
Avancement Lot1 – dec 2009

Construction et enrichissement d'ontologie



LIUPPA : Christian, Eric, Marie-Noelle,
Mauro, Tien,



IRIT : Marion, Mouna, Nathalie



07/12/2009

Geonto - Lot1



1

Plan

- Points contractuels
- Avancement IRIT
- Avancement LIUPPA
- Perspectives

Points contractuels

- Modifier le calendrier en cohérence avec la demande de prolongement du projet
- Livrables du Lot 1.1
 - 1.3 Conception d'une chaîne de traitement pour appliquer les patrons lexico-syntaxiques.
 - On a proposé une V1 à T0 + 18 et on produira une V2 à T0+24.
- Livrables du Lot 1.2
 - 1.6
 - Prévu : Version 1 du module logiciel de construction T0 + 18
 - on propose : Version 1 de l'ontologie enrichie T0 + 24
 - Tests combinant différentes techniques sur les taxonomies du COGIT T0 + 24
 - 1.7
 - Prévu : tout à T0 + 24
 - On propose : tout à T0 + 30
 - Tests sur les ontologies obtenues en résultat du sous-lot 1.1. Evaluation de la robustesse du prototype développé.
 - Validation des résultats d'un point de vue géographique.

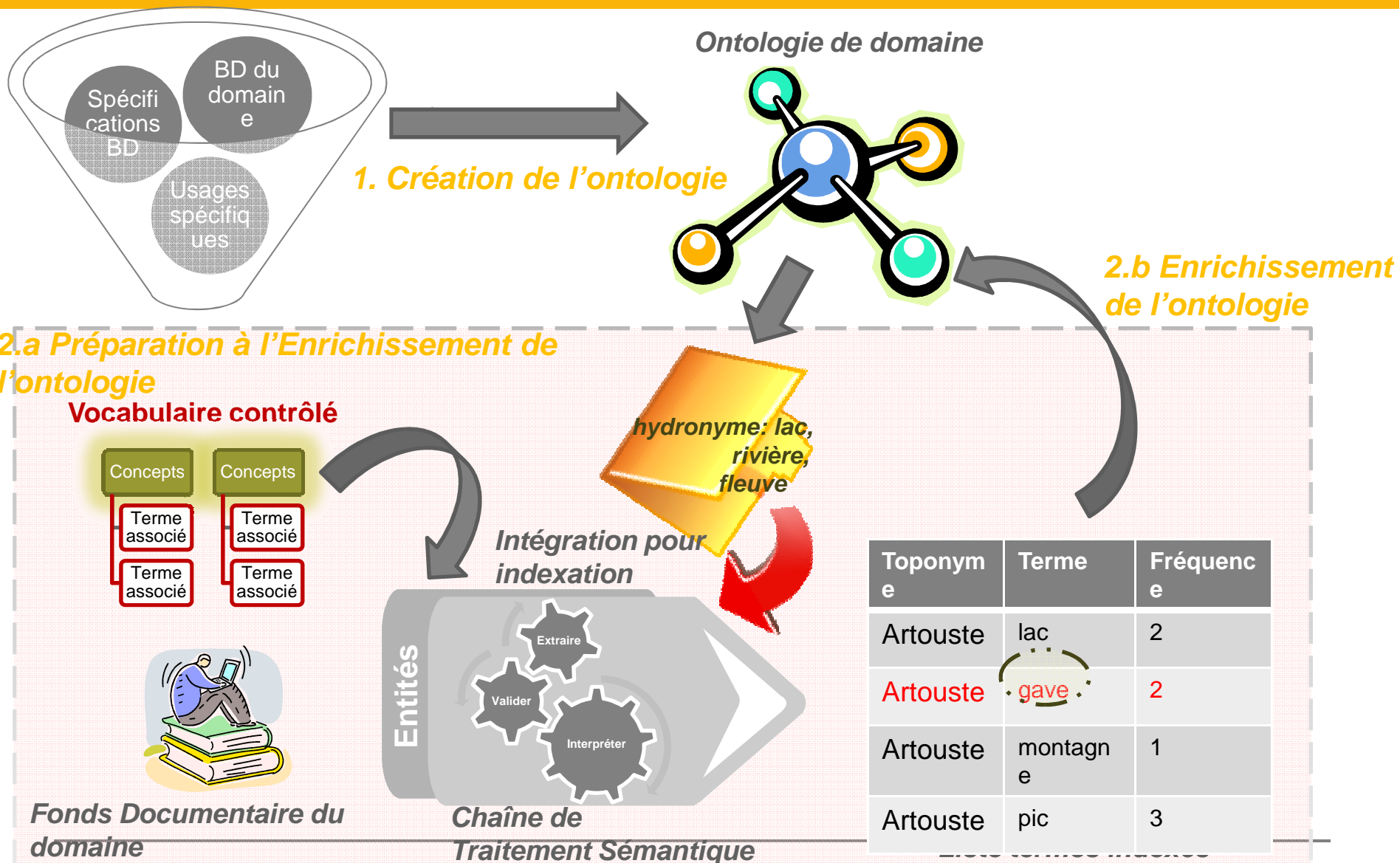
Livrables lot 1

	Libellé du livrable	Type	Responsable	Partenaires participants	Date
0	Site web du projet – Mise en place au plus tard 6 mois après le démarrage du projet et mise à jour au moins semestrielle	Web	Coordonnateur	Tous	T0+6
1	Mise au point d'outils d'extraction de concepts et de relations	Rapport intermédiaire	IRIT	IRIT, LIUPPA, COGIT	T0 + 12
2	Alignement d'ontologies	Rapport intermédiaire	LRI	LRI	T0 + 12
3	Mise au point d'outils d'extraction de concepts et de relations	Module logiciel (V1)	IRIT	IRIT, LIUPPA, COGIT	T0 + 18
4	Enrichissement d'une ontologie existante à partir de textes à l'aide des outils d'extraction et à partir de ressources lexicales	Module logiciel	IRIT	IRIT, LIUPPA, LRI, COGIT	T0 + 18
5	Intégration et accès aux schémas de bases de données	Rapport intermédiaire	COGIT	COGIT	T0 + 18
6	Indexation automatique de contenu de documents	Rapport intermédiaire	LIUPPA	LIUPPA	T0 + 18
7	Mise au point d'outils d'extraction de concepts et de relations	Rapport intermédiaire Module logiciel (V2)	IRIT	IRIT, LIUPPA	T0 + 24
4	Enrichissement d'une ontologie existante à partir de textes à l'aide des outils d'extraction et à partir de ressources lexicales	Module logiciel V2	IRIT	IRIT, LIUPPA, LRI, COGIT	T0 + 30
7	Mise au point d'outils d'extraction de concepts et de relations	Rapport final Module logiciel (V3)	IRIT	IRIT, LIUPPA	T0 + 24
7bis	Mise au point d'outils d'extraction de concepts et de relations	Rapport final Module logiciel (V3)	IRIT	IRIT, LIUPPA	T0 + 30
8	Alignement d'ontologies	Rapport final Module logiciel	LRI	LRI	T0 + 30
9	Enrichissement d'une ontologie existante à partir de textes à l'aide des outils d'extraction et à partir de ressources lexicales	Rapport final Module logiciel	LRI	IRIT, LRI	T0 + 30
10	Réconciliation d'instances pour l'alignement d'ontologies	Rapport final Module logiciel	LRI	LRI	T0 + 30
11	Restructuration d'une ontologie construite automatiquement	Module logiciel	N. Aussenac-Gilles	LRI	T0 + 36

