Journée ATALA

Le Web comme ressource pour le TAL

Exploitation de corpus médicaux extraits d'Internet : une expérience

Thierry Delbecque, thd@biomath.jussieu.fr
Pierre Zweigenbaum, pz@biomath.jussieu.fr

Introduction

- La constitution de corpus en particulier à des fins de recherche en TAL doit respecter certaines contraintes en fonction de la tâche visée;

  « *A corpus is a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language.* », *Sinclair, Corpus Typology [EAGLES 1994, p.4]*

- Intérêt des portails professionnels spécialisés sur la toile pour faciliter la construction de ces corpus :
  - Information qualifiée (important dans la recherche d'informations);
  - Possibilité de constituer des corpus dont on maîtrise la thématique;
  - Qualité rédactionnelle;
  - Représentativité des documents (des recherches à l'aide de moteurs de recherche peuvent donner un résultat biaisé du fait des principe de *ranking* adoptés);

Introduction

- L'un de ces portails (CISMeF) a été utilisé pour la constitution du corpus de la tâche médicale de la compétition EQueR 2004 (Evaluation en Questions-Réponses, ELDA/EVALDA);

- Le STIM (AP-HP) a participé à la préparation de ce corpus, et a également participé à la compétition;

- Application : expérience d'évaluation de l'utilisation de ressources terminologiques du domaine dans un cadre de Questions-Réponse; prototype de systèmes de Questions-Réponses.

CISMeF

Catalogue et Index des Sites Médicaux Francophones

http://www.chu-rouen.fr/cismef/

Plan

- L'expérience
  - Le but (utilité de l'étiquetage par des concepts de l'UMLS)
  - Les ressources (l'UMLS, constitution du corpus);
  - La démarche;
  - Les résultats;
- Post-mortem
  - Difficultés rencontrées;
  - Critiques;
- Conclusion

Expérience

Question: utilisabilité d'une ressource terminologique médicale (partie francophone de l'UMLS) comme source d'entités nommées médicales.

- Extraction d'un corpus de textes médicaux à partir de CISMeF (celui d'EQueR);
- Indexation de ce corpus par des concepts de l'UMLS, et en s'appuyant sur l'UMLS francophone;
- Évaluation de la qualité de globale de l'étiquetage ainsi obtenu;
- Utilisation pour une cartographie des documents;
- Évaluation de l'apport dans un contexte de Question-Réponses;

L'UMLS

- Vaste ressource terminologique médicale (Unified Medical Language System);
- Proposée et maintenue par la NLM (National Library of Medicine);
- 3 composants:
  - Le Metathesaurus;
  - Le réseaux sémantique;
  - SPECIALIST Lexicon;

L'UMLS : le Metathesaurus

- Le Metathesaurus organise des termes (*maladie de Crohn*) autour de concepts (CUI); les termes attachés à un même CUI sont synonymes.
- Produit de la fusion de plusieurs terminologies « sources » (MeSH, MedDRA, ...); relations entre concepts, héritées des sources ou originales;
- Multilingue;

Adrenal gland diseases	MeSH	D000307
Adrenal disorder	AOD	0000005418
Disorder of adrenal gland	Read	C15z.
Diseases of the adrenal glands	SNOMED	DB-70000

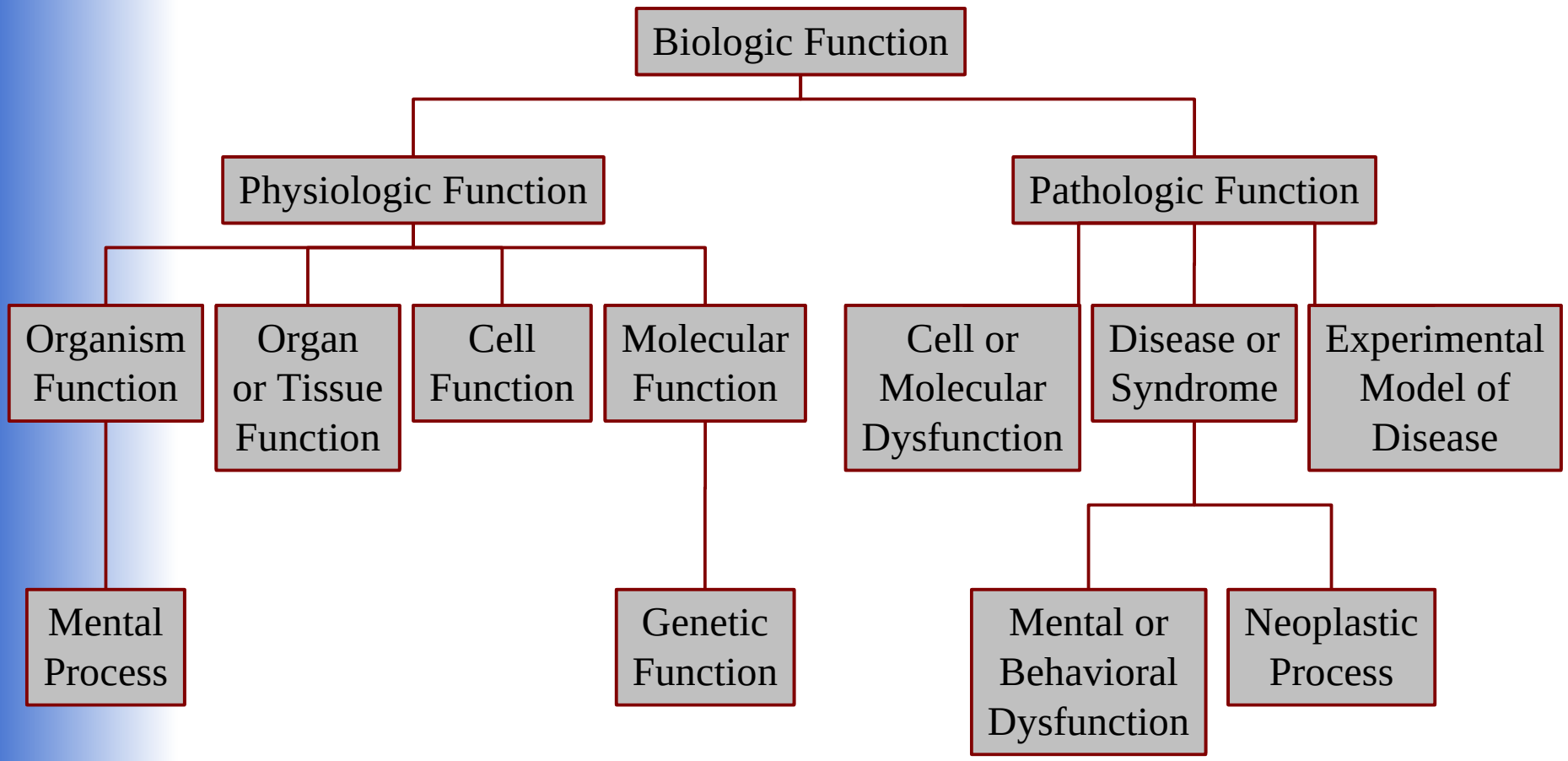
C0001621

Adrenal Gland Diseases

L'UMLS : le réseau sémantique

- A chaque concept du Metathesaurus est attaché un type sémantique;
- 134 types sémantiques dans la version 2002-AA
  - Structure arborescente (is-a);
  - 2 hierarchies
    - Entités (Entities)
      - Objets physiques (Physical Objects)
      - Entités conceptuelles (Conceptual Entities)
    - Evénements (Events)
      - Activités (Activities)
      - Phénomènes ou processus (Phenomenon or Process)

