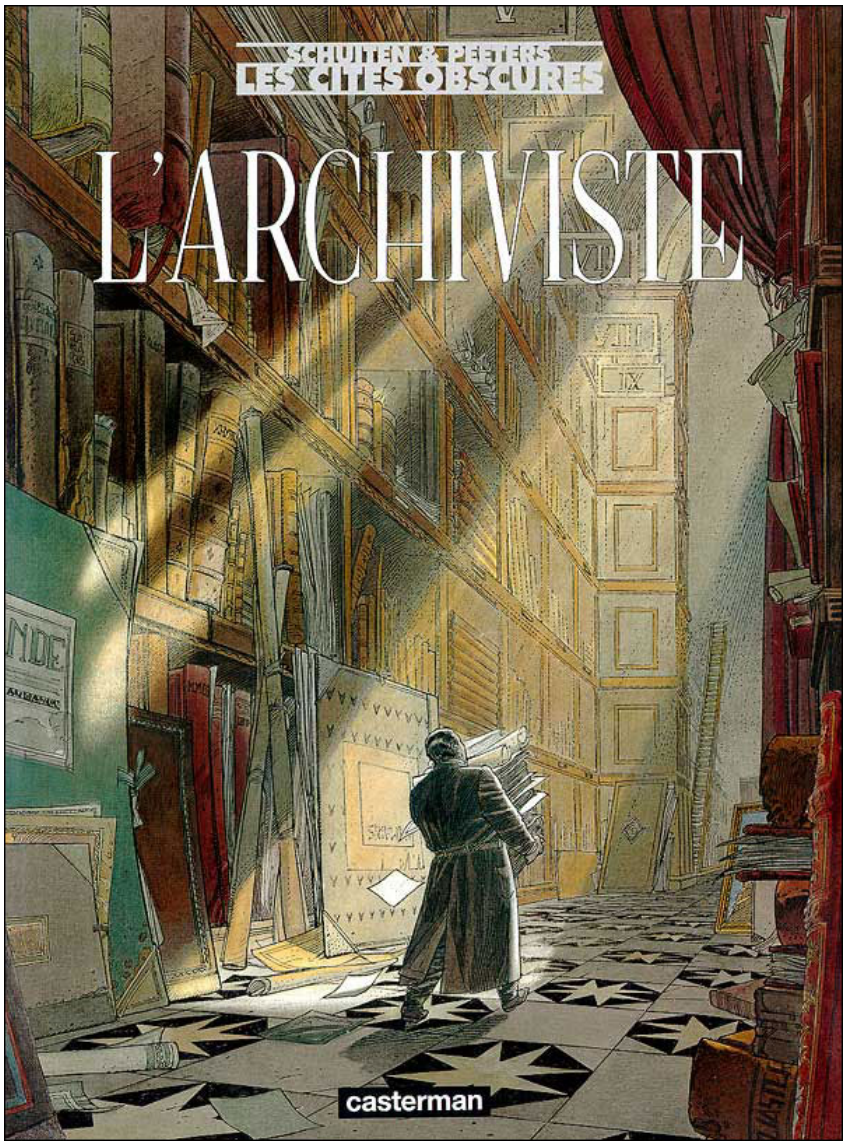
ACCÈS SÉMANTIQUE AUX BASES DE DONNÉES DOCUMENTAIRES

Techniques symboliques de traitement automatique du langage pour l'indexation thématique et l'extraction d'information temporelle

Thèse – Défense publique
Louvain-la-Neuve
31-01-2011

Contexte de la recherche

- Développement des technologies de l'information
  - ➔ La connaissance et l'information sont les nouvelles matières premières de l'activité économique
- Paradoxe de la société et de l'économie de l'information
  - ➔ Les documents sont accessibles plus facilement
  - ➔ La production et la diffusion de documents a augmenté en vitesse et en volume
  - ➔ Mais, face à une requête précise, l'information est noyée dans la masse
- **L'accès à l'information est devenu un enjeu stratégique**

