

Construction et enrichissement automatique d'ontologie à partir de ressources externes

JFO 2009
Jeudi 3 décembre 2009

E. Kergosien (LIUPPA, Pau)

M. Kamel (IRIT - UPS, Toulouse)

M. Sallabery (LIUPPA, Pau)

M.N. Bessagnet (LIUPPA, Pau)

N. Aussenac (IRIT - UPS, Toulouse)

M. Gaio (LIUPPA, Pau)



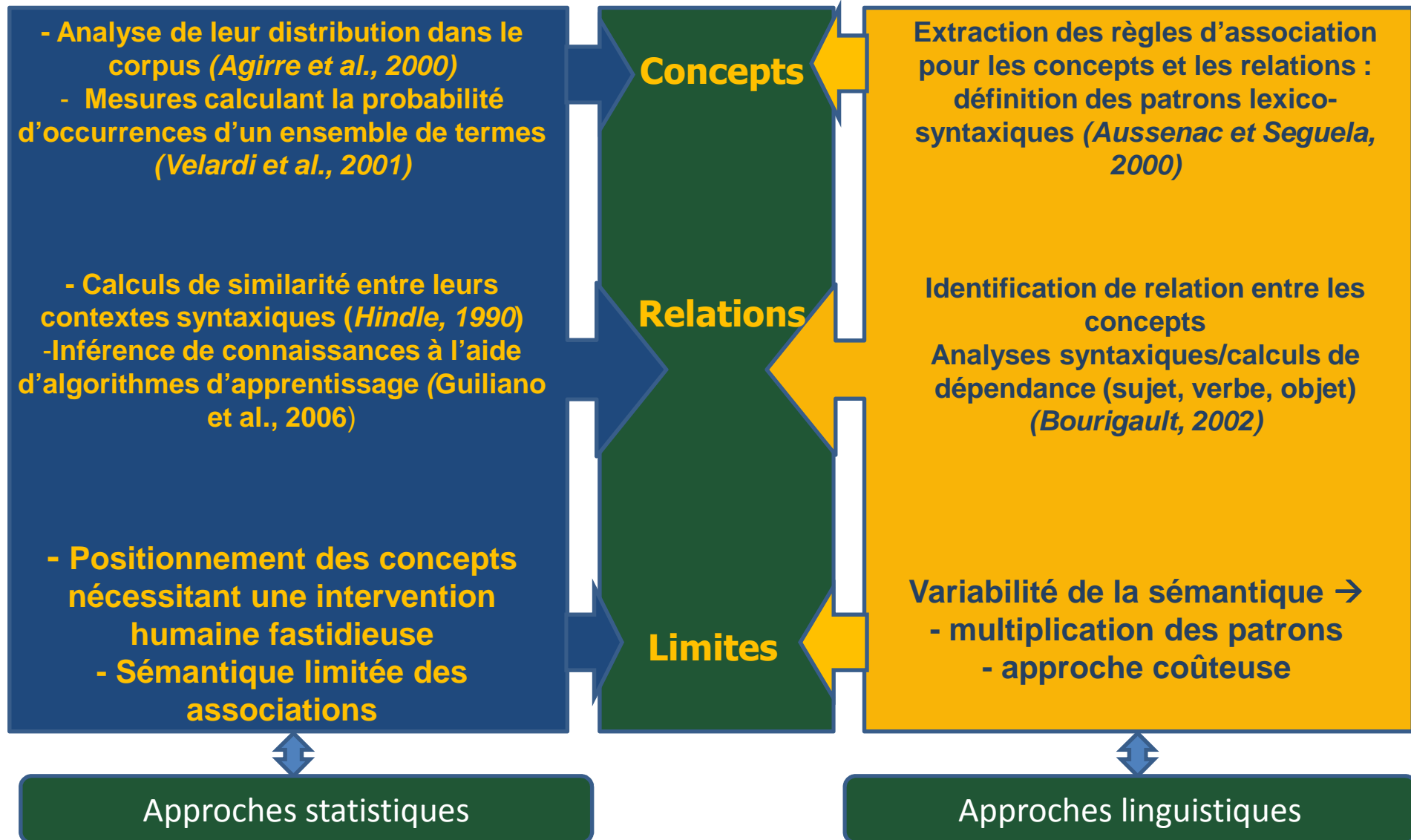
Sommaire

- Présentation du sujet & problématique
- Travaux connexes sur la création et l'enrichissement d'ontologie de domaine à partir de textes
- Démarche globale
- Application au domaine de la géographie
 - Le projet Géonto
 - Construction d'ontologie
 - Identification et Extraction des EN spatiales dans les textes grand publics
 - Enrichissement de l'ontologie
- Evaluation
- Conclusions et perspectives