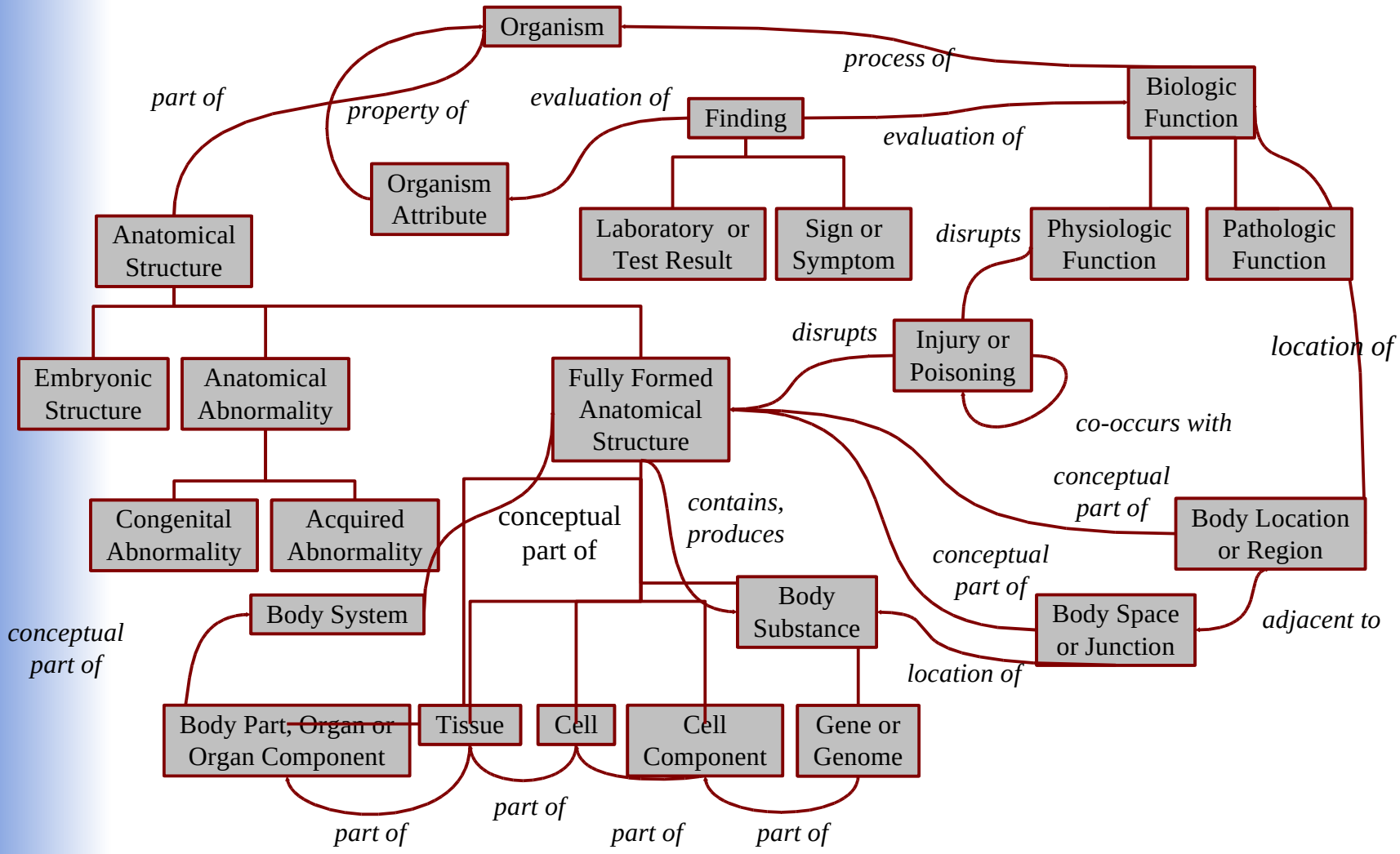
L'UMLS : les types sémantiques (extrait)



L'UMLS : le réseau sémantique

- L'ensemble des types sémantiques est structuré par 54 relations sémantiques (2002-AA):
  - relation hierarchique (isa = is a kind of)
    - entre types
      - Animal *isa* Organism
      - Enzyme *isa* Biologically Active Substance
    - entre relations
      - treats *isa* affects
  - non-hierarchiques (associatives)
    - Sign or Symptom *diagnoses* Pathologic Function
    - Pharmacologic Substance *treats* Pathologic Function

Réseau sémantique, relations associatives



L'UMLS

Indices pour l'extraction d'information médicale:

- certains types sémantiques (*sign or symptom*, ...) : types éventuels d'*entités nommées* propres à la médecine;
- certaines relations sémantiques, induites sur la base de cooccurrences de types sémantiques dans une fenêtre de texte (*diagnoses*, ...) ...

L'UMLS

Repérage de relations sémantiques

« *l'interféron a une place dans le traitement de certains cancers.* »

UMLS Knowledge Base

interféron : [pharmacological substance]
cancer: [disease or syndrome]
[disease or syndrome] isa [pathologic function]
[pharmacologic substance] treats [pathologic function]

Étiquetage de la phrase avec la relation *treats*.

L'UMLS

Représentation des différentes langues présentes dans la version 2002-AA : handicap du français.

Langue	Nb de chaînes	% / anglais	Nb concepts	Nb chaînes / concept
HEB	485	0,03	472	1,03
BAQ	695	0,04	695	1
HUN	718	0,05	718	1
NOR	722	0,05	722	1
SWE	723	0,05	723	1
DAN	723	0,05	723	1
FIN	21118	1,33	20966	1,01
ITA	23592	1,48	22698	1,04
FRE	34780	2,18	23966	1,45
DUT	36595	2,3	18375	1,99
RUS	42389	2,66	20593	2,06
POR	45806	2,88	29751	1,54
SPA	51560	3,24	37319	1,38
GER	68061	4,27	43409	1,57
ENG	1592203	100	776940	2,05

Le corpus

Documents extraits des sites suivants choisis parmi les plus gros sites indexés par CISMeF afin de simplifier l'obtention des droits :

- www.fnclcc.(fr,com) [CANCER];
- www.ladocumentationfrancaise.fr [DOCFRA];
- afssaps.sante.fr [AFSSAPS];
- www.anaes.fr [ANAES];
- www.orpha.net [ORPHA];
- www.senat.fr [SENAT];
- www.chu-rouen.fr [CHUROUEN];
- www.univ-rouen.fr [UROUEN];
- www.hc-sc.gc.ca [CANADA];

==> 5 583 documents, 18 988 705 mots.

