

Apport du Web sémantique à la recherche d'information

Pierre Molette
www.acetic.fr

I-Expo - mai 2008

Le Web sémantique, est-ce une bonne appellation ?

- Tim Berners-Lee répond ceci :
 - « Le terme Web sémantique prête un peu à confusion car la sémantique s'intéresse au sens du langage pour en déduire des constructions logiques.
 - Du coup, certains ont pensé qu'il s'agissait d'un Web qui permettrait par exemple d'effectuer des recherches sur Internet en posant des questions en langage naturel.
 - Or ce n'est pas son but. En fait, nous aurions dû l'appeler dès le départ "Web de données". Mais il est trop tard pour changer... »
-
-

Le Web sémantique propose des normes reconnues et utilisées

- Ces normes visent à rendre le Web accessible et réutilisable en utilisant des meta-données. Par exemple :
 - RDF : modèle conceptuel permettant de décrire des données ;
 - RDF Schema : langage permettant de créer des vocabulaires, ensembles de termes utilisés pour décrire des choses ;
 - OWL : langage permettant de créer des ontologies, vocabulaires plus complexes servant de support aux traitements logiques (inférences, classification, etc).
 - RDF et OWL sont des vocabulaires XML qui font déjà l'objet de nombreuses applications.
-
-

Le W3C prévoit dans le Web 3.0 des évolutions majeures d'Internet

- Certains projets du W3C prévoient de transformer Internet en une gigantesque ontologie, visible comme une base de données structurée par concepts
 - Le contenu du Web devra tenir compte du langage naturel, généraliser l'affichage en 3D et exploiter l'Intelligence Artificielle
 - Le Web 3.0 étant à l'état de projet, il faut prendre ces annonces avec précaution et attendre que cette technologie soit adoptée
-
-

OWL est une norme permettant de construire des ontologies

- OWL, qui signifie *Web Ontology Language*, permet de construire des classifications de concepts structurées et arborescentes
- OWL propose aussi des règles permettant de garantir la cohérence entre les concepts
- Les thesaurus, taxinomies (ou d'autres classifications) peuvent être exportés au format OWL, mais le contrôle de cohérence demande un travail supplémentaire

Quelques ontologies disponibles dans les Sciences naturelles

- AGROVOC / FAO (Food and Agriculture Organization, Nations Unies) ;
 - ITIS / Department of Agriculture (USDA, USA) ;
 - MeSH / National Library of Medicine (NIH, USA) ;
 - NCI Thesaurus / National Cancer Institute (NCI, USA) ;
 - Wikipedia / Wikipedia Foundation (USA) ;
 - Wordnet / Princeton University.
-
-

AGROVOC : thesaurus et ontologie

- Conçu pour standardiser l'indexation des documents relatifs à l'agriculture
 - Diffusé par la FAO (instance de l'ONU pour l'alimentation et l'agriculture) en collaboration avec certains pays membres
 - Disponible en 5 langues : anglais, français, espagnol, chinois et arabe
 - Distribué sous différents formats : MySQL, TagText, ISO2709, Microsoft Access et XML
-
-

Agrovoc est disponible en version OWL multilingue

```
<owl:Class rdf:ID="c_13251">
  <rdfs:label xml:lang="cs">peroxidázy</rdfs:label>
  <rdfs:subClassOf>
    <owl:Class rdf:ID="c_5474"/>
  </rdfs:subClassOf>
  <rdfs:label xml:lang="en">Peroxidases</rdfs:label>
  <rdfs:label xml:lang="pt">Peroxidase</rdfs:label>
  <rdfs:label xml:lang="es">Peroxidasas</rdfs:label>
  <rdfs:label xml:lang="fr">Péroxydase</rdfs:label>
  <rdfs:label xml:lang="zh">过氧化物酶</rdfs:label>
  <rdfs:label xml:lang="ar">بيروكسيداز</rdfs:label>
</owl:Class>
<owl:Class rdf:ID="c_31416">
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:someValuesFrom>
```

