

**TALN 2006- Leuven, 10-13 avril 2006**

**RNTL  ATONANT**

**Aide à l'enrichissement semi-automatique  
d'ontologies**

EADS – Sylvie BRUNESSAUX

CEA – Olivier MESNARD

LIP6 – Patrick GALLINARI

LIPN – Sylvie SZULMAN

LaRIA – Gilles KASSEL

INSA – Jean-Pierre PECUCHET

CHU Rouen CISMef – Badisse DAHAMNA

## Objectif du projet ATONANT

**« Prototyper une plate-forme d'aide à l'enrichissement semi-automatique d'ontologies »**

- Durée du projet :  
01/09/2003 – 30/11/2005
- Site Web : <http://atonant.insa-rouen.fr/>

## Synthèse des résultats obtenus

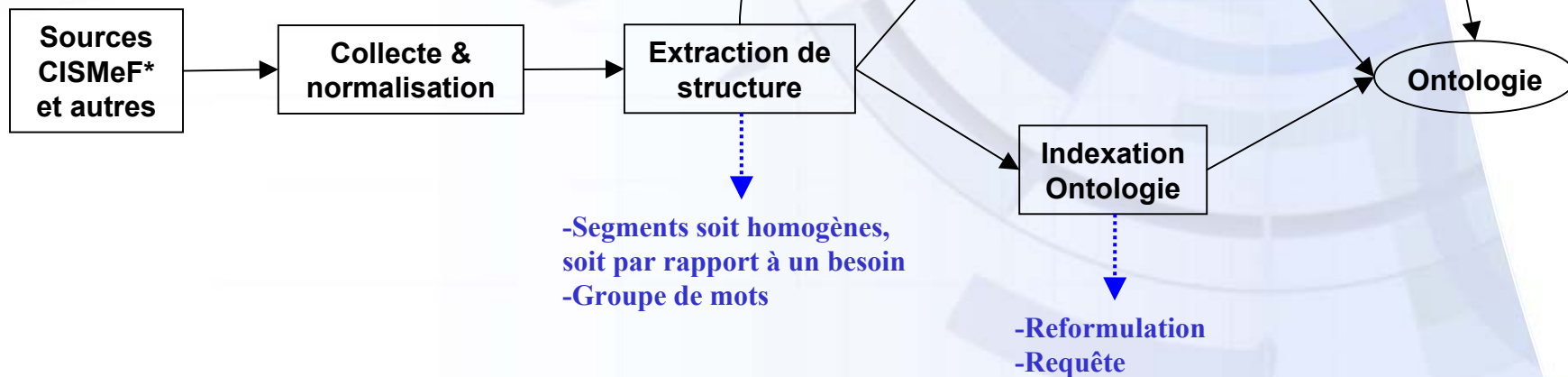
- Un outil d'aide à la recherche d'information sur le Web
- Un outil de collecte et de formatage de données
- Une méthode de conception et un outil d'édition d'ontologies
- Un outil de production de hiérarchie de termes
- Un outil de création de hiérarchie spécialisation/généralisation

# Contexte

## Les objectifs techniques du projet

- Hiérarchie de thèmes basés sur cooccurrence
- Hiérarchie sur l'usage des termes
- Intersection de clusters de termes + visualisation

Hiérarchie de « concepts »  
(définie par un groupe de termes)



\* CISMéF: Catalogue et Index des Sites Médicaux francophones

## Contexte

### Challenge scientifique

- Méthode d'aide à la conception d'ontologies formelles à partir de textes
- Construction automatique d'ontologie

## Contexte

### Challenge technologique

- Construction de ressources métier pour des applications de veille technologique
- Intégration d'outils pour la méthode
- Apprentissage automatique pour l'aide à l'enrichissement d'ontologies