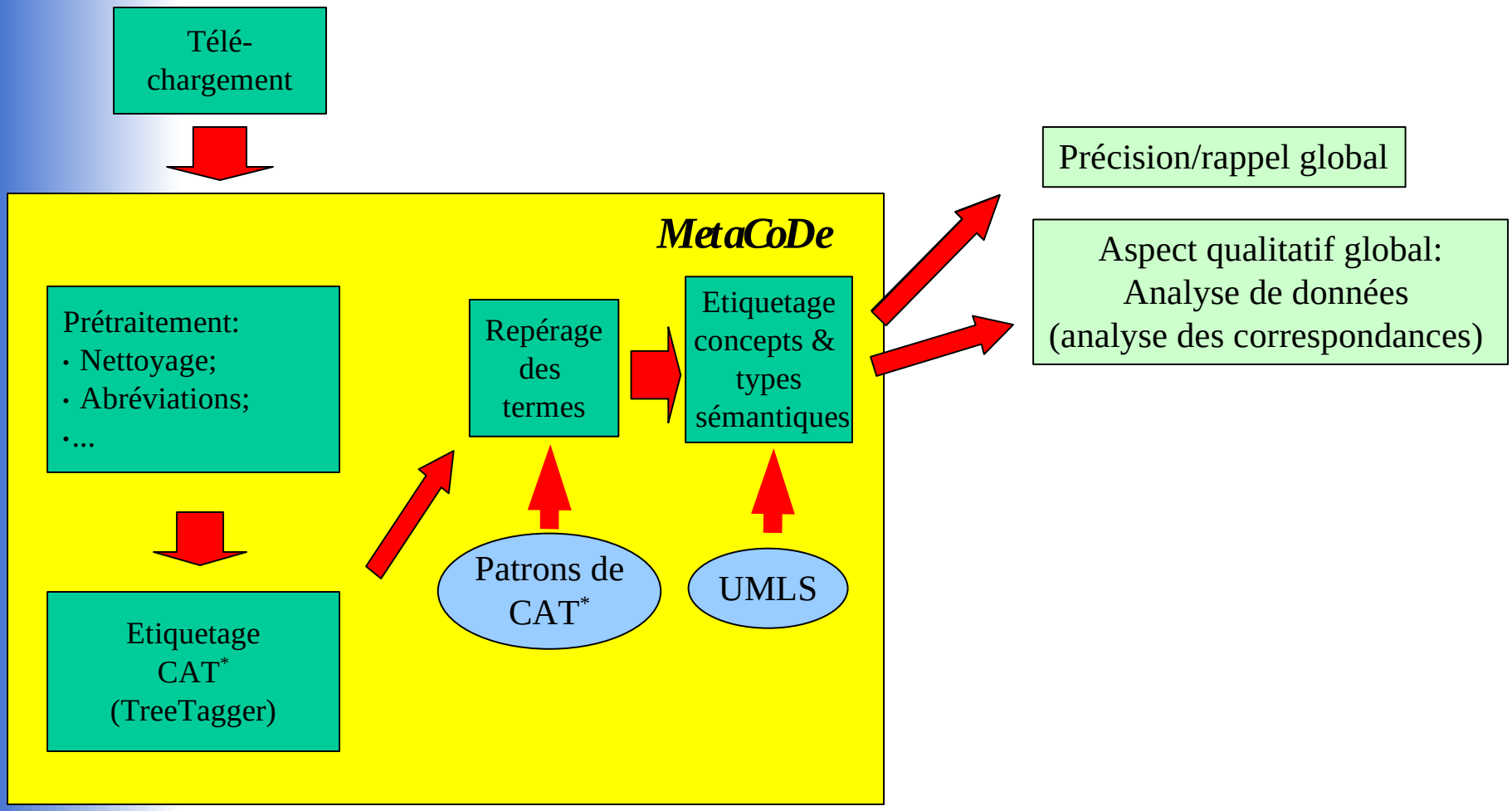
Plan

- L'expérience
  - Le but (utilité de l'étiquetage par des concepts de l'UMLS)
  - Les ressources (l'UMLS, constitution du corpus);
  - La démarche;
  - Les résultats;
- Post-mortem
  - Difficultés rencontrées;
  - Critiques;
  - Conclusion.

L'étiquetage

- Par une plate forme logicielle *ad hoc* (MetaCoDe, PERL, Linux);
- Indexation:
  - par les concepts UMLS (CUI),
  - puis par les types sémantiques,
  - enfin par les relations sémantiques postulées.

