



Extraction de connaissances à partir de sources non ontologiques

**Fabien Amarger^{1,2}, Jean-Pierre Chanet², Ollivier Haemmerlé¹,
Nathalie Hernandez¹, Catherine Roussey²**

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, 9 av. Blaise Pascal CS 200 85, 63178, Aubière, France - prenom.nom@irstea.fr

28 novembre 2013

Pour mieux
affirmer
ses missions,
le Cemagref
devient Irstea



www.irstea.fr





Plan

I. Contexte

II. Proposition

- Transformation syntaxique
- Analyseur de qualité
- Extraction ciblée sur le domaine
- Alignements et fusion des éléments filtrés
- Validations

III. Conclusion et perspectives



28 novembre 2013



I. Contexte

Motivations

- Volonté d'aider les agriculteurs à diminuer l'usage des produits phytosanitaires
- Augmentation des sources d'information dans le domaine de l'agriculture
 - Bulletins de Santé du Végétal (BSV)
 - Thésaurus AGROVOC
 - Base de données publique e-phy
 - etc.
- Mais les sources d'informations sont hétérogènes et distribuées



- Cibler la connaissance d'intérêt pour le cas d'étude
- Agréger les connaissances disponibles



I. Contexte

Motivations

- Faciliter l'accès des agriculteurs à ces connaissances
- Permettre l'interrogation de la base de connaissances en LN (SWIP)
- Rendre les connaissances extraites réutilisables dans d'autres contextes
- Volonté de contribuer au Linked Open Data (LOD)





I. Contexte

Idee générale - problématique

Problématique générale

Création d'une base de connaissances (BC) à partir de plusieurs sources non ontologiques

- Réutilisation de sources existantes
- Exploitation des avantages des différents types de formalismes associés aux sources d'information (BDD, Thésaurus)



28 novembre 2013



I. Contexte

Idée générale - Proposition

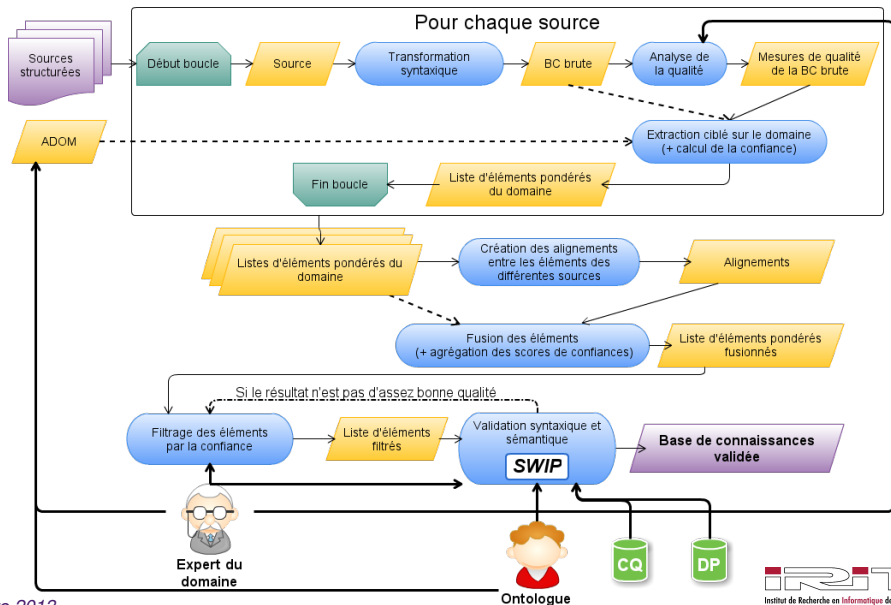
Proposition

- Évaluer la confiance à accorder à une connaissance extraite à partir de :
 - La qualité de la source de la connaissance
 - Le type de source dont elle est extraite
 - Son degré de consensus



28 novembre 2013

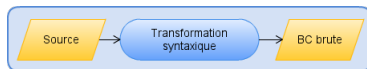
II. Proposition



28 novembre 2013

II. Proposition

1. Transformation syntaxique



Base de connaissances(BC) brute

Une BC brute est une BC obtenue par une méthode automatique

- Dans notre cas, toujours à partir d'une source non ontologique
- Utilisation du projet NOR2O (Villazon-Terrazas 2011)
 - Fait partie du projet NeOn
 - Implémentation sous forme de patrons de transformations dédiés au type de source (Thésaurus, taxonomie, ...)
 - Application disponible et open source