- Chaque descripteur fait l'objet d'un label dans une langue différente
- D'autres langues sont en cours de développement (allemand, italien, japonais, coréen, ...)
- Ce n'est pas une vraie ressource terminologique normalisée

Le projet FAO / CIRAD / ACETIC

- Transformer l'ontologie AGROVOC en ressource linguistique
 - Intégrer cette ressource dans un moteur de recherche sémantique
 - Classer automatiquement des documents scientifiques avec cette ressource linguistique
 - Tester et évaluer le résultat obtenu
-
-

Transformer une ontologie en ressource sémantique

- Constituer un dictionnaire spécialisé (vocabulaire contrôlé) qui va établir la correspondance entre les mots des textes et les concepts de l'ontologie ;
 - Disposer d'une logique permettant de résoudre les problèmes linguistiques ;
 - Transformer des descriptions vagues en descriptions concrètes.
-
-

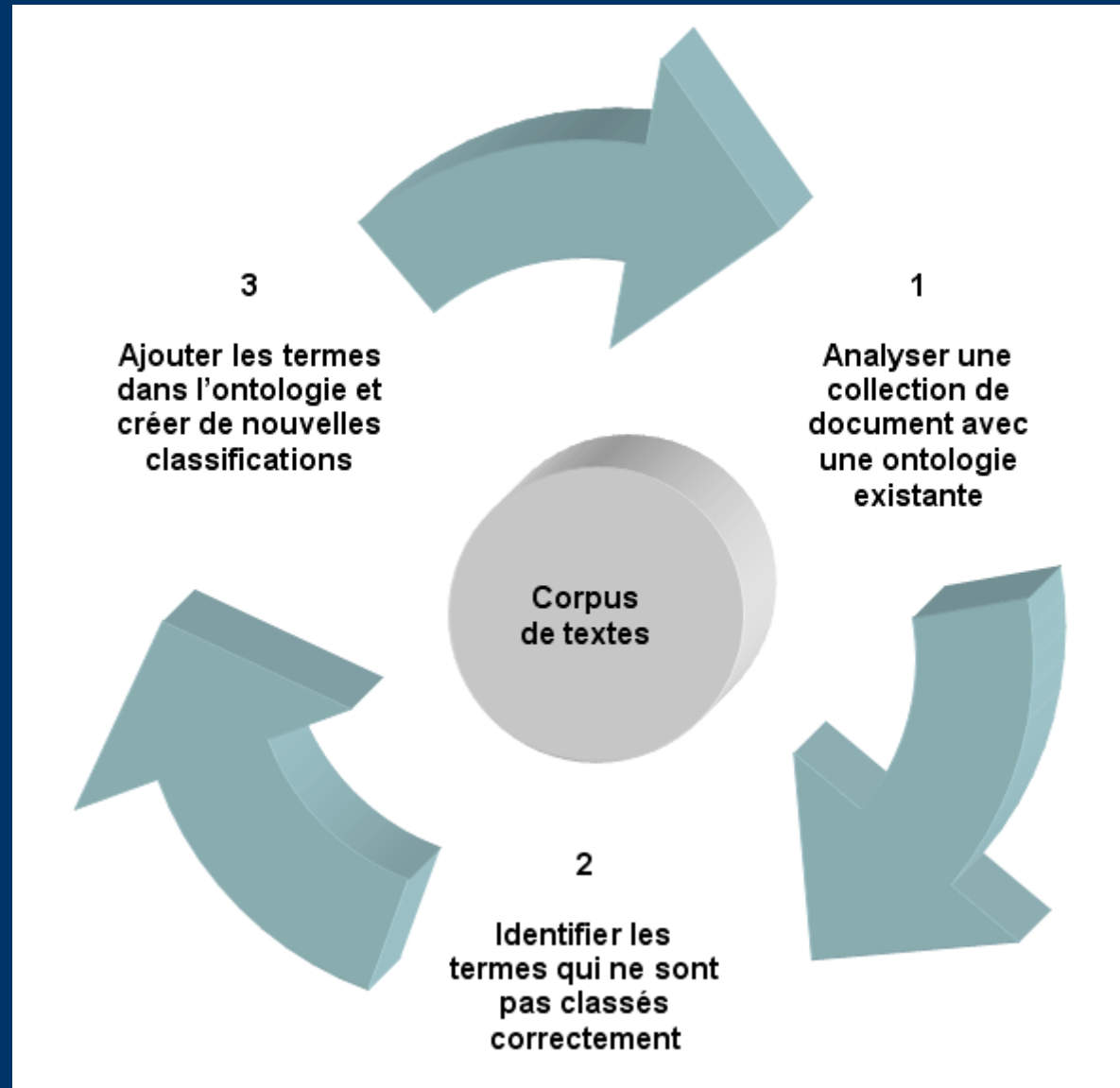
Créer une ressource sémantique = résoudre plusieurs problèmes :