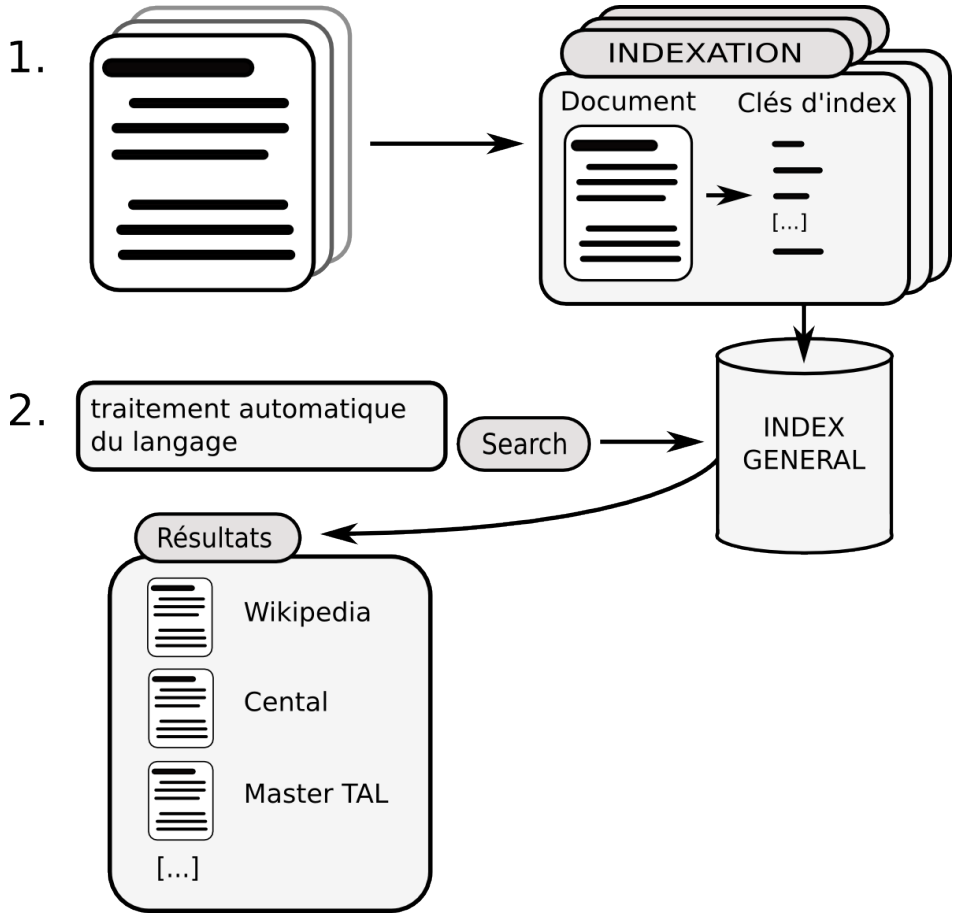


Introduction

- Différentes technologies permettent l'accès aux documents
- Globalement satisfaisantes dans de nombreux cas concrets

Recherche d'informations (RI)