## Analyse du domaine

### Spécificités de l'application à CISMef

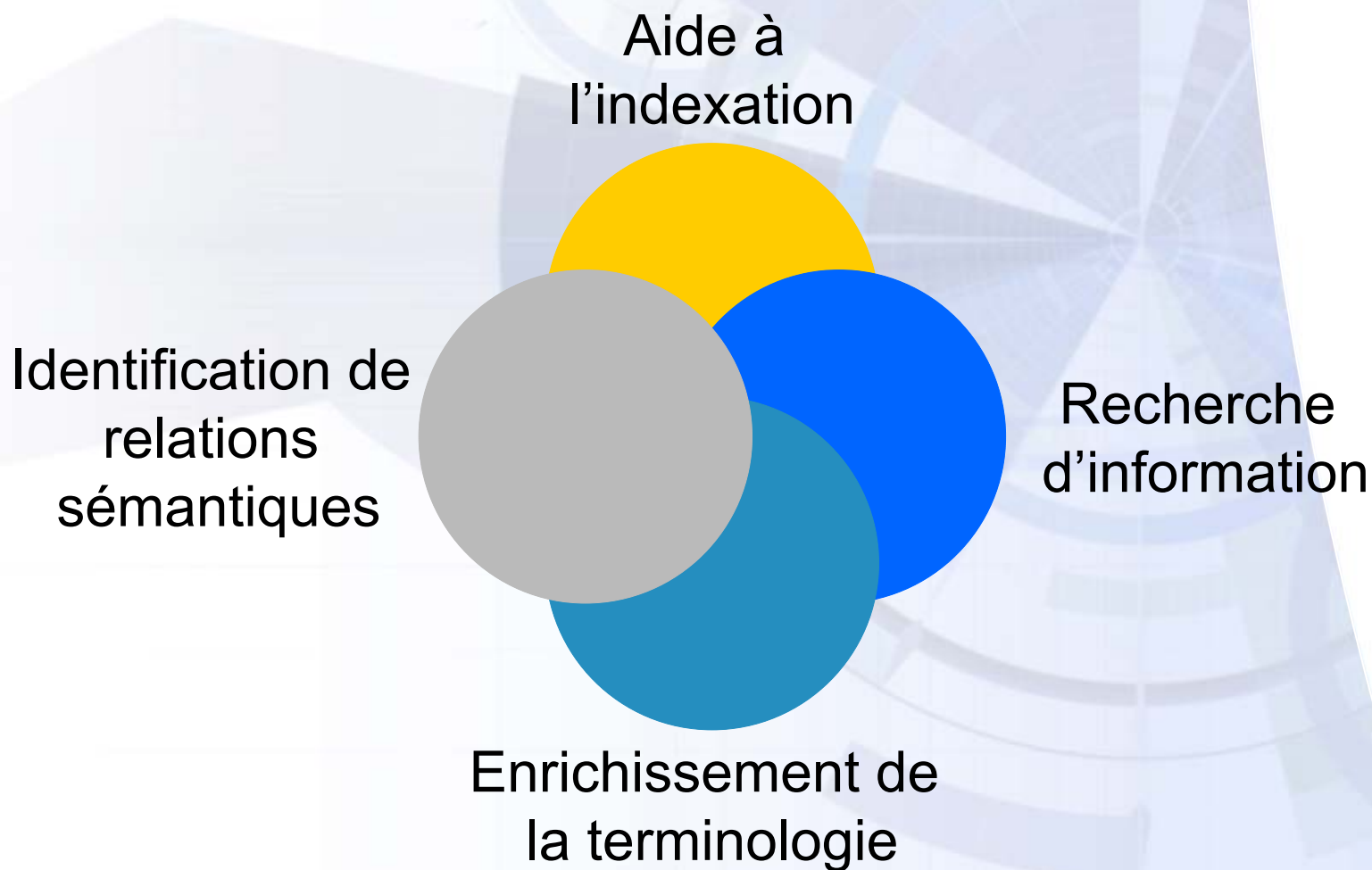
- Application unique au domaine médical
- > 13 000 ressources indexées
- > 10300 termes MeSH (Medical Subject Headings)
- Collection de données
  - Corpus et ressources ontologiques du CISMef
  - Données Thésaurus & CISMef au format XML/OWL
  - Lexique