Étiquetage par les concepts

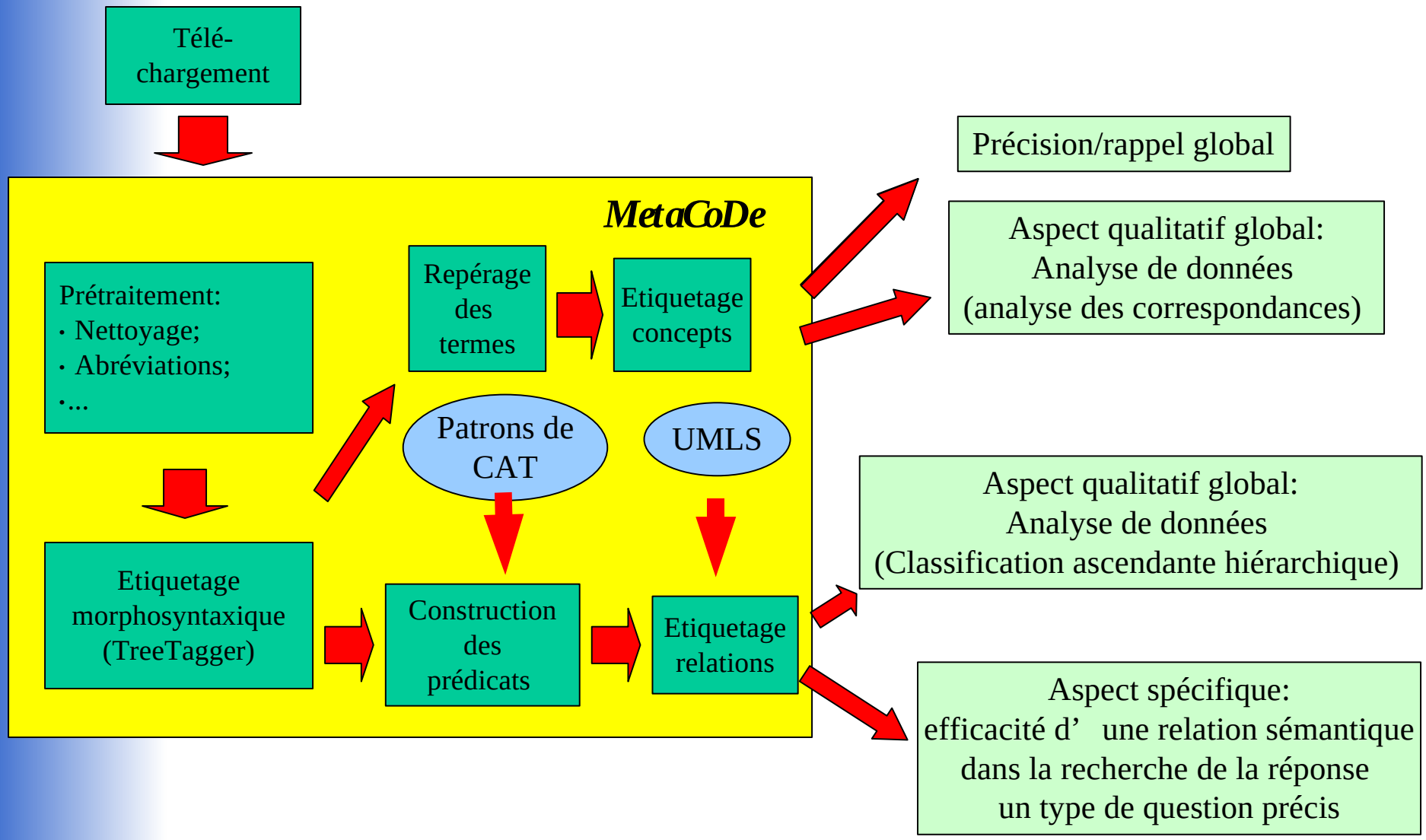


* CAT: Partie du discours

Étiquetage par les concepts

```
30      1       La
31      3       [maladie][de][Crohn]    5:S0229899.C0010346(T047)
34      1       (
35      1       [MC]
36      1       )
37      1       et
38      1       la
39      2       [rectocolite][hémorragique]     3:S0229570.C0009324(T047)
41      1       (
42      1       RCH
43      1       )
44      1       sont
45      1       les
46      1       2
47      3       [causes cause][de][MICI]
50      1       (
51      4       [maladies maladie][inflammatoires inflammatoire]
                [cryptogénétiques cryptogénétique][intestinales intestinal]
                11:S0241600.C0021390(T047)
55      1       )
56      1       .
(4 101 404 termes, 391 966 distincts)
```

Étiquetage par les relations sémantiques



Étiquetage par les relations sémantiques

```
<form id="57">
        <predicat>
                être
        </predicat>
        <moderator>
                souvent
        </moderator>
        <subject from="1339" to="1343">
                <SUI>1:S0226498</SUI>
                <CUI>C0000726</CUI>
                <TUI>T029</TUI>
                L examen de l abdomen
        </subject>
        <arguments>
                <item from="1346" to="1361">
                        <SUI>1:S0234187</SUI>
                        <CUI>C0022828</CUI>
                        <TUI>T007</TUI>
                        normal ( formes basses ) ou montre
                        parfois une sensibilité de la fosse iliaque gauche .
                </item>
        </arguments>
</form>
(495 051 prédicats)
```

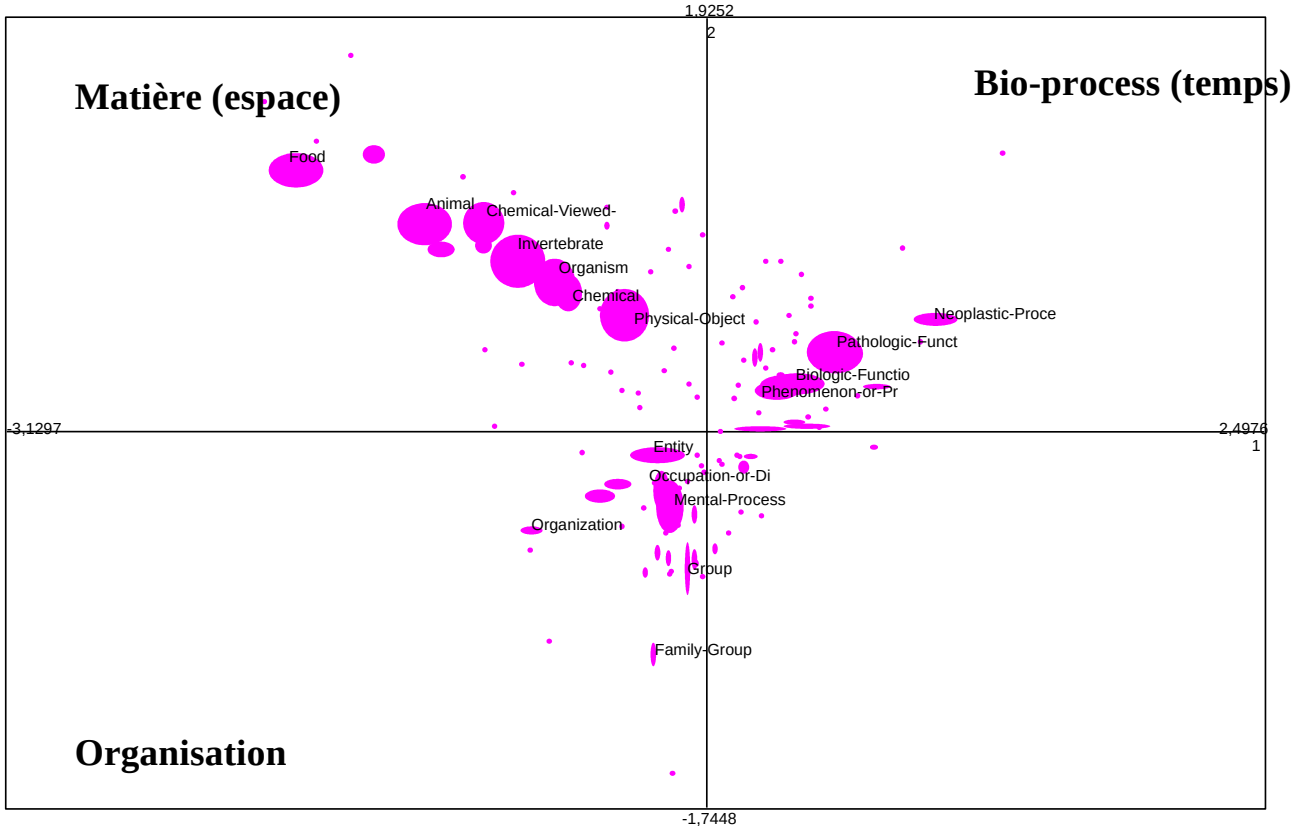
Plan

- L'expérience
  - Le but (utilité de l'étiquetage par des concepts de l'UMLS)
  - Les ressources (l'UMLS, constitution du corpus);
  - La démarche;
  - Les résultats;
- Post-mortem
  - Difficultés rencontrées;
  - Critiques;
  - Conclusion.

Cartographie

Projection sur la structure du corpus étiqueté

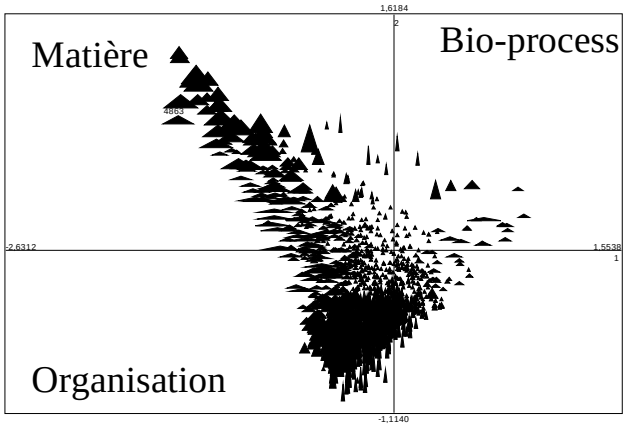
Analyse des correspondances des types sémantiques x documents.
Projection des types sur le plan 1-2