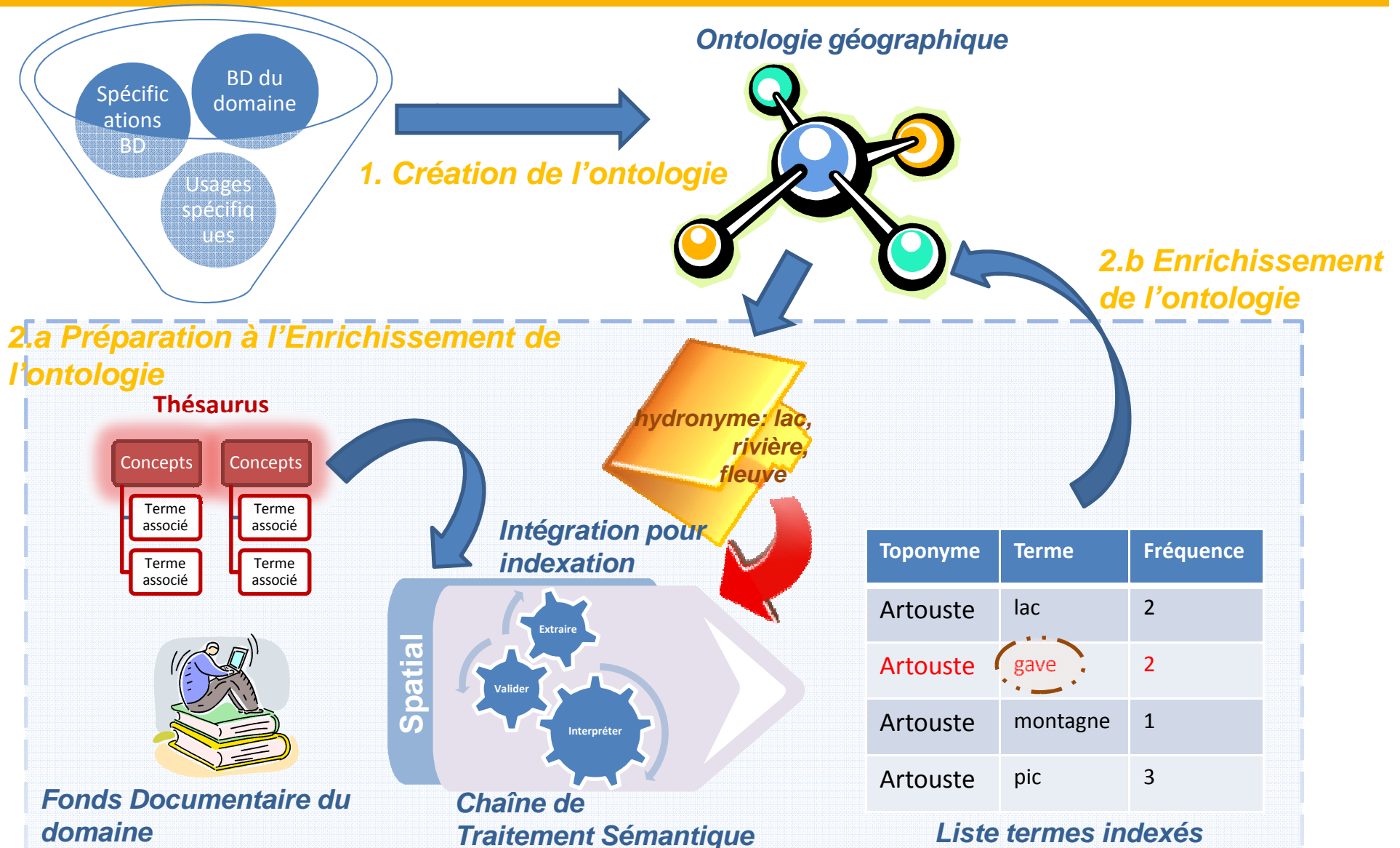
Contexte et Problématique

- Contexte
 - Chaîne d'annotation automatique d'entités nommées (EN) basée sur une analyse lexicale et syntaxique de fonds documentaires textuels
- Problématique
 - Difficulté de typer des EN candidates:
 - Ex : Difficile de proposer une représentation adaptée pour les syntagmes nominaux
 - *le lac d'Artouste et le mont d'Artouste,*
 - *le président Mitterrand et les années Mitterrand*
- Démarche globale
 - Concevoir, puis enrichir une ontologie de domaine, à partir de documents de description de ressources.
 - Spécificités : Le choix des ressources utilisées pour la validation et le calcul de représentations associées aux EN détectées

Travaux connexes



Méthodologie générale



Application au domaine de la géographie



<http://geonto.lri.fr/>



ANR-07-MDCO-005

The logo for cap-digital is a red rectangular box with the text 'cap-digital' in white, lowercase, sans-serif font. Below the main text, the words 'Paris Region' are written in a smaller, white, sans-serif font.

Le projet Géonto

- Projet ANR mené dans le cadre de l'édition 2007 du programme « Masse de Données et Connaissances »