**CISMef**  
Catalogue et Index des Sites  
Médicaux Francophones



# Analyse du domaine

## Les besoins du CISMeF

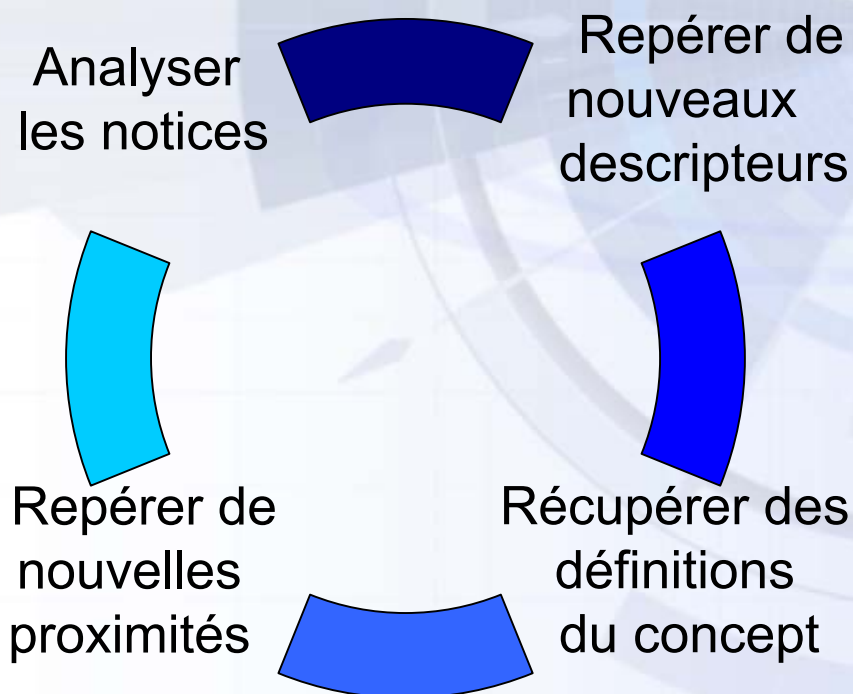




## Avancées techniques

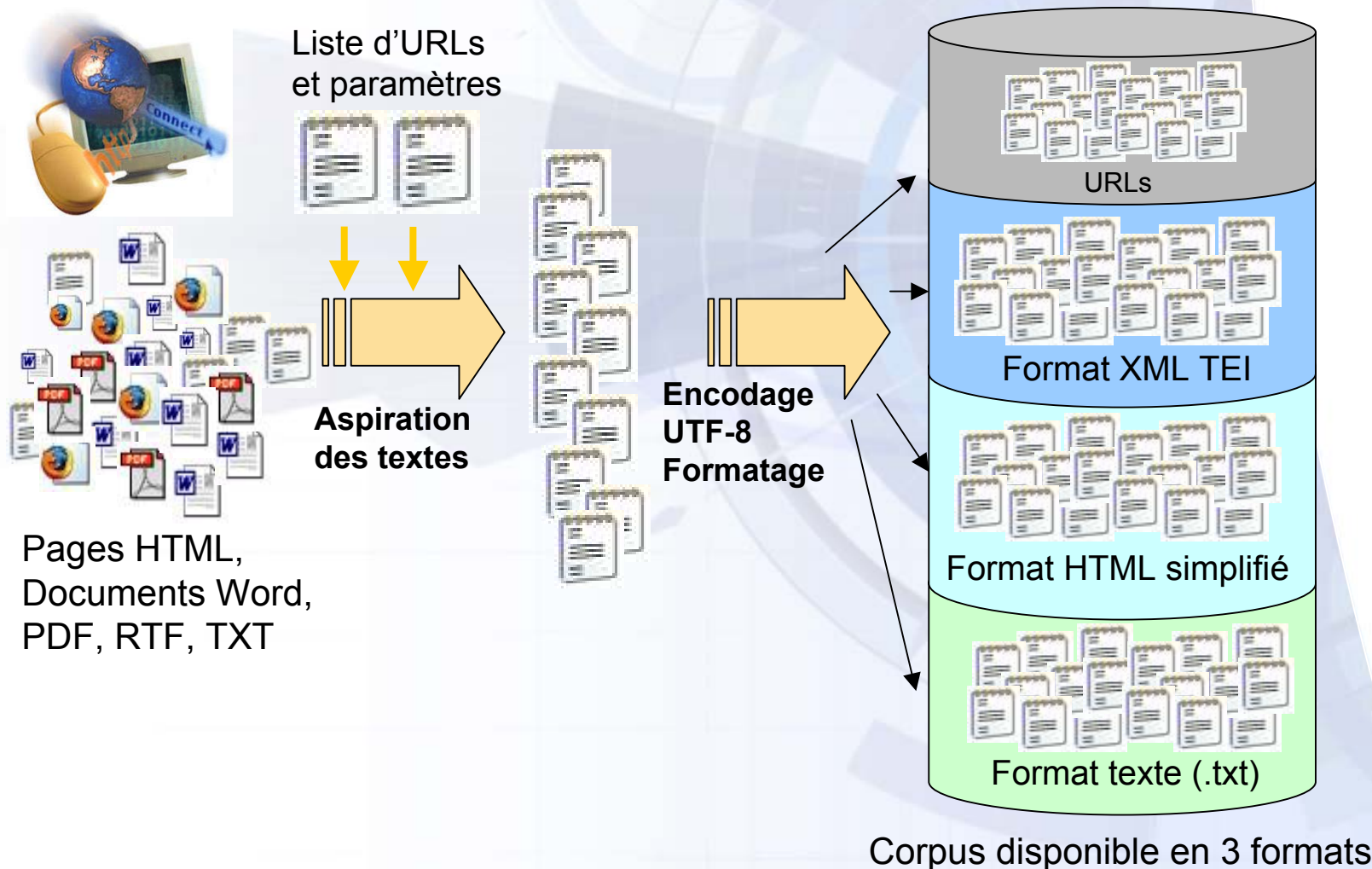
### Outil d'aide à la recherche d'information sur le Web