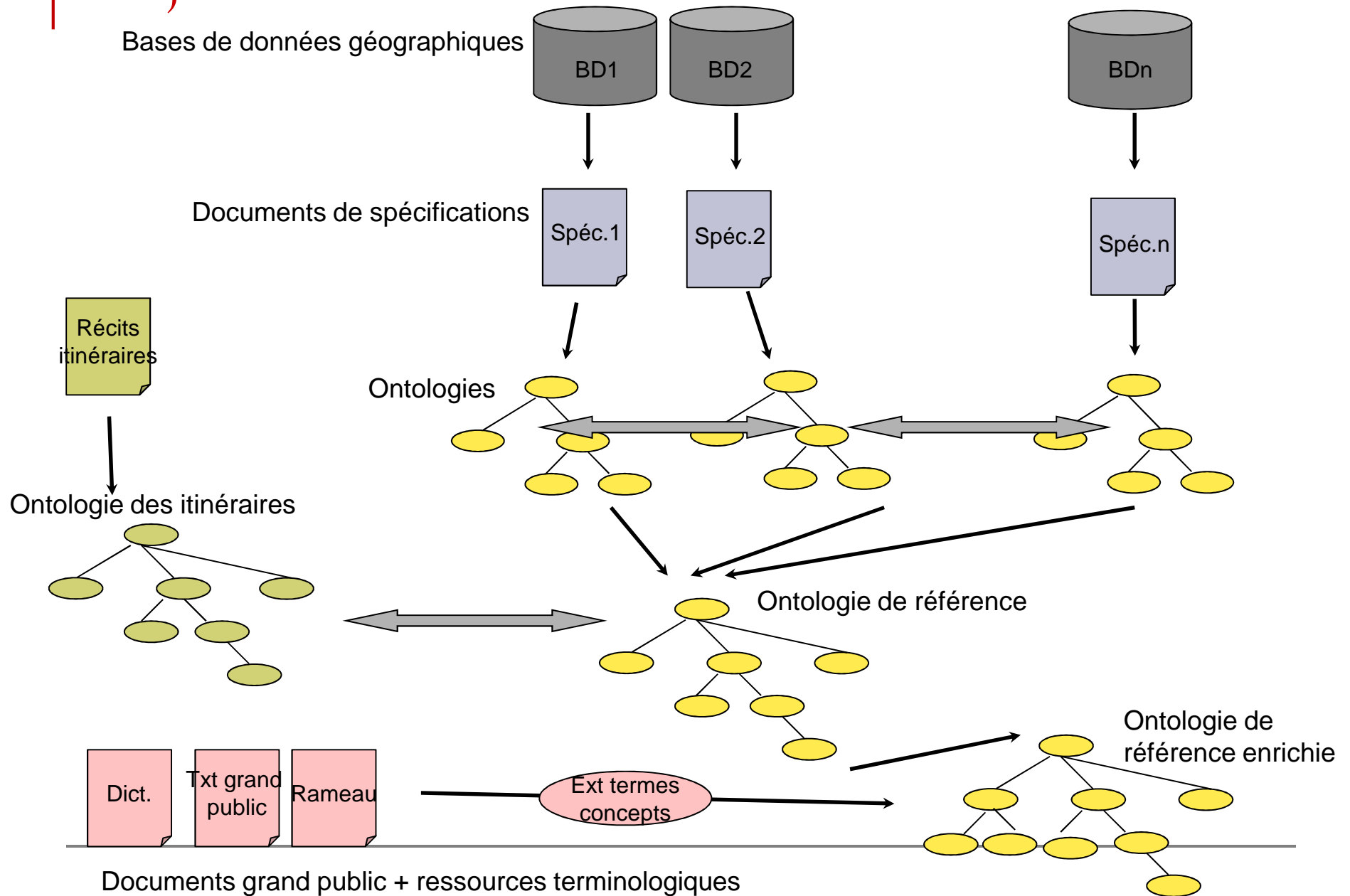
07/12/2009

Lot 1 - Module logiciel N. Aussenac-Gilles

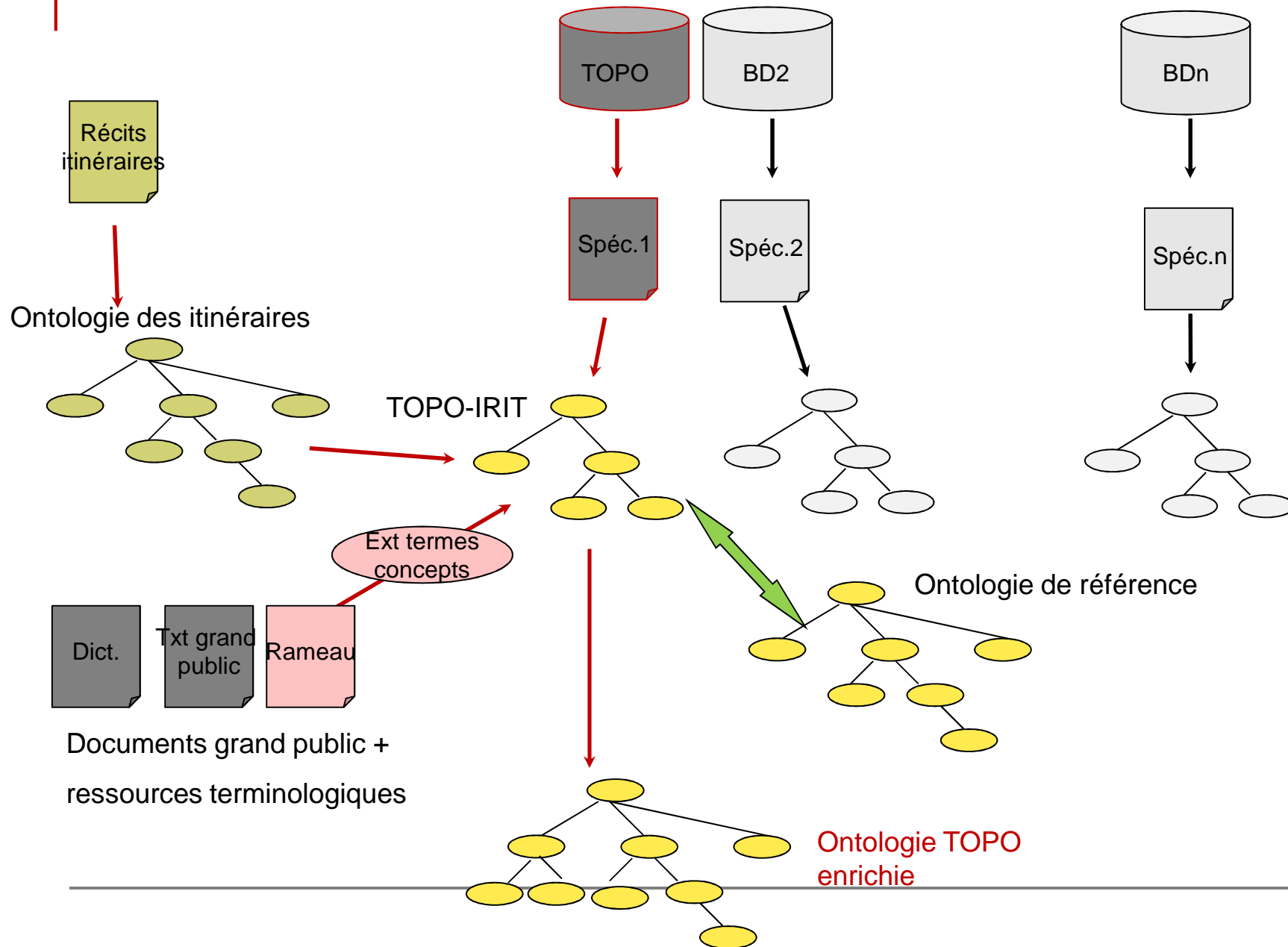
Méthodologie générale Lot 1



Objectifs lot 1



Réalisé lot 1



Projet global lot 1

- **avancement travail IRIT**

- nouvelle version des règles de génération d'ontologie à partir des spécifications / Mouna
- apports de l'analyse TAL des définitions et partie rédigées, premières pistes / Marion
- alignement / validation des ontologies apprises avant fusion / Nathalie

- **avancement travail LIUPPA**

- développements pour manipuler des ontologies depuis LinguaStream / Tien et Christian
- proposition d'enrichissement avec Rameau : point ERic et Marie-Noelle

Rappel de la méthode

- Phase 1 : Analyse de la structure
 - Ontologie
 - Approche procédurale spécifique au document de spécifications des Bases de Données COGIT