- Traitement de données cartographiques (Conception Objet et Généralisation de l'Information Topographique, COGIT à l'IGN),



- Construction d'ontologies (IRIT- Université de Toulouse),



- Annotation et indexation automatisée d'informations géographiques dans des fonds documentaires textuels (LIUPPA – Université de Pau),



- Alignement d'ontologies (LRI – Paris 11).

- Objectif : rendre interopérable les bases de données géographiques hétérogènes dont dispose le COGIT.
- Démarche :
 - Proposer une ontologie par base de données,
 - Aligner les ontologies obtenues avec une ontologie de référence construite semi automatiquement par le COGIT.

Construction de l'ontologie

Document de spécifications de bases de données au format XML

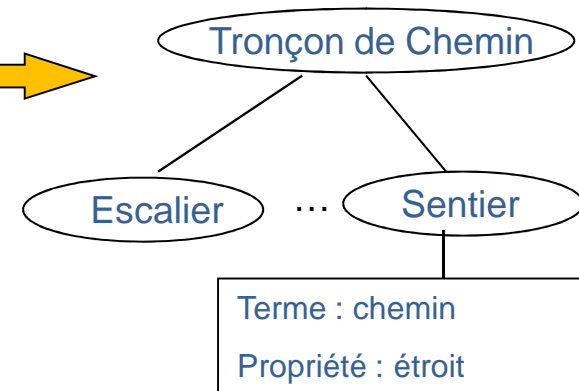
Construction d'ontologie à partir :

- De l'analyse de la structure du document

```

est-un {
  <class>
    <className> Tronçon de Chemin </className>
    <valueName> Escalier </valueName>
    <valueName> Sentier </valueName>
  </class>

```



De l'analyse linguistique du texte rédigé

```

<value>
  <valueName>Sentier</valueName>
  <description type="definition">chemin étroit ne permettant pas ... </description>
</value>