- Préparer la phase d'enrichissement manuelle de l'ontologie
- Lancer un moteur de recherche Web



# Avancées techniques

## Outil de collecte et de formatage de données



Pages HTML,  
Documents Word,  
PDF, RTF, TXT

Corpus disponible en 3 formats

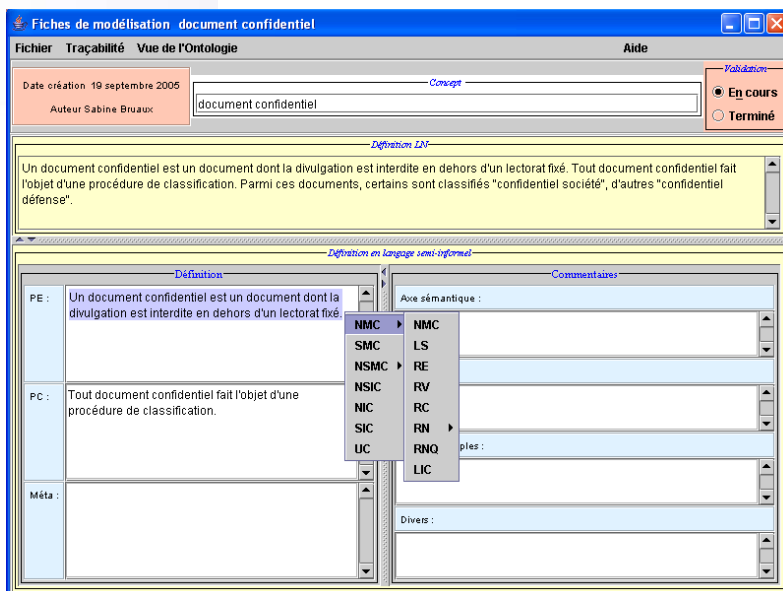
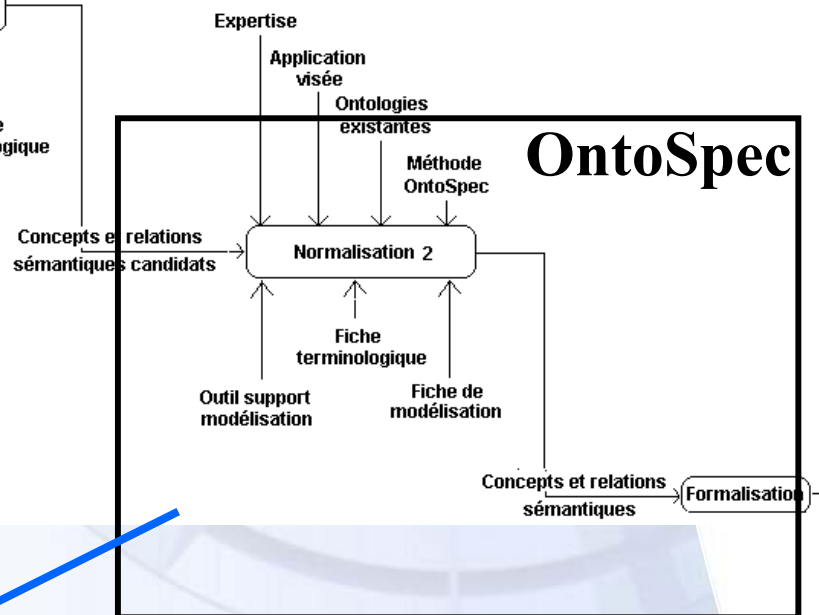
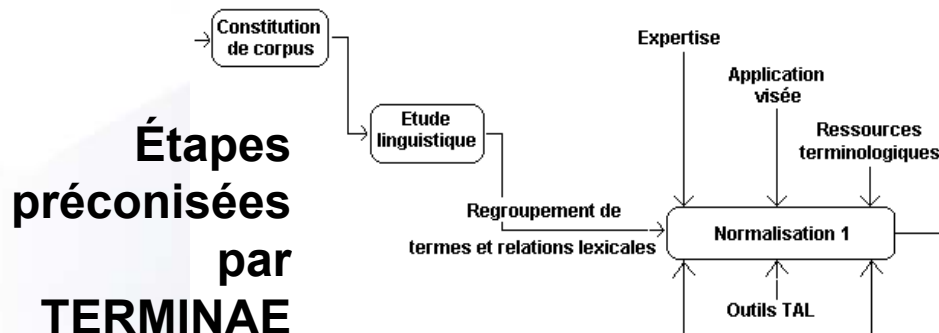
## Avancées techniques