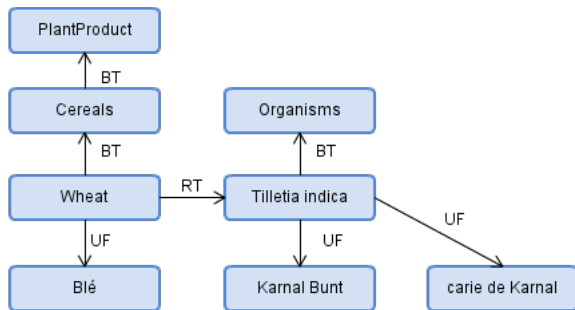


28 novembre 2013

II. Proposition

1. Transformation syntaxique

Exemple d'utilisation de NOR20



AGROVOC

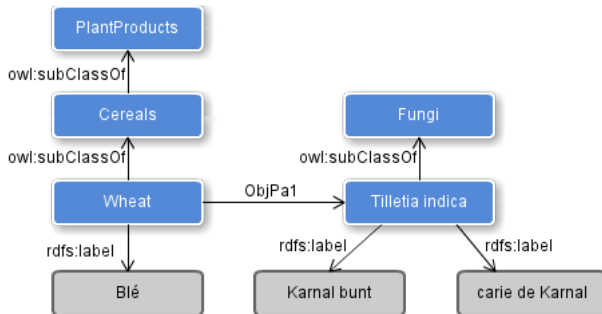


28 novembre 2013

II. Proposition

1. Transformation syntaxique

Exemple d'utilisation de NOR2O



AGROVOC

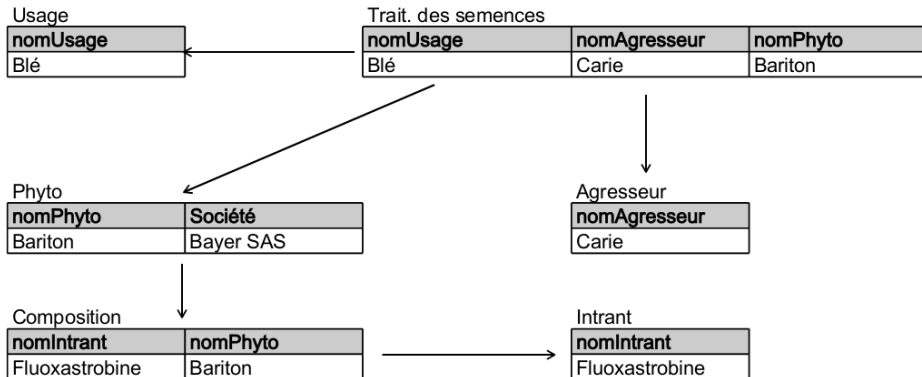


28 novembre 2013

II. Proposition

1. Transformation syntaxique

Exemple d'utilisation de NOR20

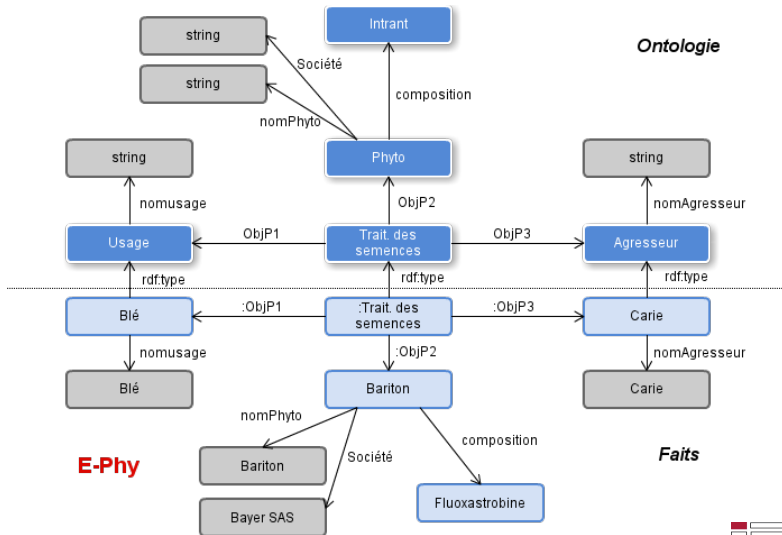


E-Phy

II. Proposition

1. Transformation syntaxique

Exemple d'utilisation de NOR20

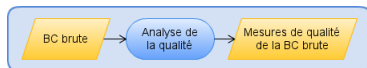


E-Phy

Facts

II. Proposition

2. Analyseur de qualité



En réflexion ...

- Définir les critères de qualité d'une source
- Prise en compte du point de vue de l'utilisateur
- Méthode de représentation de la qualité d'une source





II. Proposition

2. Analyseur de qualité *Définitions - Qualité*

traduction de ISO 9000 - Quality management systems - 2005

Ensemble de caractéristiques d'un produit, d'un système ou d'un processus qui lui confèrent l'aptitude à satisfaire des besoins exprimés ou implicites.

- La définition de la qualité, pour nous, est :
 - Dépendante de l'application
 - Évolutive





II. Proposition

2. Analyseur de qualité

Critères de qualité

Citations

[String et al. 1997][Jayawardene et al. 2013][Gómez-Pérez et al. 1999][Poveda-Villalón et al. 2012][Gil et al. 2007][Suominen et al. 2012][Mader et al. 2011]