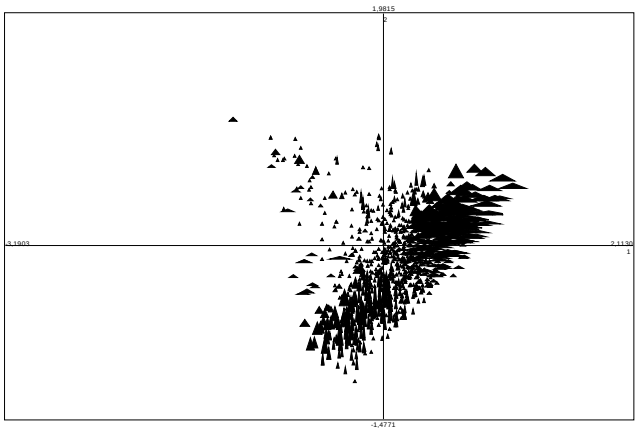
Cartographie

Projection sur la structure du corpus étiqueté

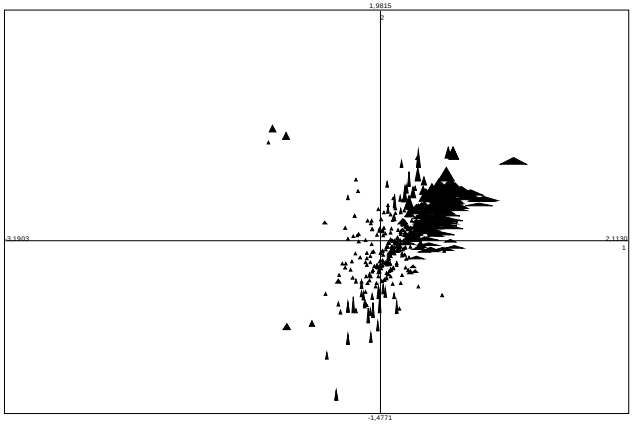
Analyse des correspondances des types sémantiques x documents.
Projection des documents sur le plan 1-2



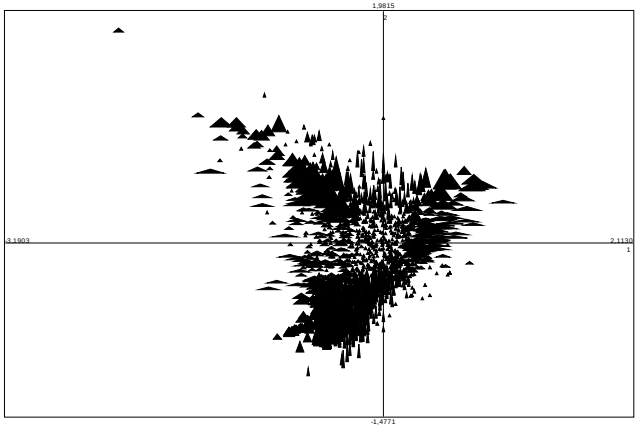
[SENAT][DOCFRA]



[ANAES][AFSSAPS]



[ORPHA][CANCER][UROUEN]



[CHUROUEN][CANADA]

Étude dans un contexte de QR

Question: quelle est la précision obtenue par l'indice donné par la présence d'un étiquetage par *treats* pour des questions du type:

« quel traitement pour tel signe/pathologie »

Vrai positif:

« *Chez les patients au stade de cirrhose, l'utilité du traitement par interféron alpha est démontrée* »

Faux positif:

« *L'ostéoporose peut être secondaire à un état pathologique ou à la prise de certains médicaments* »

Étude dans un contexte de QR

Mesures sur la base d'un échantillon par source.

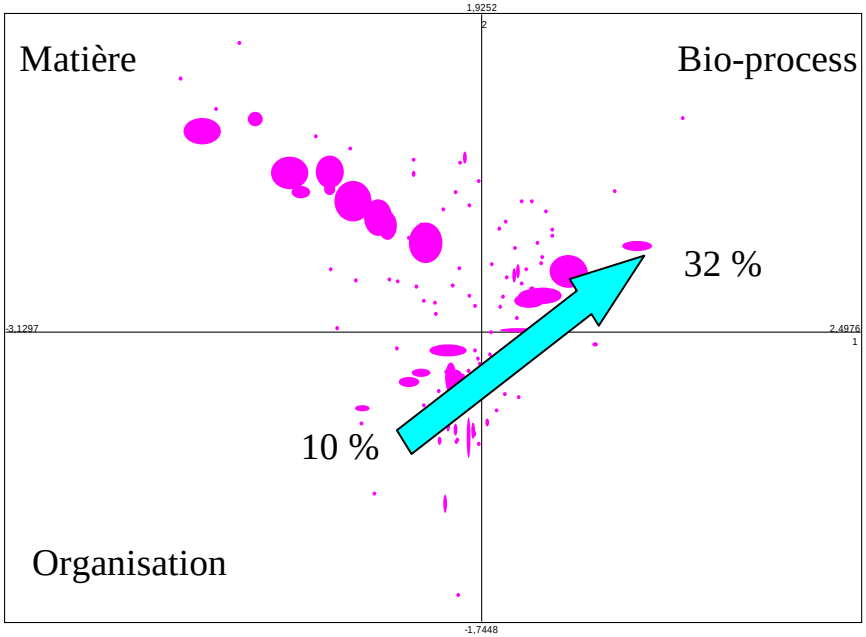
	Nombre de phrases	Nombre de phrases avec l'étiquette *treats*	Echantillon	% vrais positifs	Ecart-type
SENAT	199372	1265 (0,6 %)	200	10	2,1
CANADA	90986	2743 (3,0 %)	200	16	2,6
CHUROUEN	10232	230 (2,2 %)	200	19	2,8
UROUEN	14799	621 (4,2 %)	200	20	2,8
AFSSAPS	5187	202 (3,9 %)	202	20	0
ANAES	125659	4174 (3,3 %)	200	22	2,9
ORPHA	1460	25 (1,7 %)	25	27	0
CANCER	47356	2325 (4,9 %)	200	32	3,3

- L'efficacité de l'étiquette *treat* croit suivant un axe « sources généralistes » => « sources spécialisées »;
- Phénomène dû à des différences de langue, et non à des différences de densité d'étiquetage.