- Deux phases :
  1. Indexation des documents
  2. Recherche des documents

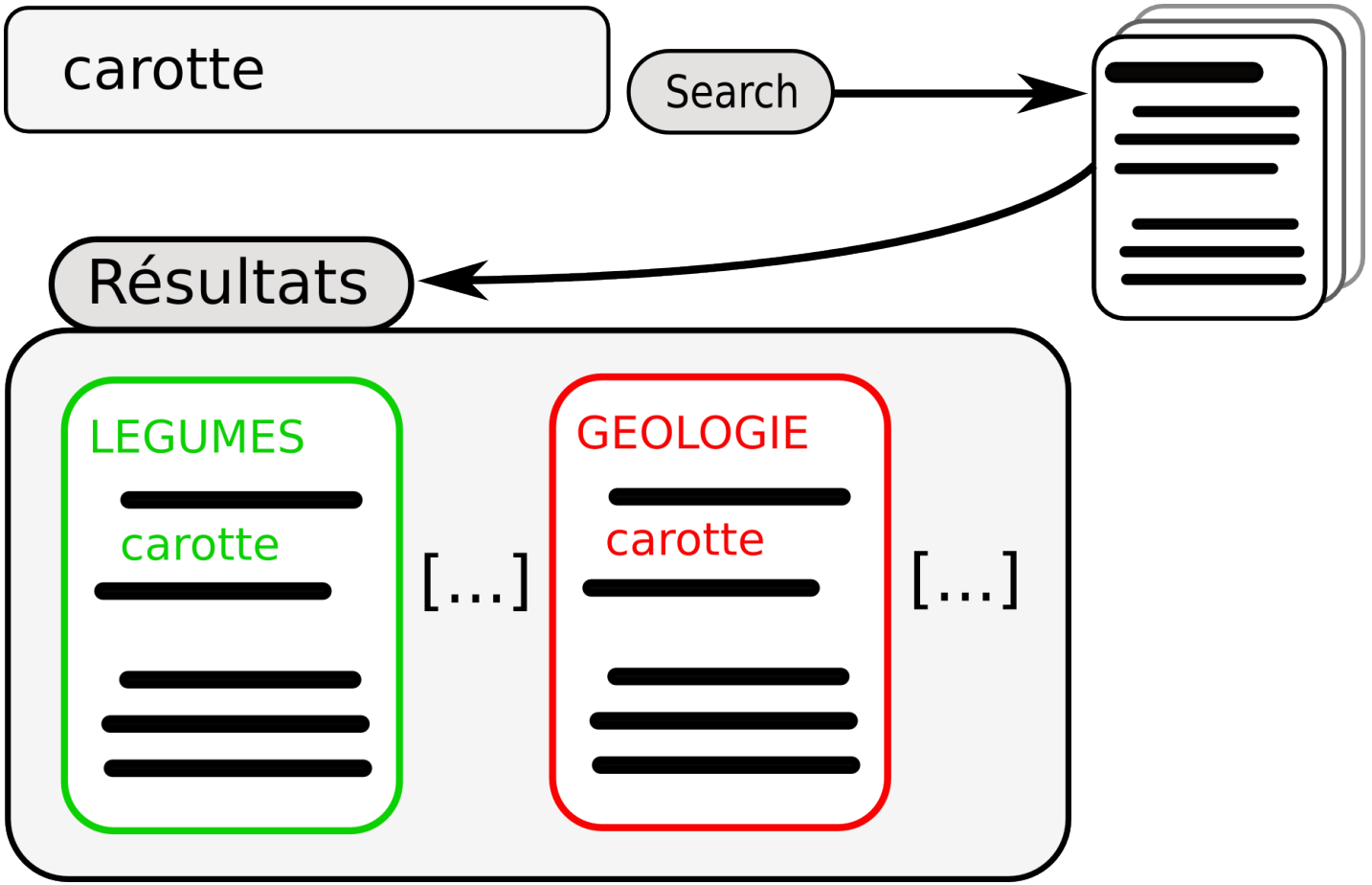


Recherche d'informations (RI)

- Il reste des possibilités d'amélioration pour rendre les résultats plus complets et précis!
- La recherche d'informations (RI) est traditionnellement basée sur un espace de mots, ce qui entraîne certaines difficultés
  - Ambiguïté lexicale
  - Variété lexicale
  - Prise en compte du sens des expressions (e.a. temporelles)
  - Informations complexes, composées
  - ...