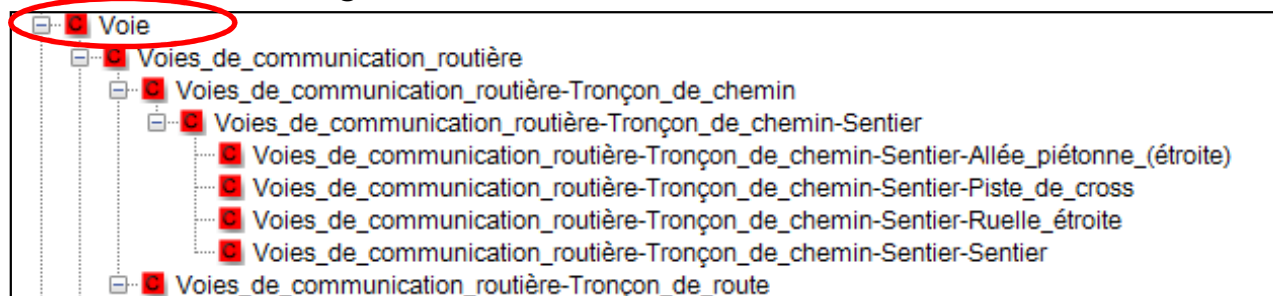
C:\Mouna\IRIT\Projets\ANR\Géonto\Développement\Construction_Ontologie\Geonto.jape

→ Approche ad-hoc

- Phase 2 : Analyse Linguistique des champs "*définition*"
 - Ontologie → Approche générique
 - Définition de patrons lexico-syntaxiques (usages de la langue)
-

Rappel des résultats obtenus

Extrait de l'ontologie



Propriétés du concept *Sentier*

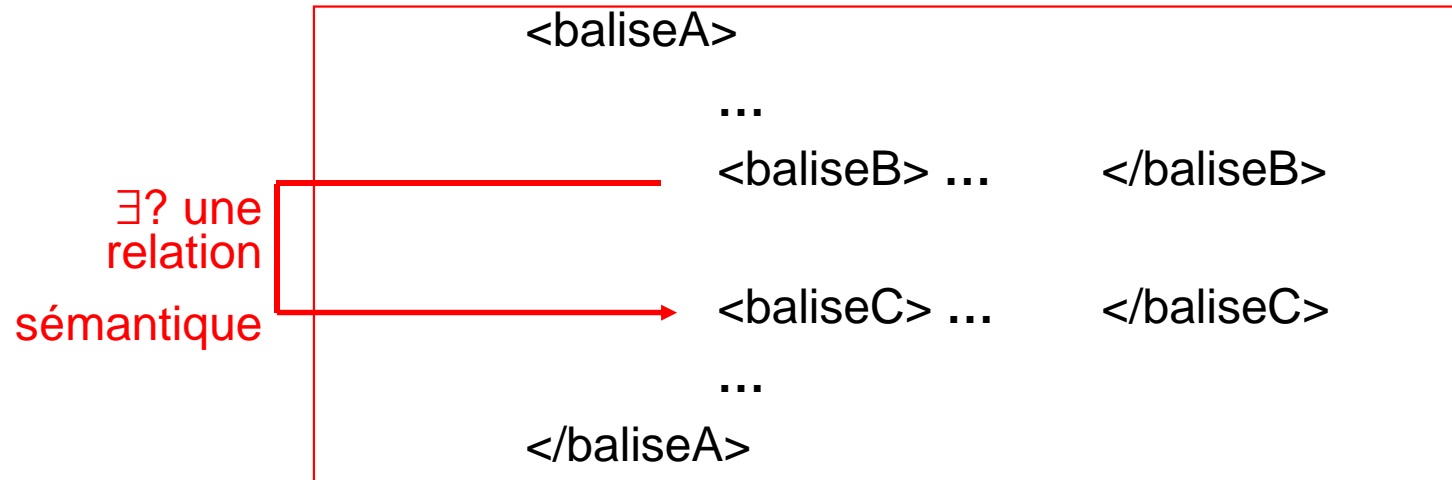
▼ Property Values	
propriete	étroit
Terme_plus_Generique	Chemin
label	Voies de communication routière-Tronçon de chemin-Sentier
Terme	Sentier
Definition	Chemin étroit ne permettant pas le passage de véhicules.
Origine	Structure
Reference	Voies de communication routière-Tronçon de chemin-Nature

Les mappings comme aide à la validation

- 1-- un même terme désigne plusieurs concepts à différents niveaux de l'hierarchie
- 2 -- énumérations
- 3 -- concepts identifiés par des adjectifs

Approche générique concernant la structure

principe de base :

