation des relations hiérarchiques
- **Validation** : Une méthode de validation est elle appliquée (Automatique ou manuelle)

III. Etat de l'art

3. Utilisation de la hiérarchie

Analyse

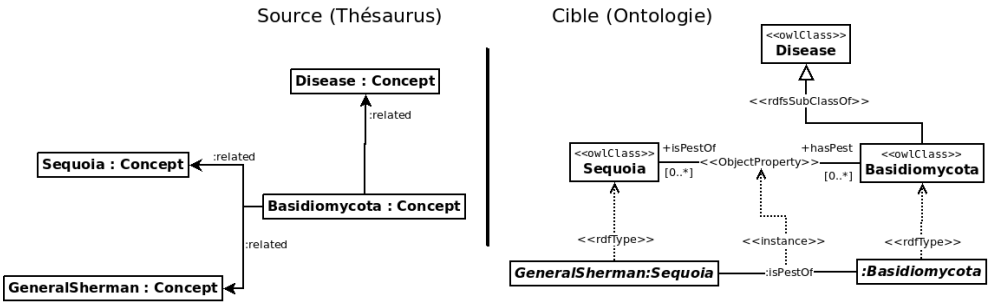
Rérérence	Sél.	Trait.	Désamb.	Val.
Charlet et al. 2012	oui	Man.	Man.	Auto.
Kless et al. 2012	oui	Man.	Man.	Man.
Li et al. 2012	non	Auto.	non	non
Villazon-Terrazas et al. 2010	non	Auto.	Ressources	non
Chrisment et al. 2008	non	Auto.	non	Man.
Hepp et al. 2007	non	Auto.	non	non
Soergel et al. 2004	non	Auto.	Patrons	non
Van Assem et al. 2004	non	Auto.	non	non
Hahn et al. 2003	oui	Auto.	non	Auto.
Wielinga et al. 2001	non	Auto.	non	non

- Peu de sélection
- Tendance actuelle : extraction manuelle
- Deux méthodes de désambiguïsation automatique
- Peu de validation

III. Etat de l'art

4. Utilisation des associations

Exemple



- **Utilisation** : Prise en compte des relations "related"
- **Traitement** : Méthode d'extraction automatique ou manuelle
- **Désambiguïsation** : Technique de désambiguïsation des relations d'association
- **Validation** : Une méthode de validation est elle appliquée (Automatique ou manuelle)

III. Etat de l'art

4. Utilisation des associations

Analyse

Rérérence	Util.	Trait.	Désamb.	Val.
Charlet et al. 2012	oui	Man.	Man.	Auto.
Kless et al. 2012	oui	Man.	Man.	non
Li et al. 2012	oui	Man.	Man.	non
Villazon-Terrazas et al. 2010	oui	Auto.	Ressources	non
Chrisment et al. 2008	oui	Semi-Auto.	Corpus	Man.
Hepp et al. 2007	non	non	non	non
Soergel et al. 2004	oui	Auto.	Patrons	non
Van Assem et al. 2004	non	non	non	non
Hahn et al. 2003	oui	Auto.	Man.	non
Wielinga et al. 2001	non	non	non	non

- Certains auteurs ne les traitent pas du tout
- Tendance actuelle : extraction manuelle
- Désambiguïsation automatique

IV. Analyse générale

- Peu de validation
  - Validation de la conceptualisation manuelle (Charlet et al. 2012, Chrisment et al. 2008, Kless et al. 2012)
  - Validation structurelle automatique (Charlet et al. 2013, Hahn et al. 2003)
- Désambiguïsation des relations (hiérarchiques et d'associations)
  - Désambiguïsation naïve
  - Désambiguïsation par patrons (Soergel et al. 2004)
  - Désambiguïsation par utilisation d'une ressource externe (Villazon-Terrazas et al. 2011)
  - Désambiguïsation par traitement d'un corpus (Chrisment et al. 2008)
- Tendance actuelle
  - Traitement manuel (Charlet et al. 2012, Kless et al. 2012, Li et al. 2012)
  - Extraction d'axiomes par Traitement du Langage Naturel de définitions (Kless et al. 2012)

V. Perspectives

- Validation des classes générées grâce à un alignement sur diverses ressources (DBPedia, Yago, etc ...)
- Désambiguïsation de toutes les relations (related, broader, narrower) en utilisant plusieurs ressources externes
- Confrontation d'informations : extraction parallèle sur plusieurs sources
  - Valeur de confiance d'une information

VI. Remerciement

Merci pour votre attention.
Avez-vous des questions ?