- Grammaticaux (par exemple "livre" correspond à un nom dans "une pile de livres" et à un verbe dans "on nous livre du béton") ;
 - Sémantiques (par exemple "livre" indique une œuvre littéraire "un livre", une monnaie "la livre" ou un poids "une livre de beurre") ;
 - Qualitatifs (les fautes de frappe, de typographie ou d'orthographe peuvent créer certains contresens fâcheux.
-
-

Protocole de test et d'évaluation

- Partir d'une première classification
 - Analyser une collection conséquente de documents représentatifs du sujet traité
 - Identifier tous les termes qui ne sont pas pris en compte dans la classification et qui sont pertinents par rapport à la problématique
 - Rajouter les termes pertinents dans la classification et repartir à la première étape
-
-

Test cyclique d'ontologie



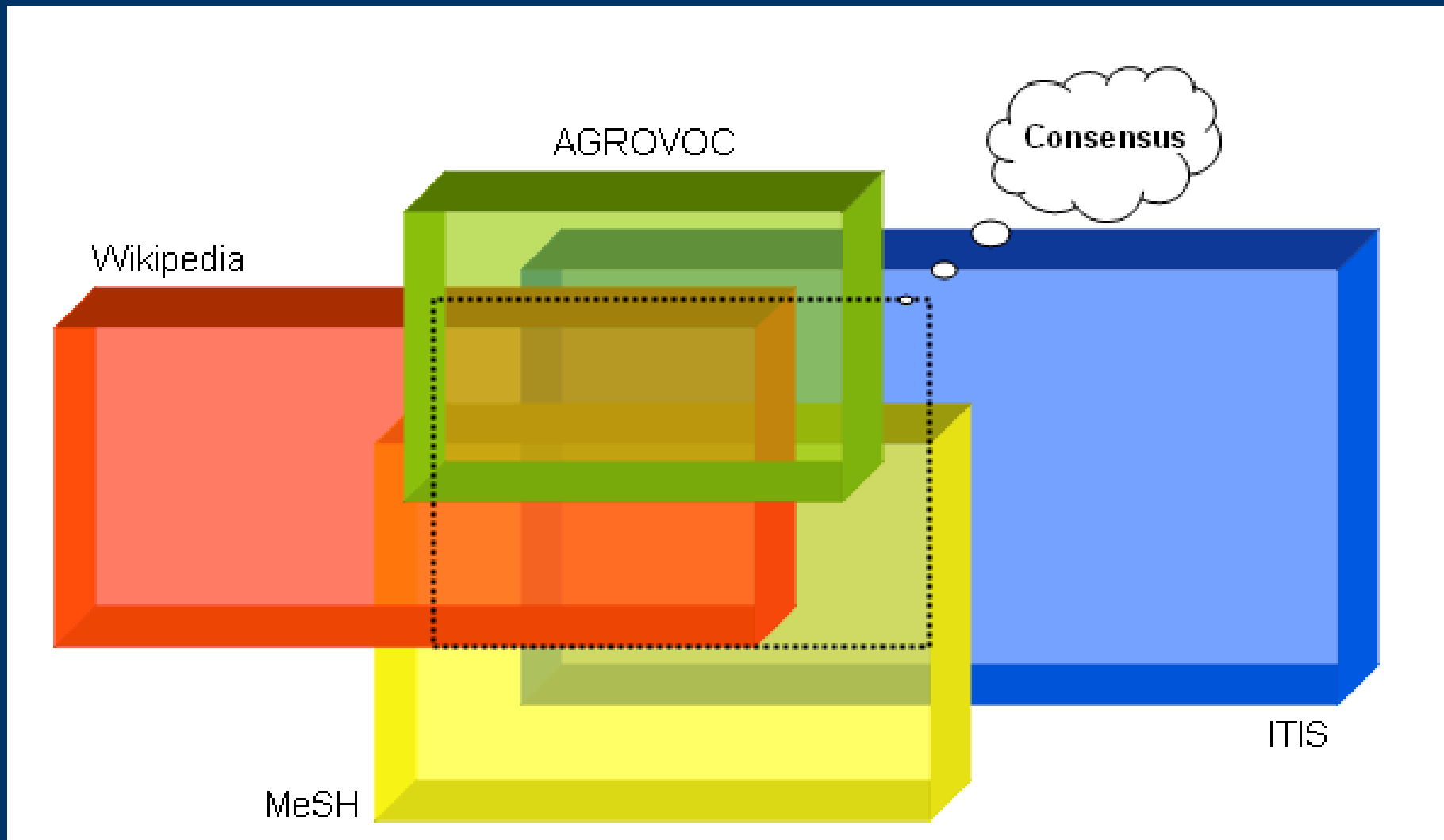
Nécessité d'une ontologie correcte et suffisamment exhaustive