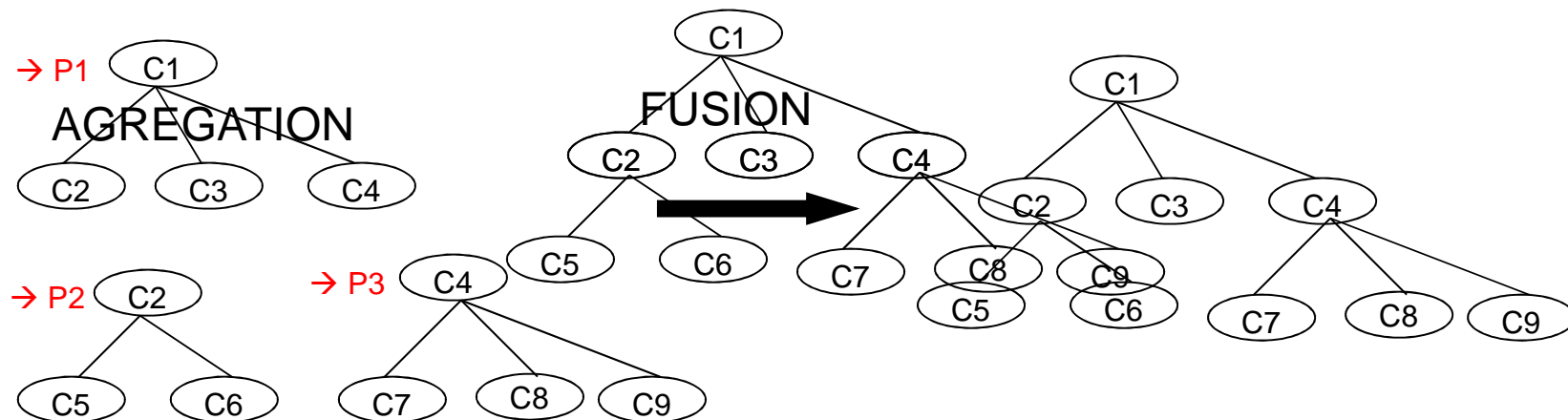


- proposer à l'expert toutes les possibilités (reste à charge pour l'expert d'identifier la relation)
- analyse de l'expert qui identifie les balises et les relations pertinentes

Approche générique concernant la structure

■ Méthode :

- Proposer un langage de contraintes pour que l'ontologue puisse exprimer les liens existant entre les balises imbriquées
- Générer automatiquement la règle (Patron Structurel) correspondante
- Projeter le patron sur le document et obtenir un fragment d'ontologie (1 père + ses fils)



→ La règle s'appuie sur le fait que le concept père existe déjà dans l'ontologie

Patron Structurel (PS)


- Caractérise la sémantique portée par un élément structurel en prenant en compte :
 - les balises correspondant aux éléments de structure
 - les noms de ces balises
 - leur niveau d'imbrication ou de hiérarchisation
- Sachant que :
 - plusieurs balises peuvent porter le même nom
 - avoir des localisations différentes dans le document

PS : motif basé sur les annotations → instructions

Exemple de contrainte et de PS

Document XML

```
<specification>
  <featureCatalogue>
    <package name="A - Voies de communication routière">
      <packageName>A - Voies de communication
routière</packageName>
      <class name="Tronçon de route">
        <className>Tronçon de route</className>
      ...
    </package>
  </featureCatalogue>
</specification>
```



Contraintes

- *packageName* sous /specification/featureCatalogue/package
- *className* sous /specification/featureCatalogue/package/class
- *packageName* et *className* sous la portée de *package*

→ Processus de Ré-annotation des balises

Exemple GEONTO

P1 /specification/featureCatalogue/package/packageName

P2 /specification/featureCatalogue/package
packageName/className
package/class

P3 /specification/featureCatalogue/package
packageName/className/attributeName
package/class/ attributes-attribute

P4 /specification/featureCatalogue/package
packageName/className/attributeName/valueName
package/class/ attributes-attribute/enumeratedValues-value

P5 /specification/featureCatalogue/package
packageName/className/attributeName/valueName/TermList
class/ attributes-attribute/enumeratedValues-value/description

Evaluation de la nouvelle approche

■ Inconvénients :

- Ne tient pas compte des spécificités du document
 - Interprétation des attributs en fonction de leur type
 - Liste des attributs dans *Regroupement* qui traduisent une relation *est-un*
 - Tous les attributs ont le même statut

Exemple :

dans OntoV1 : sentier ***est-un*** Tronçon de chemin
 Tronçon de chemin ***a-pour-Franchissement*** Pont

dans OntoV2 : Tronçon de chemin ***a-pour-Nature*** sentier
 Tronçon de chemin ***a-pour-Franchissement*** Pont

■ Avantages :

- Approche modulaire
 - Réutilisable sur d'autres documents
 - Aide à la conception d'ontologies à partir de textes
 - Dualité Patrons Structurels / Patrons lexico-syntaxiques
-

Post-doc de Marion

- analyse du texte
 - pousser plus loin l'étude des définitions (cf. travaux de Mouna)
 - Identifier d'autres types de relations
- analyse des marqueurs de relations spatiales
 - exploitation du champ "modélisation géométrique", lien avec des textes grand public (Pau)

Sous LinguaStream : création d'une chaîne simple - 1

- 1 je prends BDTOPO
- 2 je sélectionne uniquement les noeuds xml sur lesquels je souhaite travailler :
 - ▶ packageName/ className/ AttributeName/ valueName/ description-extensionalDefinition = "terme"
 - ▶ description-type :définition = "définition"
- 3 tokenisation simple (celle de LS par défaut → il faudra en faire une mieux)
- 4 Treetagger (idem : corriger quelques bugs soit en amont, soit en sortie du module)

Sous LinguaStream : création d'une chaîne simple - 2

- 5 lexique de termes/concepts extraits du documents lui-même
- 6 découpage des phrases et des parenthèses
- 7 repérage de SN divers (simples, coordonnés, avec adj ou participe passé, ect)
 - ▶ un traitement pour les termes/concepts
 - ▶ un autre pour les parties définition
- 8 repérage de "synonymie stricte" *i.e.* sans verbe dans la partie droite (C/T / définition : C/T unique)
- 9 visualisations diverses (concordancier, coloration de texte en fonction des analyses, ...)