```

Construction de l'ontologie

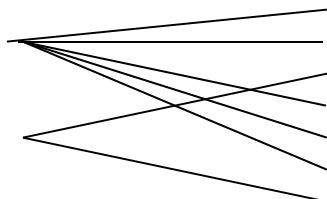
Extrait de l'ontologie

- ▼ ● <http://www.owl-ontologies.com/unnamed.owl#Voie>
 - ▼ ● http://www.owl-ontologies.com/unnamed.owl#Voies_de_communication_routière
 - ▶ ● http://www.owl-ontologies.com/unnamed.owl#Voies_de_communication_routière-Surface_de_route
 - ▼ ● http://www.owl-ontologies.com/unnamed.owl#Voies_de_communication_routière-Tronçon_de_chemin
 - ▶ ● http://www.owl-ontologies.com/unnamed.owl#Voies_de_communication_routière-Tronçon_de_chemin-Chemin
 - ▶ ● http://www.owl-ontologies.com/unnamed.owl#Voies_de_communication_routière-Tronçon_de_chemin-Chemin_empierré
 - ▶ ● http://www.owl-ontologies.com/unnamed.owl#Voies_de_communication_routière-Tronçon_de_chemin-Escalier
 - ▶ ● http://www.owl-ontologies.com/unnamed.owl#Voies_de_communication_routière-Tronçon_de_chemin-Piste_cyclable
 - ▼ ● http://www.owl-ontologies.com/unnamed.owl#Voies_de_communication_routière-Tronçon_de_chemin-Sentier
 - ▶ ● http://www.owl-ontologies.com/unnamed.owl#Voies_de_communication_routière-Tronçon_de_chemin-Sentier-Allée_piétonne_étroite
 - ▶ ● http://www.owl-ontologies.com/unnamed.owl#Voies_de_communication_routière-Tronçon_de_chemin-Sentier-Piste_de_cross
 - ▶ ● http://www.owl-ontologies.com/unnamed.owl#Voies_de_communication_routière-Tronçon_de_chemin-Sentier-Ruelle_étroite
 - ▶ ● http://www.owl-ontologies.com/unnamed.owl#Voies_de_communication_routière-Tronçon_de_chemin-Sentier-Sentier
 - ▶ ● http://www.owl-ontologies.com/unnamed.owl#Voies_ferrées_et_autres_moyens_de_transport_terrestre
 - ▶ ● http://www.owl-ontologies.com/unnamed.owl#Voies_de_communication_routière-Tronçon_de_chemin-Franchissement
 - ▶ ● http://www.owl-ontologies.com/unnamed.owl#Voies_de_communication_routière-Tronçon_de_route-Franchissement
 - ▶ ● http://www.owl-ontologies.com/unnamed.owl#Voies_ferrées_et_autres_moyens_de_transport_terrestre-Tronçon_de_voie_ferrée-Franchissement
 - ▶ ● http://www.owl-ontologies.com/unnamed.owl#Voies_ferrées_et_autres_moyens_de_transport_terrestre-Tronçon_de_voie_ferrée-Largeur_de_voie_ferrée
 - ▶ ● http://www.owl-ontologies.com/unnamed.owl#Voies_ferrées_et_autres_moyens_de_transport_terrestre-Tronçon_de_voie_ferrée-Électrifié
 - ▶ ● http://www.owl-ontologies.com/unnamed.owl#Zonages_techniques_et_administratifs
 - ▶ ● http://www.owl-ontologies.com/unnamed.owl#Zonages_techniques_et_administratifs-Commune-Multi-canton

Propriétés du concept *sentier*

Structure

Texte

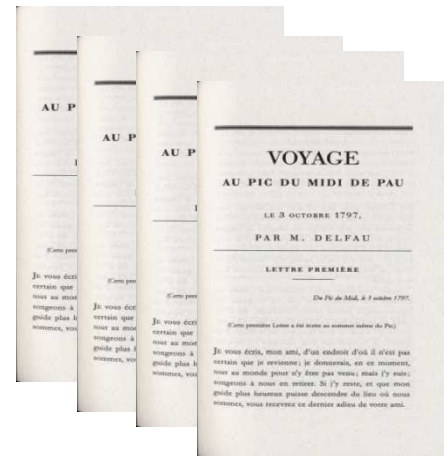


| Property | Value |
|--|---|
| <input type="checkbox"/> rdfs:comment | |
| <input checked="" type="checkbox"/> Definition | Chemin étroit ne permettant pas le passage de véhicules. |
| <input checked="" type="checkbox"/> Origine | Structure |
| <input checked="" type="checkbox"/> propriete | étroit |
| <input type="checkbox"/> rdfs:label | Voies_de_communication_routière-Tronçon_de_chemin-Sentier |
| <input checked="" type="checkbox"/> Reference | Voies de communication routière-Tronçon de chemin-Nature |
| <input checked="" type="checkbox"/> Terme | Sentier |
| <input checked="" type="checkbox"/> Terme_plus_Generique | Chemin |

Enrichissement de l'ontologie

- Proposition

- utiliser les termes associés à des toponymes dans des textes grand public afin d'enrichir l'ontologie géographique



14 livres (récits de voyages dans les Pyrénées) faisant partie d'un corpus fourni par la MIDR

- Méthode

- recoupement de termes du thésaurus RAMEAU avec les concepts de l'ontologie IGN-IRIT
- Pourquoi RAMEAU?
 - utilisé dans un grand nombre de bibliothèques françaises depuis les années 80 pour indexer manuellement des fonds documentaires territorialisés.
 - source de connaissance (+ de 111000 termes) couvrant les disciplines scientifiques, des loisirs, des arts, etc.
 - Gère la synonymie