### Un éditeur d'ontologies formelles

- La méthode **OntoSpec** a été complétée et intégrée à la méthode **TERMINAE**
- Intégration d'un éditeur de fiches de modélisation dans la plate-forme **TERMINAE**

# Avancées techniques

## Caractérisation des concepts au moyen d'OntoSpec



Intégration d'un éditeur de fiches de modélisation

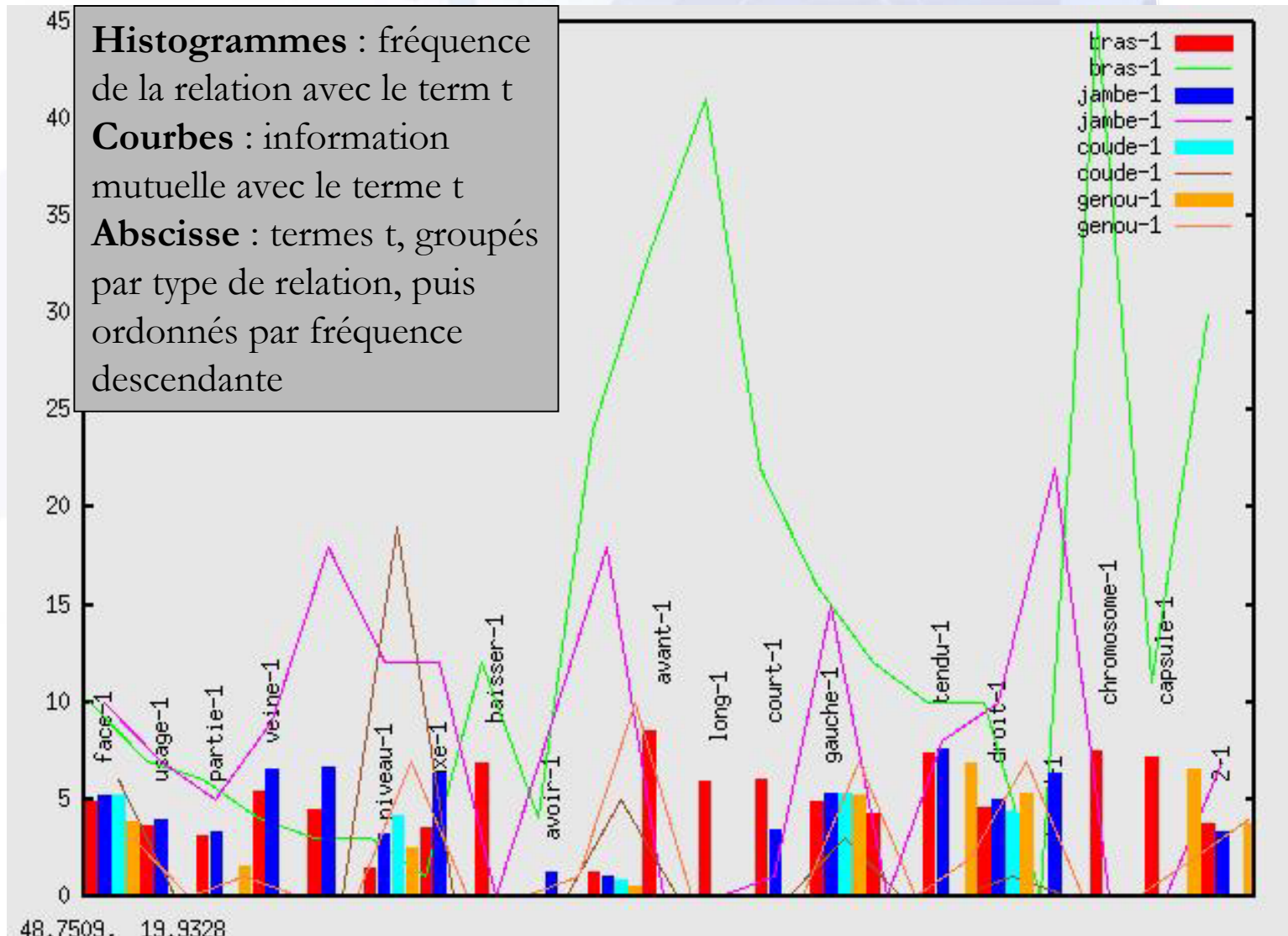
## Avancées techniques

### Outil de production de hiérarchie de termes

- Structuration de la terminologie par l'étude des collocations basées sur les relations de dépendances syntaxiques

#### 4 étapes:

1. Analyse du corpus
2. Sélection des termes candidats
3. Calcul de la similarité
4. Analyse des distributions





# Avancées techniques

## Création de hiérarchie spécialisation/généralisation

**d m o z** open directory project

[about dmoz](#) | [suggest URL](#) | [help](#) | [link](#) | [editor login](#)

Search [advanced](#)

### Arts

[Movies](#), [Television](#), [Music](#)...

### Games

[Video Games](#), [RPGs](#), [Gambling](#)...

### Kids and Teens

[Arts](#), [School Time](#), [Teen Life](#)...

### Reference

[Maps](#), [Education](#), [Libraries](#)...

### Shopping

[Autos](#), [Clothing](#), [Gifts](#)...

### World

[Deutsch](#), [Español](#), [Français](#), [Italiano](#), [Japanese](#), [Nederlands](#), [Polska](#), [Dansk](#), [Svenska](#)...

### Business

[Jobs](#), [Real Estate](#), [Investing](#)...

### Health

[Fitness](#), [Medicine](#), [Alternative](#)...

### News

[Media](#), [Newspapers](#), [Weather](#)...

### Regional

[US](#), [Canada](#), [UK](#), [Europe](#)...

### Society

[People](#), [Religion](#), [Issues](#)...