Problèmes du lexique dans sa version actuelle - 3

Pb avec certains des termes/concepts extraits dans "AttributeName" et "valueName"

Pour certains, on peut se demander si ce sont des concepts du meme type que "voie de communication".
"nature", "franchissement", "nombre de voies"

Pour d'autres, ce ne sont pas des concepts (termes) mais plutot une caractéristique/ propriété qui s'applique à un concept (?)

- ▶ pour l'attributeName "*électrifié*" la valueName est *Oui/non*
- ▶ des valeurs numériques

Projet global lot 1

- **avancement travail IRIT**

- nouvelle version des règles de génération d'ontologie à partir des spécifications / Mouna
- apports de l'analyse TAL des définitions et partie rédigées, premières pistes / Marion
- alignement / validation des ontologies apprises avant fusion / Nathalie

- **avancement travail LIUPPA**

- développements pour manipuler des ontologies depuis LinguaStream / Tien et Christian
- proposition d'enrichissement avec Rameau : point ERic et Marie-Noelle

1-- un terme /plusieurs concepts à différents niveaux de hiérarchie

- Aide 1 : regroupe tous les concepts ayant le même label (ex tous les « péages »)
- Aide 2 : rapproche ces concepts de concepts de l'onto cible (comme cela se fait sur la base du terme, tous les mappings sont les mêmes pour ces différents concepts)
 - ex Péage isClose Aire de Péage
 - Comme il n'y en a qu'un dans l'onto cible, c'est le seul, le "meilleur" et il n'y a pas mieux.
- Pistes : exploiter la définition pour départager les sens

2-- Pb alignement avec des mots « vides »

■ Suggestion de mots vides

- ❑ Nature, regroupement, valeur, qualité, de type
- ❑ surface, tronçon, aire, zone, bâtiment ou bassin ...

■ Constat

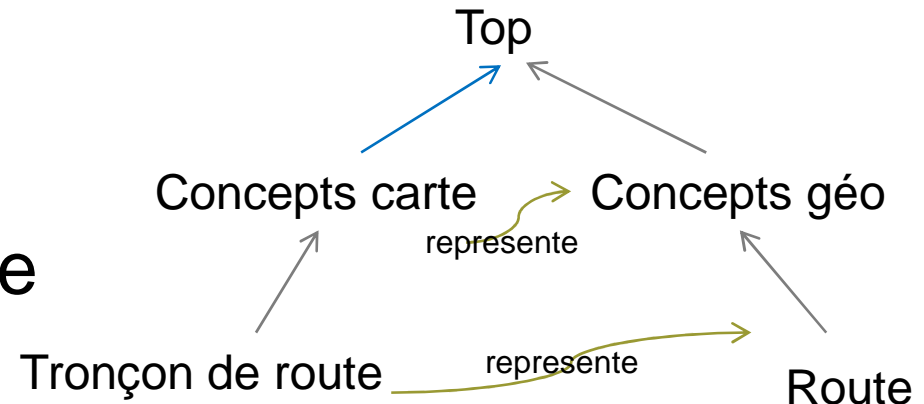
- mots qui ne changent pas la nature des objets géographiques ou qui servent à regrouper des données mais ne sont pas des concepts de l'onto du COGIT

■ Suggestion de traitement

- ❑ Ne pas les utiliser pour l'alignement
- ❑ Séparer 2 parties dans l'ontologie ?

Vers 2 branches dans l'ontologie

- Voir si on arrive à placer les concepts correctement dans une des branches
- Un même terme -> découpé différemment -> 2 concepts
- Refait le modèle de données d'un SIG ?



Projet global lot 1

- **avancement travail IRIT**

- nouvelle version des règles de génération d'ontologie à partir des spécifications / Mouna
- apports de l'analyse TAL des définitions et partie rédigées, premières pistes / Marion
- alignement / validation des ontologies apprises avant fusion / Nathalie

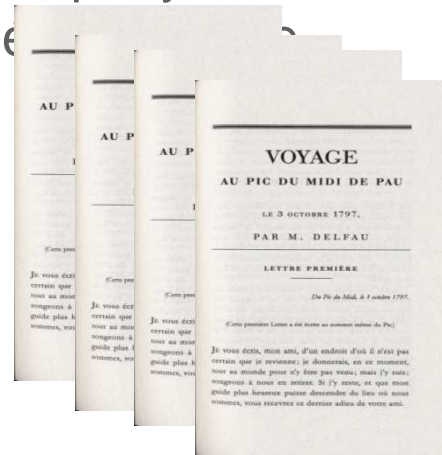
- **avancement travail LIUPPA**

- développements pour manipuler des ontologies depuis LinguaStream / Tien et Christian
- proposition d'enrichissement avec Rameau : point ERic et Marie-Noelle

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 [1 subd. géogr.]