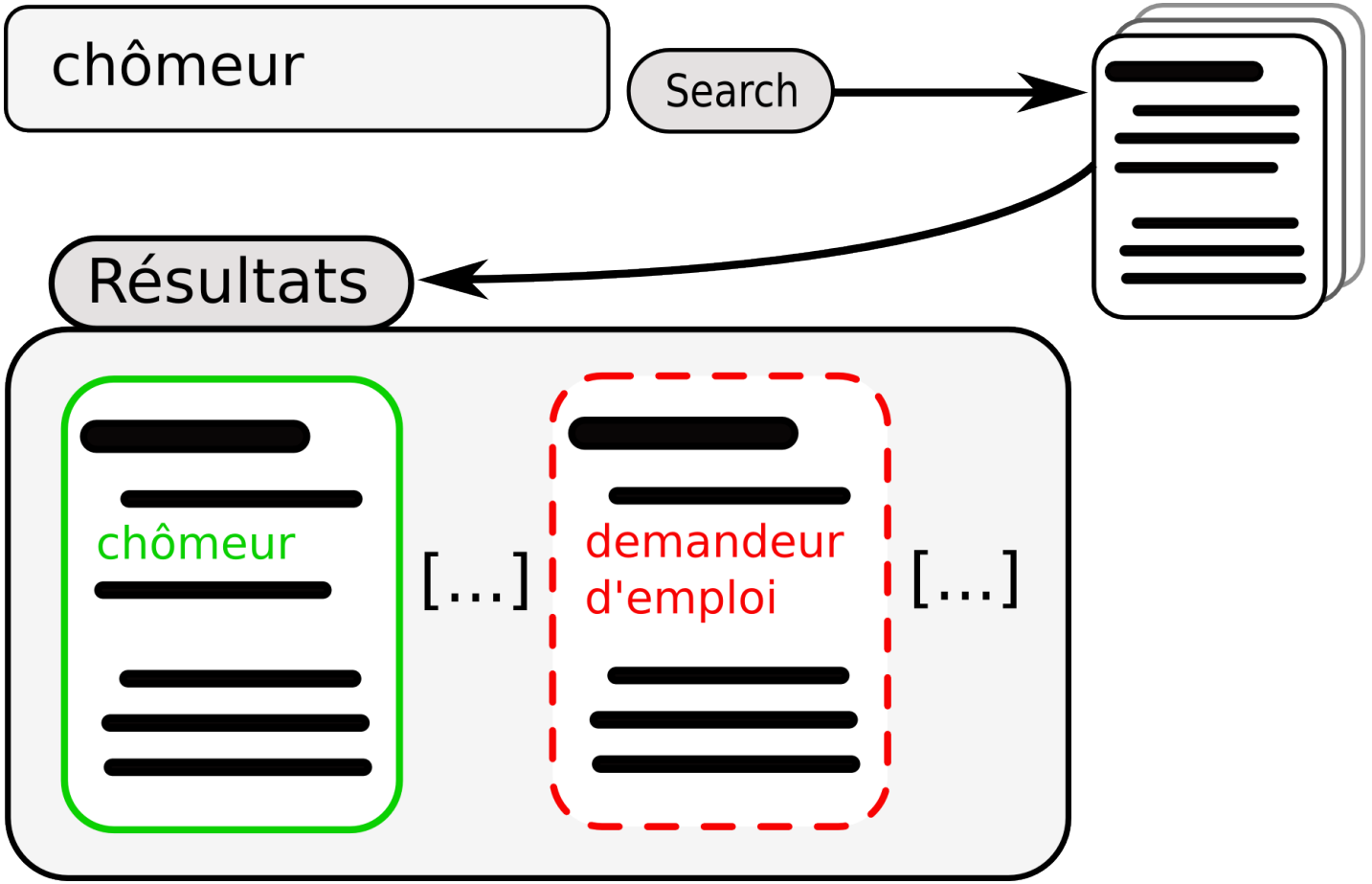
Difficultés en RI

- Ambiguïté lexicale