- Un travail de classification n'est acceptable que s'il est suffisamment exhaustif (la majorité des utilisateurs trouvent ce qu'ils cherchent) ;
 - Ceci explique l'abandon des thesaurus et des annuaires internet au profit des moteurs de recherche en texte intégral ;
 - Les moteurs de recherche sémantiques permettent de combiner les deux mondes (en ajoutant des ontologies au full-text).
-
-

Extension d'AGROVOC par ajout d'ontologies complémentaires

- Conserver toutes les arborescences de classifications jugées pertinentes ;
 - Remplacer les classifications généralistes (peu pertinentes) par les ontologies utilisées en standard dans le moteur Tropes d'Acetic ;
 - Etendre ou remplacer certaines classifications scientifiques en utilisant à la fois ITIS, MeSH, Wordnet et d'autres sources (pertinentes).
-
-

Trouver un consensus entre plusieurs ontologies



Résultat de la fusion d'ontologies

- Le nombre de termes contrôlés a pu être augmenté de plus de 200% (par rapport à la version française AGROVOC) ;
 - La classification est jugée satisfaisante, dans la majorité des cas, bien qu'elle ne soit pas complète (et elle ne le sera jamais parce que le vocabulaire évolue continuellement) ;
 - L'absence d'informations terminologiques précises dans les ontologies OWL nécessite un lourd travail d'arbitrage.
-
-

Perspectives d'avenir

- Selon le W3C, les ontologies devraient constituer un socle important du Web 3.0, en le structurant très fortement de façon sémantique
 - Les normes actuelles, comme OWL, ne permettent pas une intégration rapide des classifications, parce qu'elles ignorent certains problèmes linguistiques
 - Les normes d'ontologies devront donc évoluer vers un modèle unifié et réellement sémantique
 - Toutefois OWL est, dès aujourd'hui, une norme exploitable dans de nombreux logiciels propriétaires
-
-