Verdette 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 trouloulytique](#)
- [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](#)

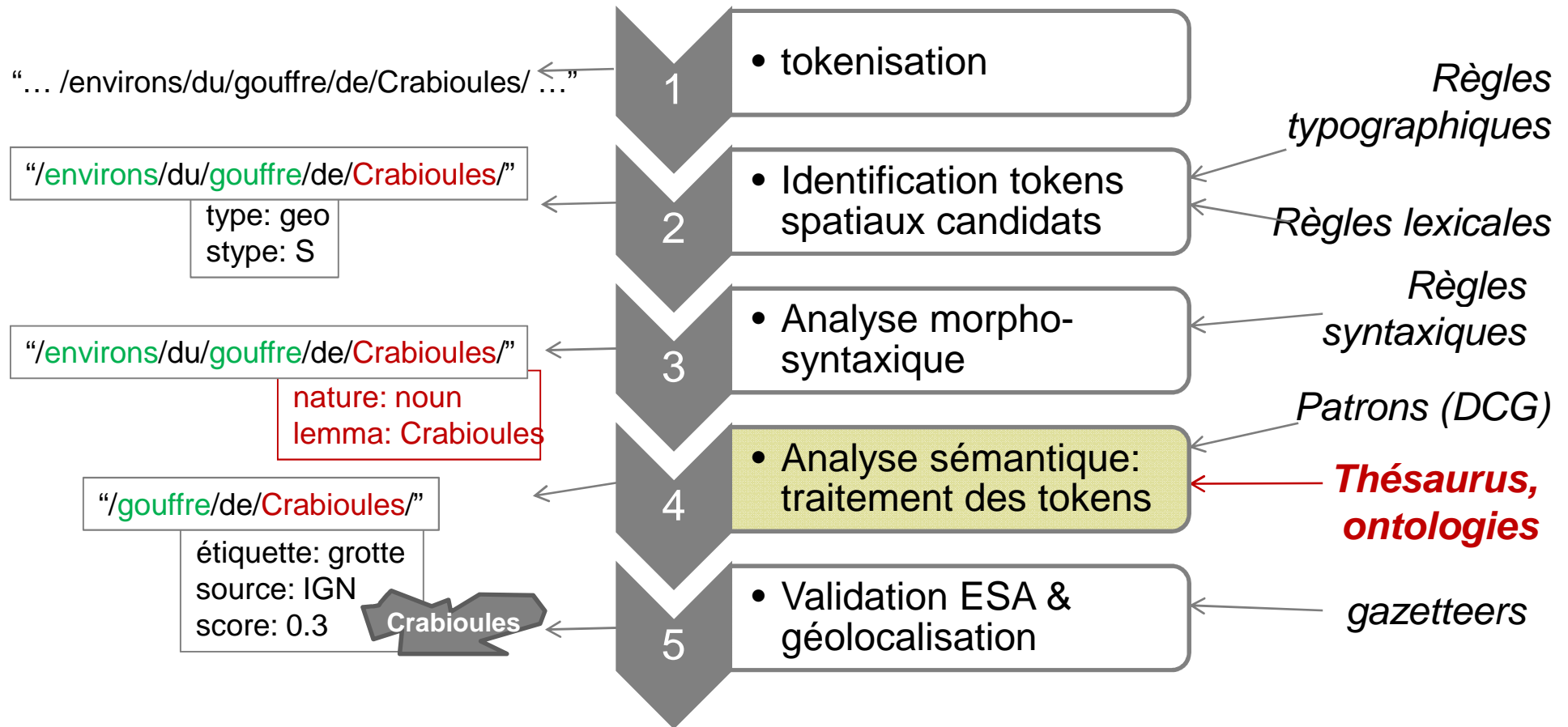
Source(s) : - Grand Larousse universel. - Grand Rubric de la langue française, 2001
 - Dict. de géologie / A. Foucault, J. F. Raoult, 2006. - Dict. de la géographie / P. George, 1671 - Les mots de la géographie : dictionnaire critique / R. Brunet, 1993. - La préhistoire : hist. et dict. / D. Valou, 2004. - Dict. de la préhistoire / A. Leroi-Gourhan, 1994