RAMEAU : Exemple de notice

L'autorité matière *Grottes*

Grottes [+ subd. géogr.]

vedette matière nom commun - S'emploie en tête de vedette

<Employé pour :

- Abîmes
- Antres
- Avers
- Cavernes *Ancienne vedette* ----- Termes « employé pour »
- Cavernes préhistoriques
- Cavités souterraines
- Gouffres
- Grottes unies
- Grottes préhistoriques
- Préhistoire -- Grottes
- Spélniques

<<Terme(s) générique(s) : ----- Termes « génériques »

- [Habitat préhistorique](#)
- [Relief \(géographie\)](#)
- [Zones souterraines](#)

>>>Terme(s) associé(s) : ----- Termes « associés »

- [Abris-sous-roche](#)
- [Architecture troulodytique](#)
- [Art pariétal](#)
- [Écologie des cavernes](#)
- [Kart](#)
- [Spéléologie](#)

Voir aussi aux noms des grottes particulières, par ex. : Lascaux, Grotte de (Dordogne) ; Arago, Caune de l' (Pyrénées-Orientales)

>>Terme(s) spécifique(s) : ----- Termes « spécifiques »

- [Grottes de jardin](#)
- [Grottes marines](#)
- [Spéleothèmes](#)

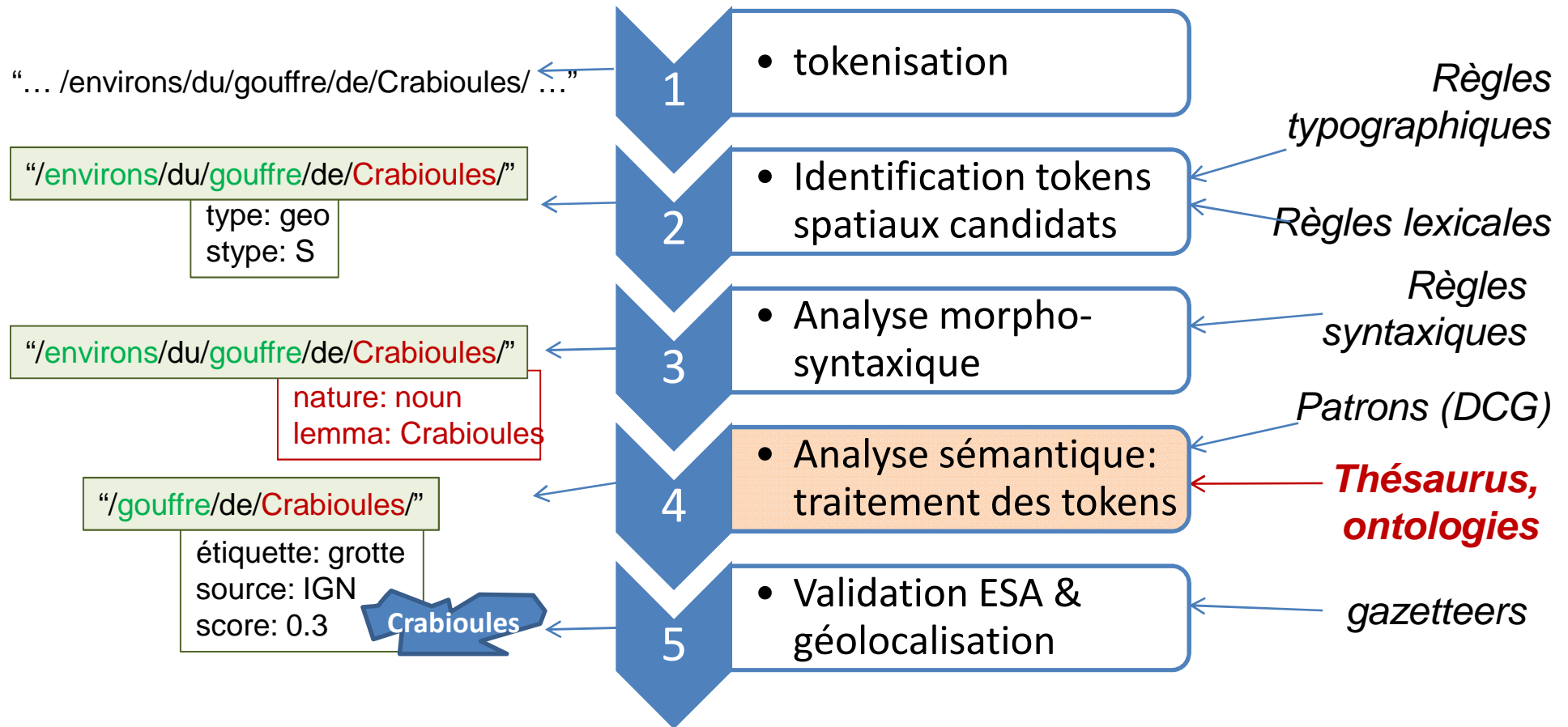
Source(s) : - Grand Larousse universel. - Grand Rubric de la langue française, 2001
 - Dict. de géologie / A. Foucault, J.-F. Raoult, 2005. - Dict. de la géographie / P. George, 1974. - Les mots de la géographie : dictionnaire / R. Brunet, 1993. - La préhistoire : hist. et géogr. / D. Valou, 2004. - Dict. de la préhistoire / A. Leroi-Gourhan, 1994