Organisation des critères

- Conception de la source
- Utilisation de la source
- Contenu de la source





II. Proposition

2. Analyseur de qualité

Critères sur la conception de la source

- Origine
 - La provenance des données est indiquée et sûre
 - L'expertise des auteurs est reconnue
 - L'intérêt des auteurs pour la qualité des données
- Accessibilité
 - Données ouvertes
 - Format simple, flexible et ouvert
 - Qualité de la présentation / Qualité de la documentation





II. Proposition

2. Analyseur de qualité

Critères sur l'utilisation de la source

- Popularité
 - Les données sont beaucoup utilisées
 - Les données sont liées
 - La source fait référence dans le domaine
 - La source a une bonne réputation





II. Proposition

2. Analyseur de qualité

Critères sur le contenu de la source

- Fraicheur
 - Dernière MAJ
 - Fréquence de MAJ
 - Nécessité de MAJ
- Validité
 - Valide syntaxiquement
 - Respect les contraintes du domaine



28 novembre 2013



II. Proposition

2. Analyseur de qualité

Critères sur le contenu de la source

- Précision
 - La granularité est conforme aux besoins
 - La source contient le moins possible d'ambiguïtés
- Couverture
 - Les données couvrent le domaine étudié

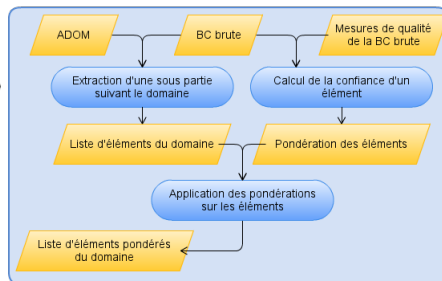
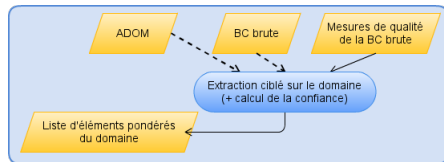


28 novembre 2013

Amarger F., Chanet J.P., Haemmerlé O., Hernandez N., Roussey C.