Étude dans un contexte de QR

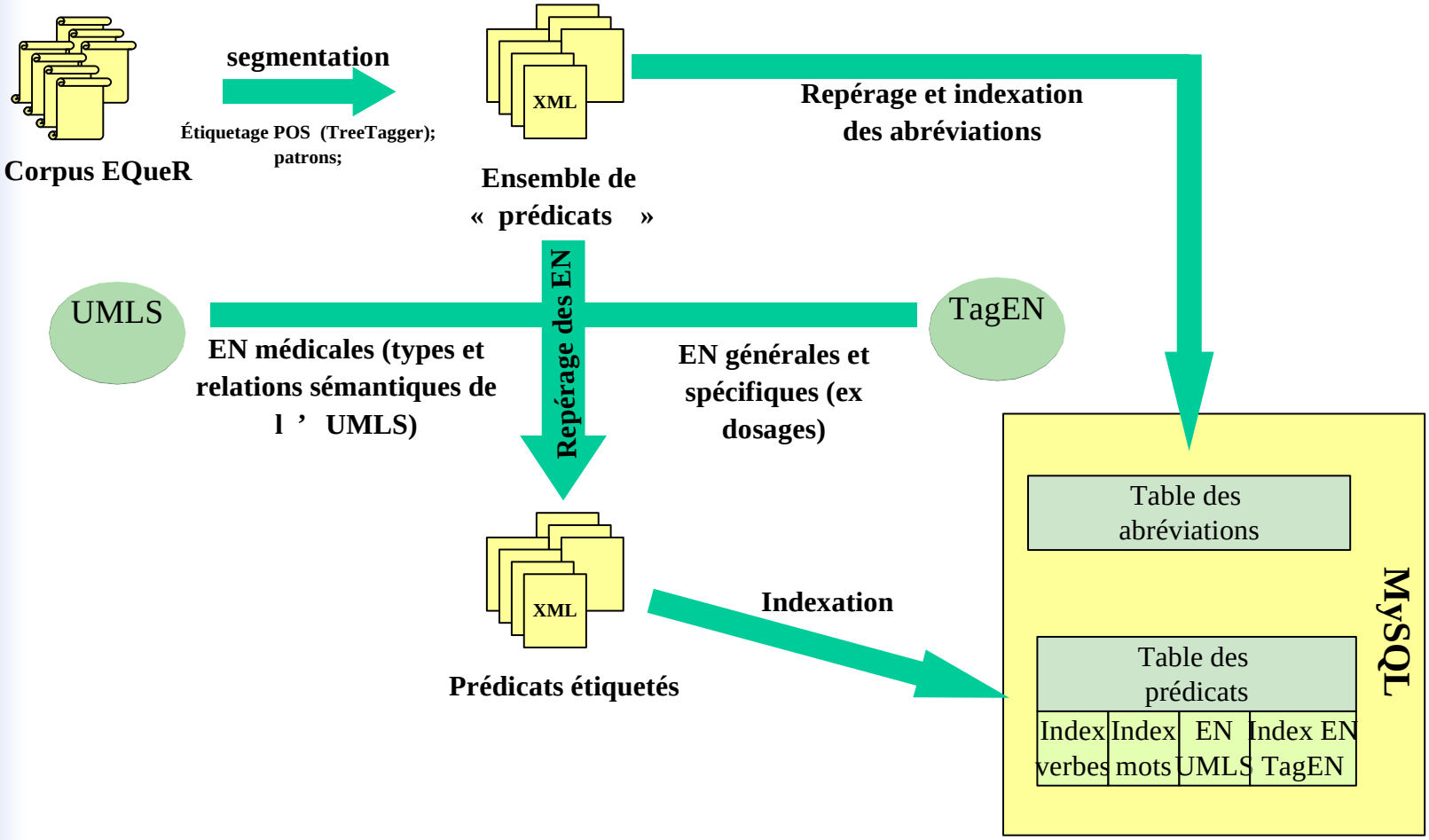
Projection sur la structure du corpus étiqueté

Analyse des correspondances des types sémantiques x documents.
Efficacité de l'étiquette *treat*.