### Computers

[Internet](#), [Software](#), [Hardware](#)...

### Home

[Family](#), [Consumers](#), [Cooking](#)...

### Recreation

[Travel](#), [Food](#), [Outdoors](#), [Humor](#)...

### Science

[Biology](#), [Psychology](#), [Physics](#)...

### Sports

[Baseball](#), [Soccer](#), [Basketball](#)...

[Become an Editor](#) Help build the largest human-edited directory of the web



Copyright © 1998-2004 Netscape

over 4 million sites - 66,064 editors - over 590,000 categories

# Principaux points de notre méthodologie de structuration hiérarchique

## Problématique étudiée:

Structuration des collections autour du lien de type spécialisation/généralisation

- **Création d'une hiérarchie de thèmes**

- Extraction de thèmes d'un corpus
- Description des thèmes
- Détection de la relation spécialisation/généralisation

- **Hiérarchie de documents**

- Affectation des documents aux thèmes
- Relation de spécialisation/généralisation



# Conclusion

## Résultats

- Forte complémentarité des outils
- Problématique centrale dans le cadre du Web sémantique
- De nombreux problèmes restent ouverts

## Conclusion

### Perspectives

- Intégration de ces outils dans un schéma d'exploitation complet
- Poursuite de l'intégration des outils développés par le LIPN et le LaRIA
- Réutilisation des avancées dans le cadre du projet de plate-forme WebContent (CEA, EADS, LIP6, etc.)
- Exploration d'approches complémentaires dans VODEL (INSA, EADS, etc.)
- Réflexions sur la poursuite des travaux sur un domaine restreint

**TALN 2006- Leuven, 10-13 avril 2006**

**RNTL Technolanguage ATONANT  
Aide à l'enrichissement semi-automatique  
d'ontologies**

**Merci de votre attention**