II. Proposition

3. Extraction ciblée sur le domaine



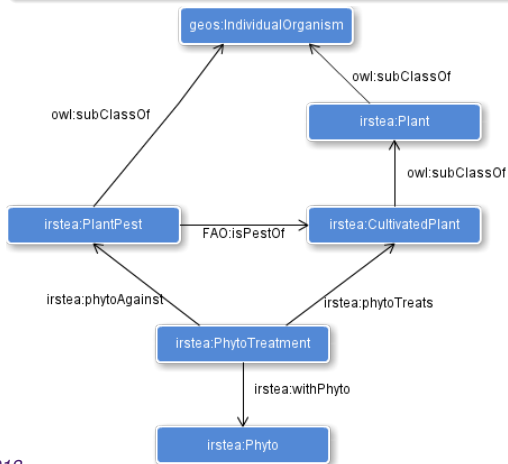
II. Proposition

3. Extraction ciblée sur le domaine

Exemple ADOM

ADOM

Abstract Domain Ontological Module

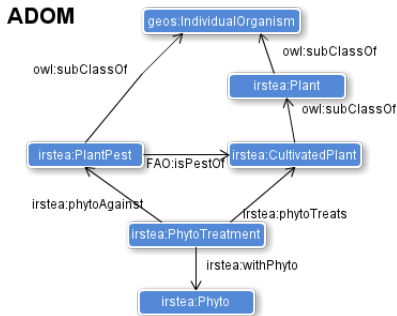


28 novembre 2013

II. Proposition

3. Extraction ciblée sur le domaine

Exemple



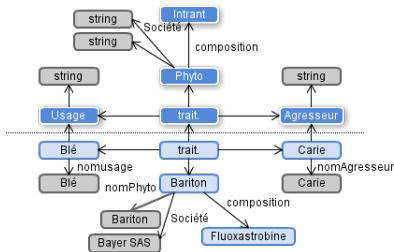
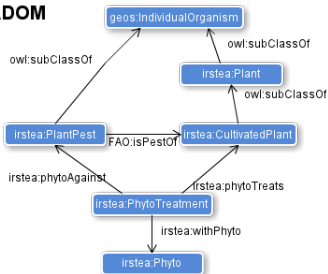
28 novembre 2013