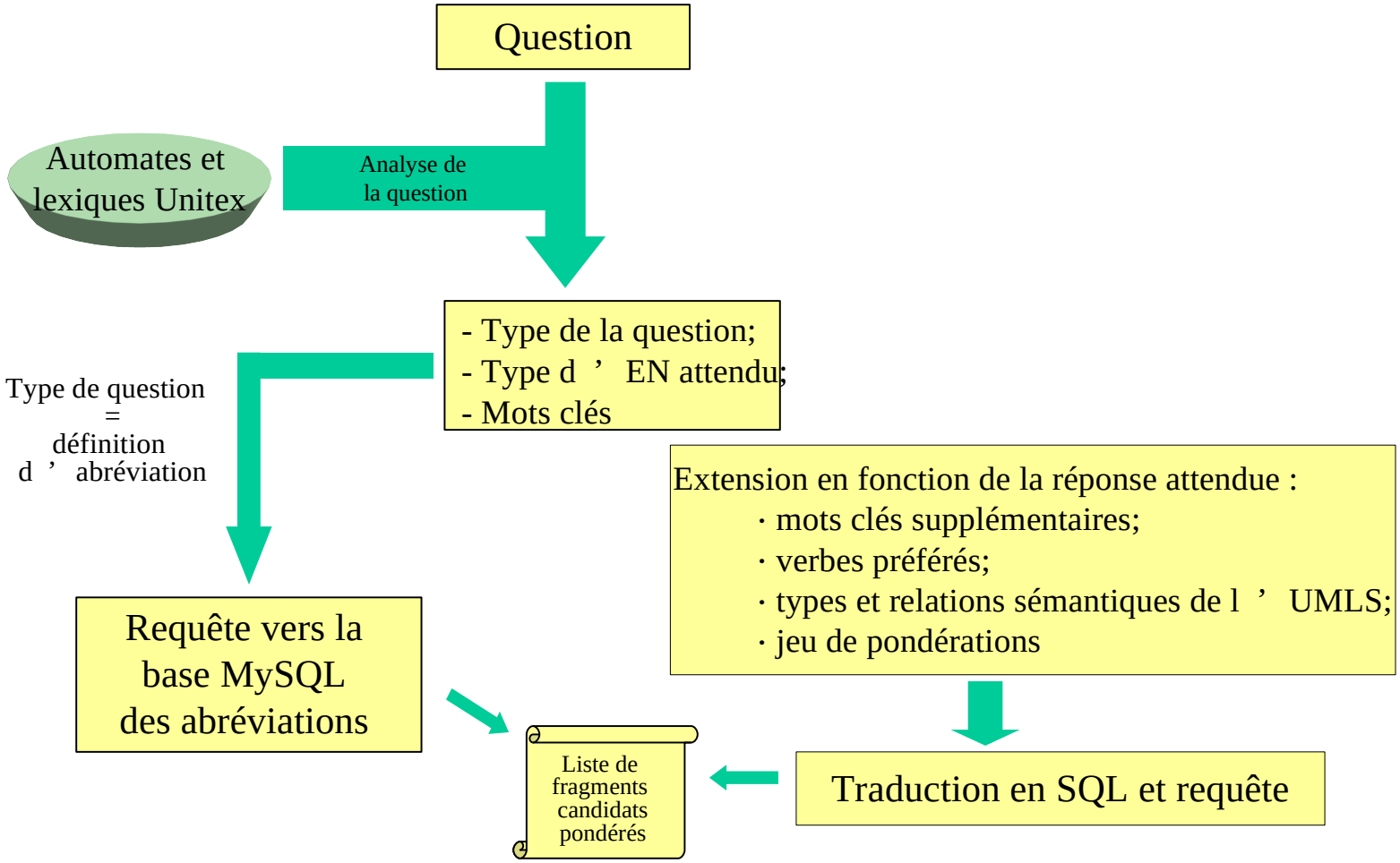
Application : prototype EQueR

Pré-traitement du corpus



Application : prototype EQueR

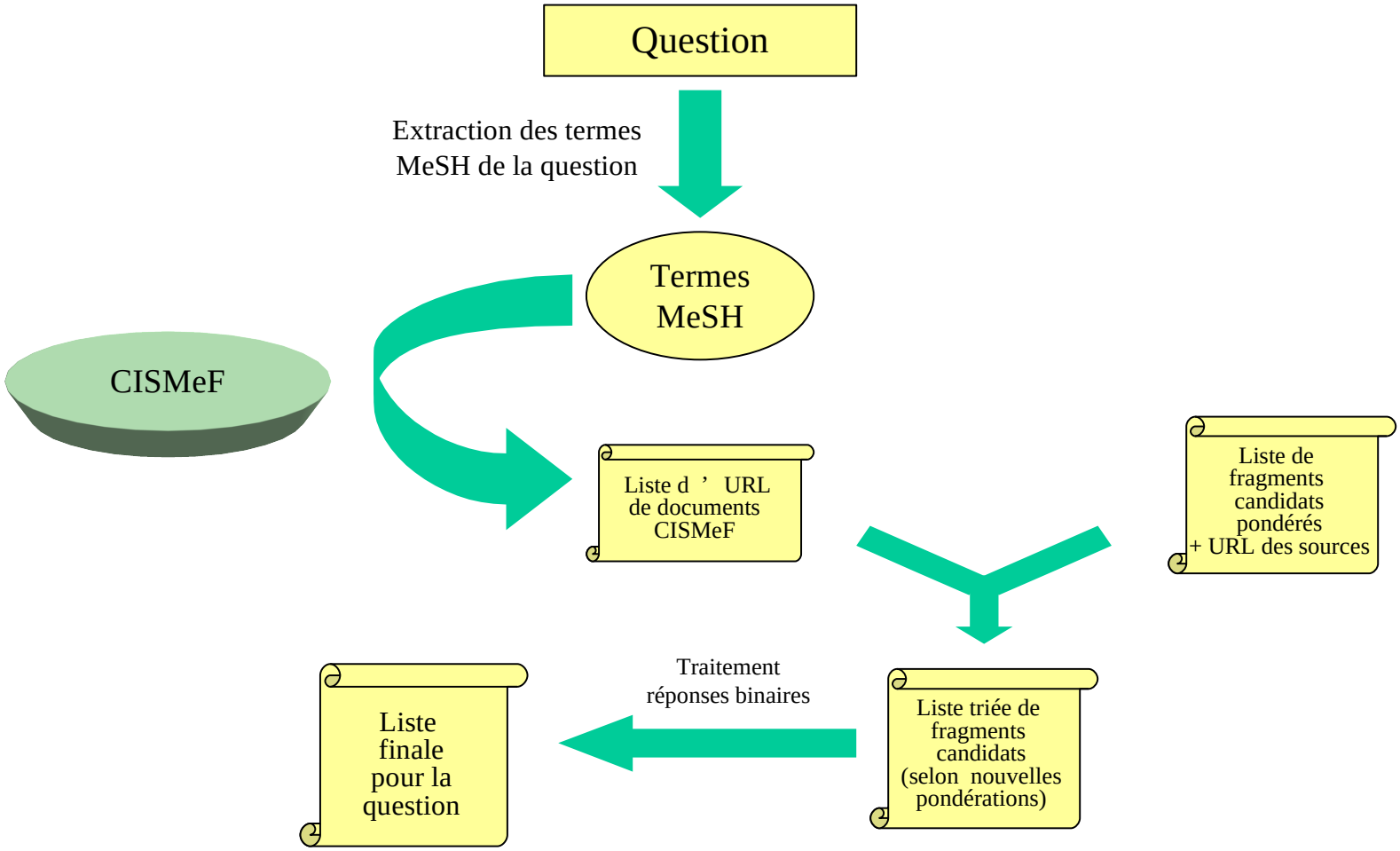
Déroulement d’ une requête



Application : prototype EQueR

Traitement final des fragments candidats:

(Thématique des documents.)



Conclusion de l'expérience

- L'étude nous a permis de mesurer la qualité de l'étiquetage d'un corpus médical par l'UMLS;
- pour cela, la confrontation de la cartographie donnée a priori et celle fournie par l'étiquetage a été précieuse;
- l'étude nous a également apporté des informations sur la manière d'utiliser les concepts de l'UMLS projetés dans le corpus (les étiquettes) comme types d'entités nommées;
- prototype d'un système de QR exploitant un corpus préparé (spécialement annoté et indexé).

Post mortem : bénéfice du portail

Bénéfice du recours à CISMeF pour la constitution du corpus:

- Richesse et variété du contenu, pour un domaine donné;
- Pertinence des sources indexées (thèmes et qualité du contenu);
- Cartographie a priori disponible:
  - Soit par le domaine auquel appartient le document (utilisé dans l'expérience);
  - soit par l'indexation MeSH effectuée par l'équipe CISMeF);
  - soit par les métadonnées RDF/RDFS attachées aux documents par CISMeF;

Post mortem : difficultés

Difficultés liées aux spécificités de la source (WEB)

La mise à disposition des documents souvent en PDF rend l'extraction du texte difficile:

- agglutination: ... *leseultraitementàviséecurative* ...
- dislocation: ... *R a d i o g r a p h i e t h o r a c i q u e* ...
- beaucoup de bruit lié aux tableaux, images, etc ...
- perte de la structuration du texte (titres, sections, etc.), problèmes de segmentation ...
- ...