Equiv. LCSH : Caves

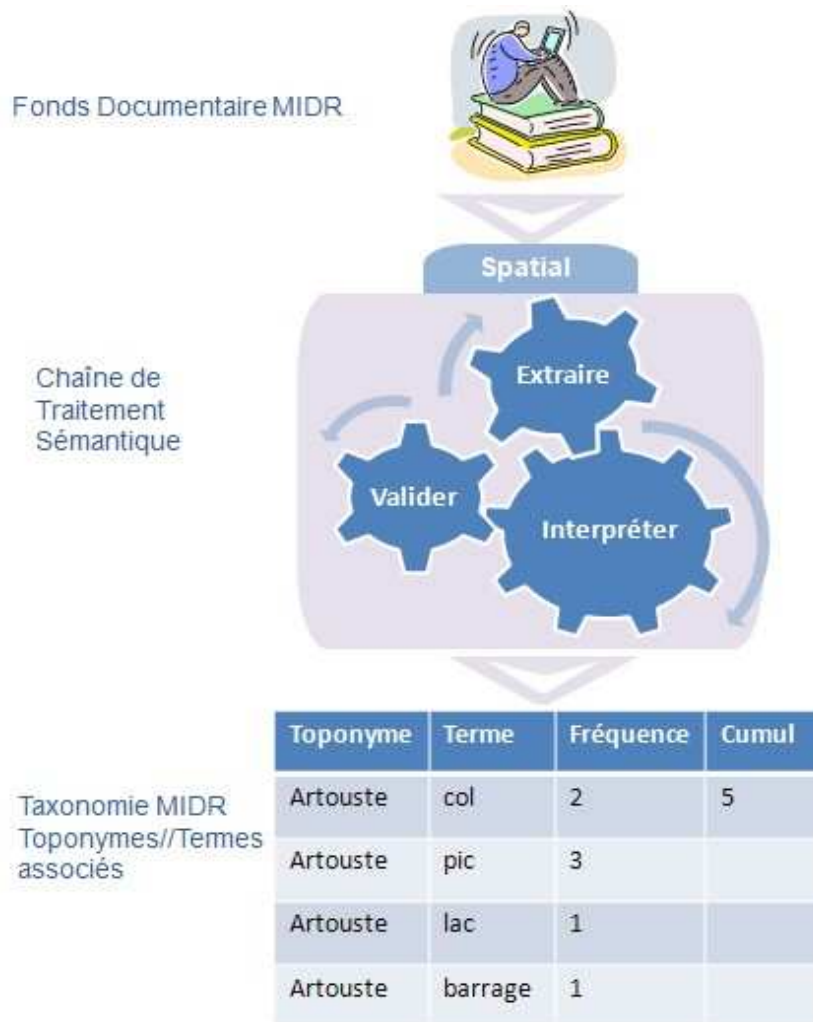


Identification et Extraction des EN spatiales dans les textes grand publics

“... dans les environs du gouffre de Crabioules ...”