II. Proposition

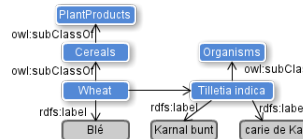
3. Extraction ciblée sur le domaine

Exemple

ADOM



BC brute e-phy



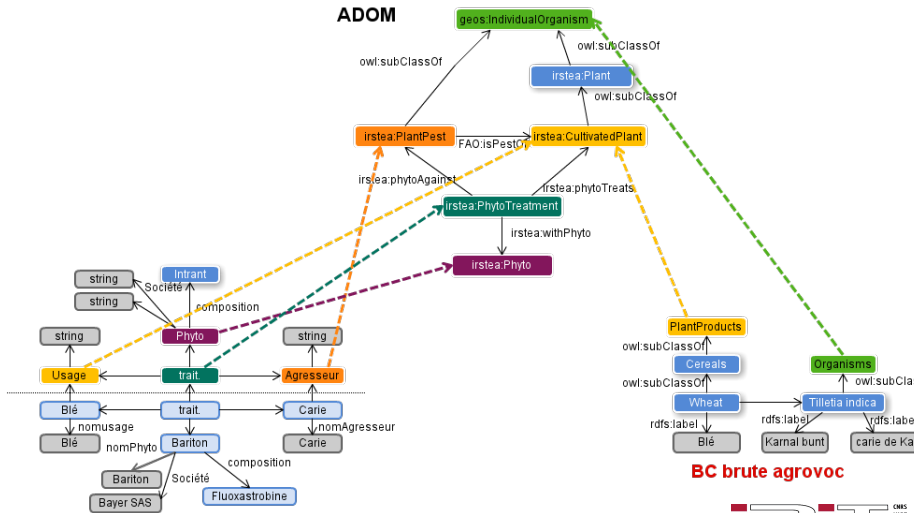
BC brute agrovoc



II. Proposition

3. Extraction ciblée sur le domaine

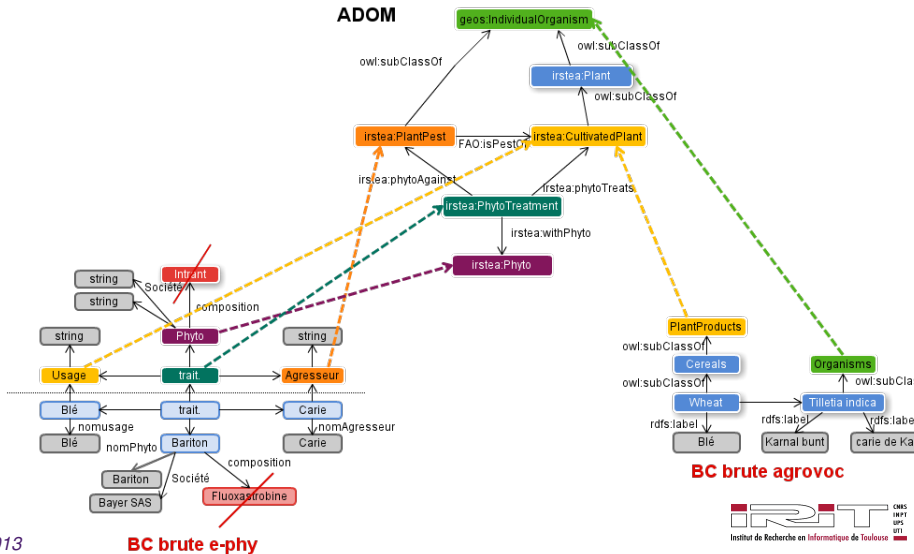
Exemple - Alignement manuels des sources avec le ADOM



II. Proposition

3. Extraction ciblée sur le domaine

Exemple - Filtrage des sous parties





II. Proposition

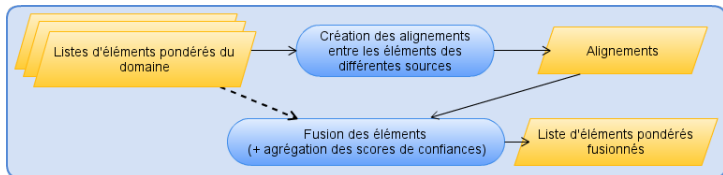
3. Extraction ciblée sur le domaine *Calcul des scores de confiance*

- Calcul des scores de confiances à partir de la qualité de la source d'information
- Exemple :
 - Augmentation du score de base pour les classes provenant d'un thésaurus
 - Diminution du score des relations ObjectProperties provenant d'un thésaurus
 - Augmentation du score des relations ObjectProperties provenant d'une base de données
 - ...



II. Proposition

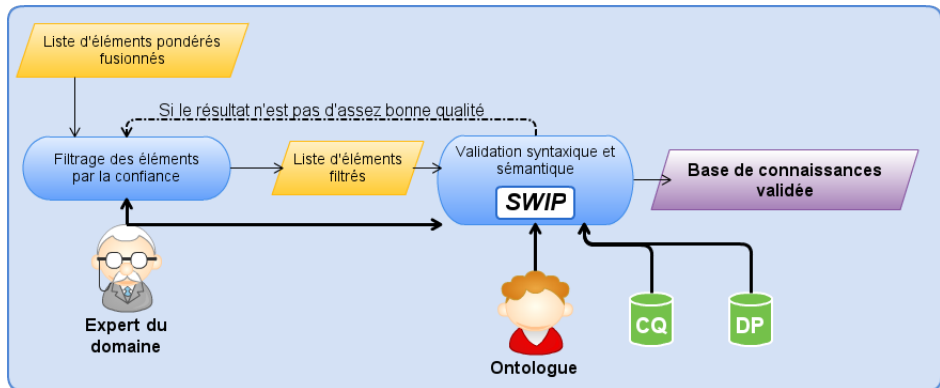
4. Alignements et fusion des éléments filtrés



- Création des alignements d'équivalence entre un maximum d'éléments des éléments filtrés
 - Alignement des relations, classes, dataproperties et instances
 - Méthode encore à déterminer
- Utilisation de ces alignements pour fusionner les listes d'éléments
 - Agrégation des scores de confiances

II. Proposition

5. Validations





III. Conclusion et perspectives

- Proposition d'une méthode d'extraction de connaissances à partir de sources non ontologiques
- En cours de réflexion :
 - Critères de qualités d'une source
 - Calcul de la confiance d'un élément à partir de la qualité de la source
 - Méthode de représentation de la confiance
 - Méthode d'agrégation de la confiance lors de la fusion
- Perspectives :
 - Méthode d'extraction de sous parties suivant le domaine (ADOM)
 - Méthode d'alignement et de fusion
 - Méthodes de validation