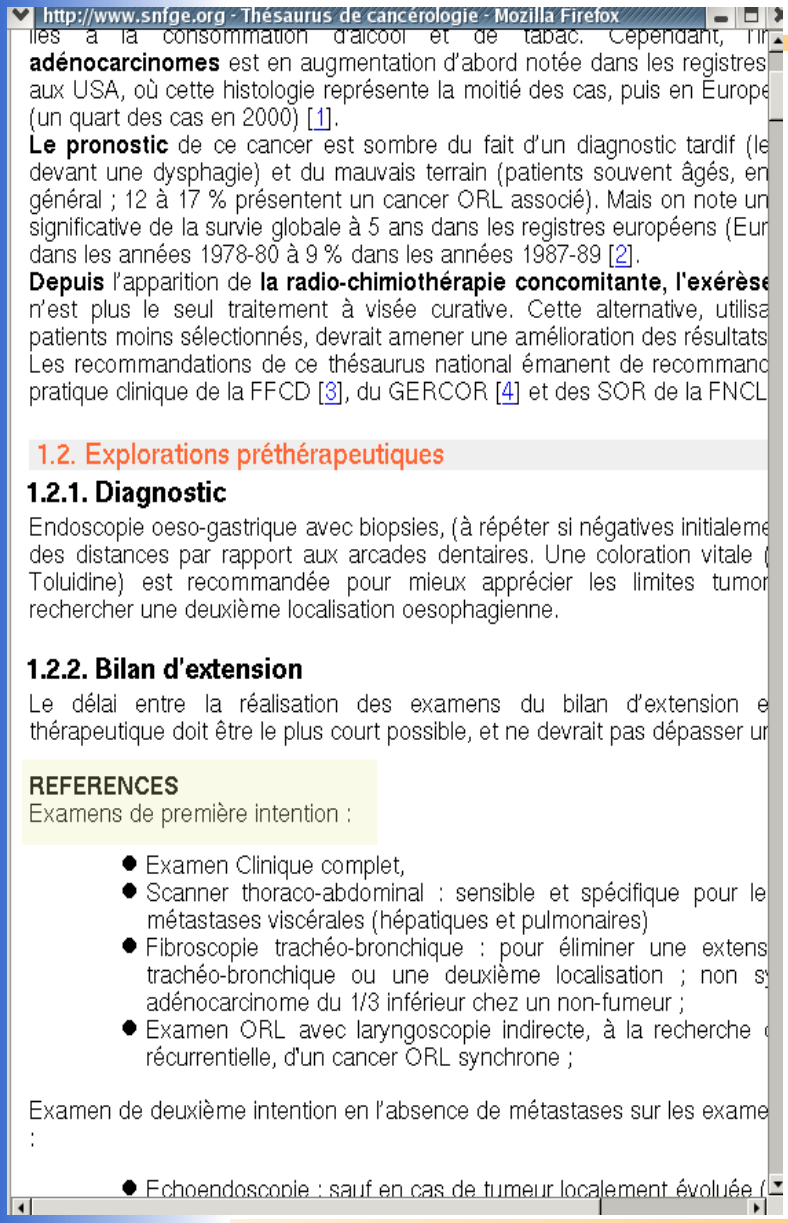
Post mortem : difficultés

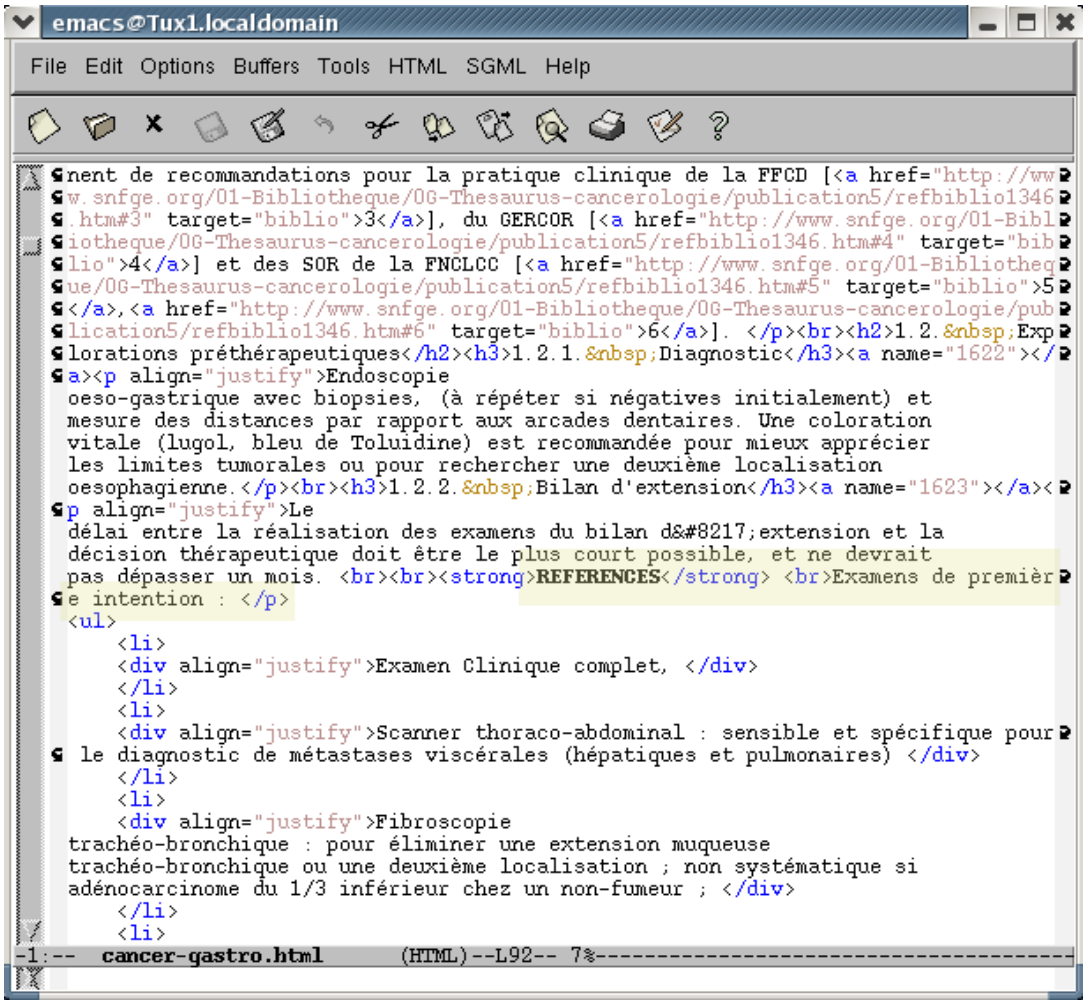
Difficultés liées aux spécificités de la source (WEB)

L'extraction du texte à partir des fichiers HTML n'est pas complètement satisfaisante:

- manque de principes généraux d'utilisation des balises HTML, qui rend le repérage de l'organisation du texte difficile (paradoxal aujourd'hui ...);
- segmentation (phrases, paragraphes) à retrouver;
- extraction parfois partielle (par exemple lors de recours aux tableaux pour la mise en forme); utilisation insuffisante de CSS;
- ...

Post mortem : difficultés





Post mortem

Particularités liées aux spécificités d'un domaine spécialisé :

une terminologie particulière ...

- Termes peu courants, non reconnus et mal étiquetés par les étiqueteurs morphosyntaxiques;
- Noms de substances, de produits, de personnes, etc ...;
- Abréviations;
- Sens particuliers attachés aux termes (indiquer, sombre, étrier, frein, ...)

Post mortem

Particularités liées aux spécificités d'un domaine spécialisé :

... mais des ressources spécifiques disponibles:

- thésaurus (MeSH, SNOMED, CIM10, ...), ontologies, ... (cependant lacunaires en français);
- connaissances morphologiques (ex : morphologie flexionnelle et dérivationnelle de Fiammeta Namer);

Post mortem : critiques de l'expérience

Différentes étapes de traitement et d'annotation ont été enchaînées :

- extraction du texte; nettoyage;
- segmentation en mots et étiquetage TreeTagger;
- repérage de termes; étiquetage par l'UMLS;
- repérage des prédicats;

mais :

- les produits de chaque étape sont « indépendants » (pas de références entre eux);
- les résultats ne sont généralement pas représentés sous une forme normalisée: rarement sous forme XML, ou alors selon un « schéma » propriétaire;
- les ressources créées sont difficilement réutilisables par des tiers;
- ...

Conclusion

Utilité de la disponibilité d'un corpus de textes médicaux annotés:

- Composés de documents **issus du Web** (référencés par CISMeF, par exemple);
- Annotations selon des méthodes proches des standards (TEI/XCES : documents primaires et annotations *stand-off*, métadonnées (header TEI), ...);
- Annotations disponibles comme des ressources en soi;
- Étiquetage morphosyntaxique des termes étendus aux termes spécialisés;

Conclusion

Utilité de la disponibilité d'un corpus de textes médicaux annotés:

- Repérages de concepts médicaux issus de thésaurus (avec l'avantage de pouvoir considérer ces concepts comme des « entités universelles » (MATE*));

- Traitement spécifique des balises HTML, lorsque le document est au format HTML; faut-il un nouvel élément, par exemple ? :

```
<BALISE><![CDATA[<html>]]></BALISE>
<BALISE><![CDATA[<head>]]></BALISE>
<BALISE><![CDATA[<title>}}></BALISE>
<p>Thésaurus de cancérologie</p>
<BALISE><![CDATA[</title>]]></BALISE>
```

* Multilevel Annotation Tool Engineering

Conclusion

Utilité de la disponibilité d'un corpus de textes médicaux annotés:

- Une branche « Web médical » intégrée à la FReeBank ? (corpus « de recherche »);

- Avantage du maintien par un portail comme CISMeF de telles sources annotées, pour la recherche d'informations médicales sur Internet.

Conclusion

Merci