Equiv. LCSH : Caves

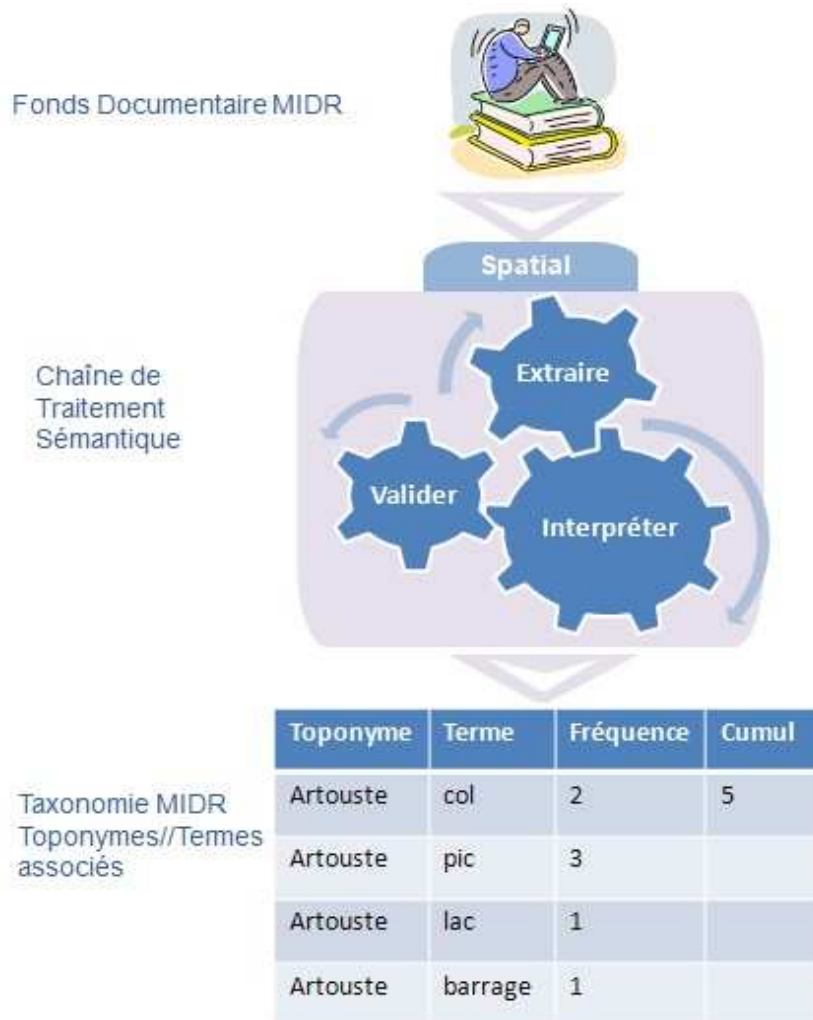


Identification et Extraction des EN spatiales dans les textes grand publics

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



Identification et Extraction des EN spatiales dans les textes grand publics



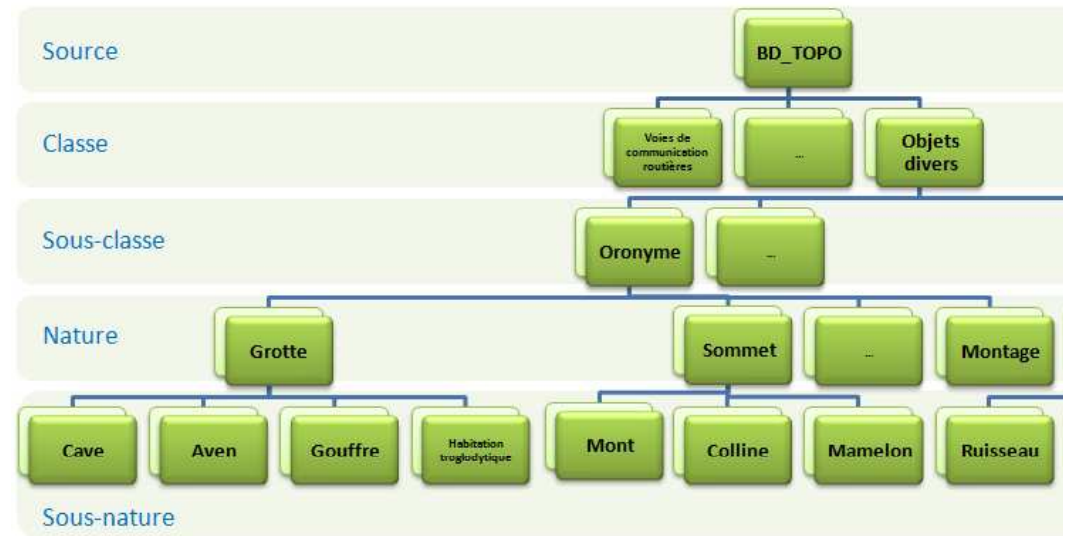
Exemple de résultats:

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:

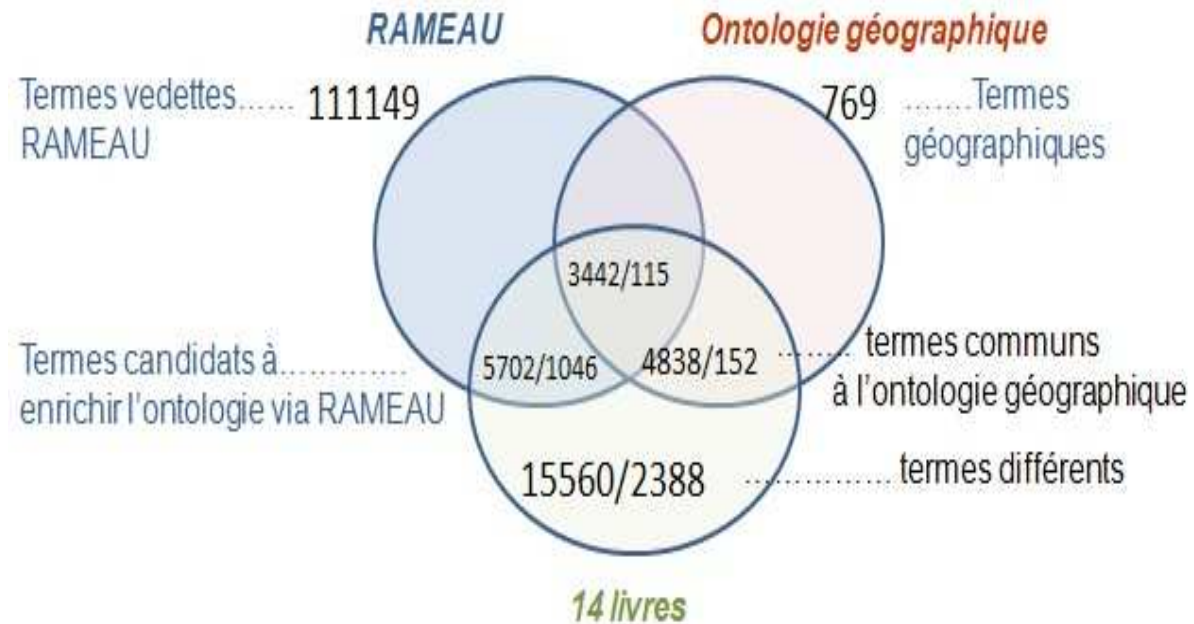
Si « Abîme » existe dans RAMEAU

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.

Réalisé lot 1