Identification et Extraction des EN spatiales dans les textes grand publics



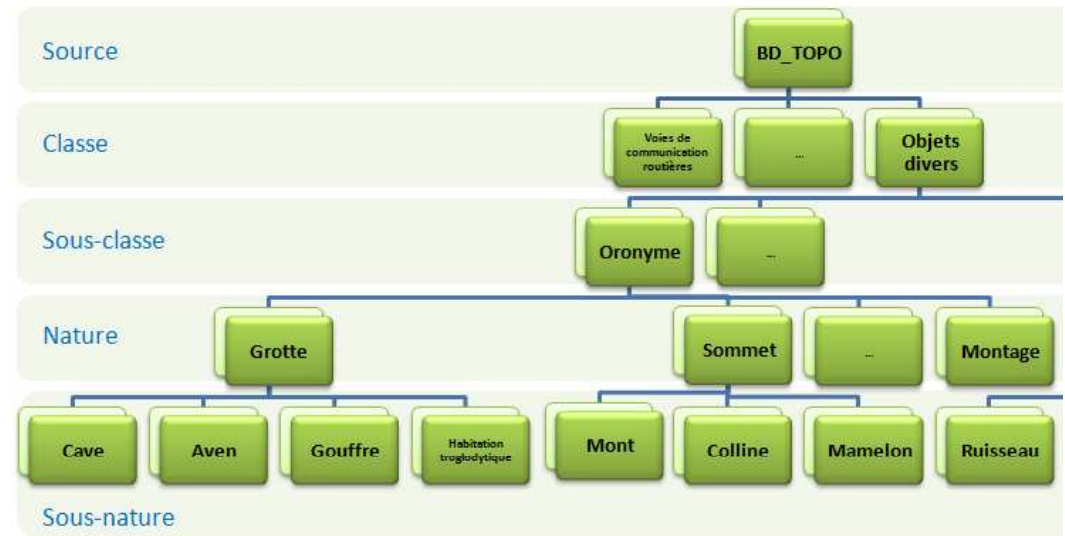
Exemple de résultats:

| Crabioules | | | |
|-------------|---------------|------------------------|------------------|
| Occurrences | Terme associé | Ontologie Géographique | Thesaurus RAMEAU |
| 1 | abîme | | X |
| 2 | col | X | X |
| 1 | corniche | | X |
| 1 | crête | X | X |
| 1 | mont | X | X |
| 1 | promenade | | X |
| 1 | route | X | X |
| 1 | sommet | X | X |

→ *Abîme*, *Corniche* et *Promenade* sont candidats à enrichir l'ontologie:

Enrichissement de l'ontologie (2)

Extrait de l'ontologie IGN



Extrait d'une notice RAMEAU

Grottes [+ subd. géogr.]

Vedette matière nom commun . S'emploie en tête de vedette

<Employé pour :

Abîmes

Antres

Avens

Cavernes *Ancienne vedette*

Cavernes préhistoriques

Cavités souterraines

Gouffres

Grottes ornées

Grottes préhistoriques

Préhistoire – Grottes

Spélonques

Recoupement par classe d'équivalence:

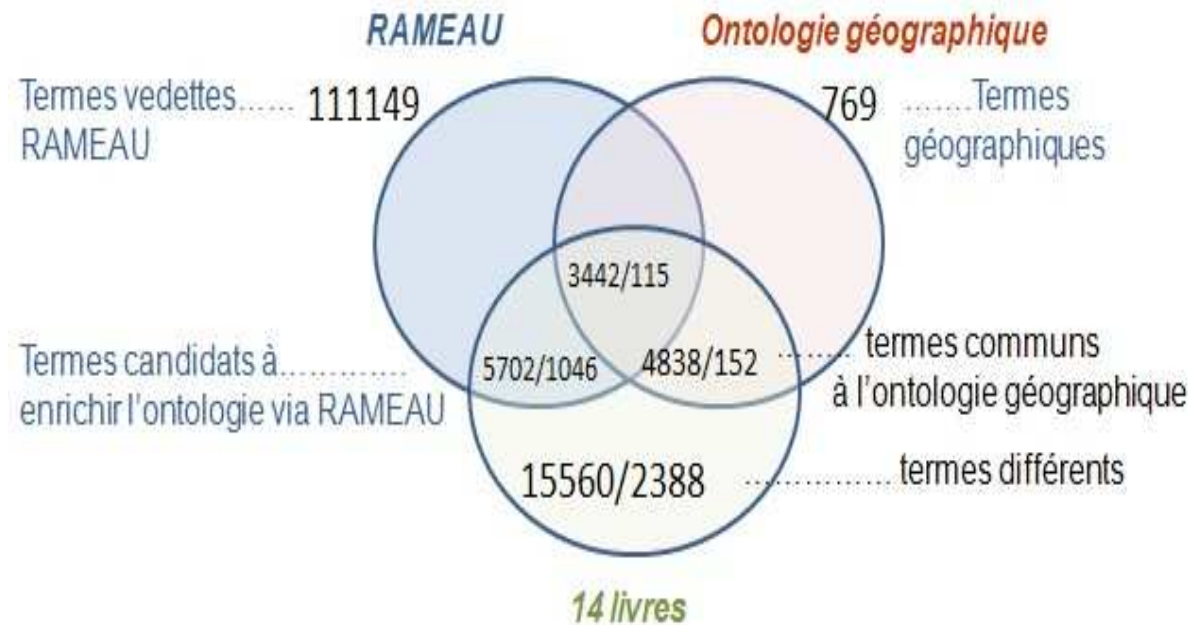
Si « Abîme » existe dans RAMEAU

Si au moins un des termes liés dans RAMEAU existent dans Ontologie IGN

- proposition d'enrichissement de l'ontologie avec le concept de « nature » qui a le plus grand nombre d'équivalence(s).

Ici « Abîme » est proposé comme sous-nature de Grotte car on peut identifier 3 termes équivalences (Grotte, Aven, et Gouffre)

Estimations



Toponymes candidats & termes associés.

→ 1046 termes RAMEAU sont candidats à l'enrichissement de notre ontologie.

Limites actuelles

- Construction d'ontologie
 - La qualité de l'ontologie obtenue dépend entièrement de la qualité des spécifications
 - lorsque des incohérences existent au niveau des spécifications, une intervention humaine (expert) s'impose pour corriger l'ontologie.
Ex : énumérations de concepts qui ne sont pas forcément de même niveau, comme *Rues* et *Rues piétonnes*.
- Enrichissement d'ontologie
 - RAMEAU ne nous permet pas d'apporter de réponse pour l'ensemble des termes candidats à l'enrichissement
 - certains des termes engendrent des contres sens qui peuvent amener à générer des résultats erronés.
Ex : les termes *glacier* et *gorges* notamment sont présents deux fois dans RAMEAU (sens géographique et autre). Travail à réaliser sur le contexte dans RAMEAU.

Conclusions

- Proposition d'une méthode pour la construction d'ontologie couvrant au mieux un domaine sémantique cible
 1. Travail sur le contenu et la structure de documents de description de ressources pour la création automatique d'ontologie de domaine;
 2. Enrichissement de l'ontologie obtenue à partir d'une analyse de textes du même domaine.
 - Expérimentation dans le domaine Géographie
 1. Traitement de corpus textuels composés de récits de voyages dans les Pyrénées.
 2. Intégration de l'ontologie géographique dans une chaîne de traitement pour l'annotation automatique d'EN spatiales dans un corpus qui décrit un territoire (récits de voyage, journaux, etc.)

Premiers résultats : diminution de plus de moitié des ambiguïtés pouvant apparaître lors de la validation et du calcul de représentations associées aux EN détectées.
- L'ontologie géographique générée offre un moyen efficace d'optimiser l'accès à des ressources grand public de type récits de voyages.

Perspectives

- Construction d'ontologie
 - Approche plus générique basée sur la structure du document (approche par patrons structurels --> complémentarité entre les patrons lexico-syntaxiques et patrons structurels)
 - Analyse linguistique plus fine du contenu: exploitation des conjonctions/disjonctions, des exceptions ("sauf ..."), etc.
- Enrichissement d'ontologie
 - Affinement des résultats obtenus afin de détecter les termes RAMEAU porteur d'un sens géographique.
 - identification des termes pouvant engendrer des contres sens, dus à l'hétérogénéité des ressources utilisées, qui peuvent amener à générer des résultats erronés.
 - Prise en compte de contexte de chaque terme dans le vocabulaire contrôlé utilisé afin d'identifier la bonne vedette à exploiter.
- Evaluation
 - Tests sur un fonds documentaire conséquent pour valider la démarche



pour votre attention

Eric KERGOSIEN - LIUPPA