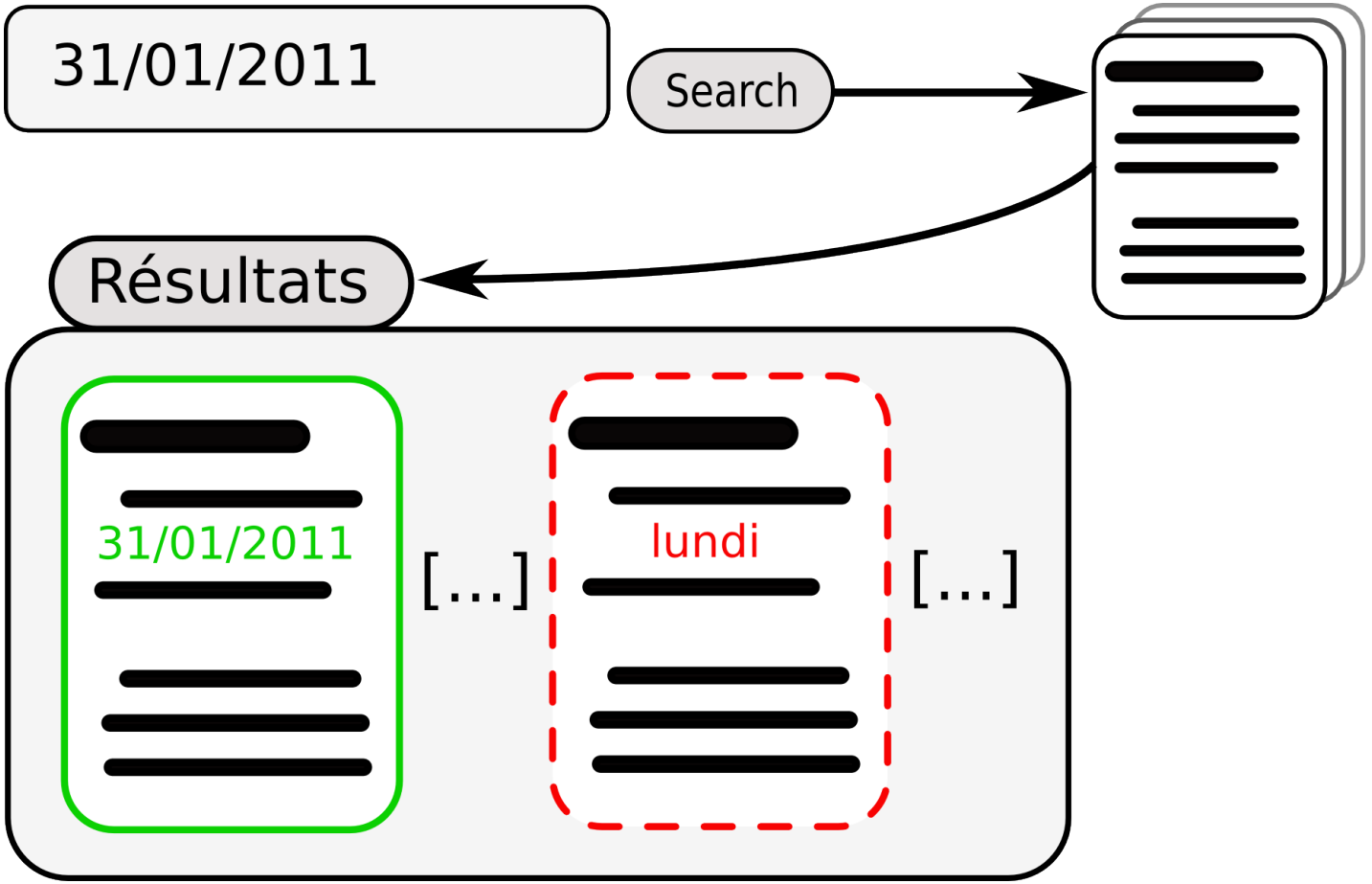
Difficultés en RI

- Variété lexicale



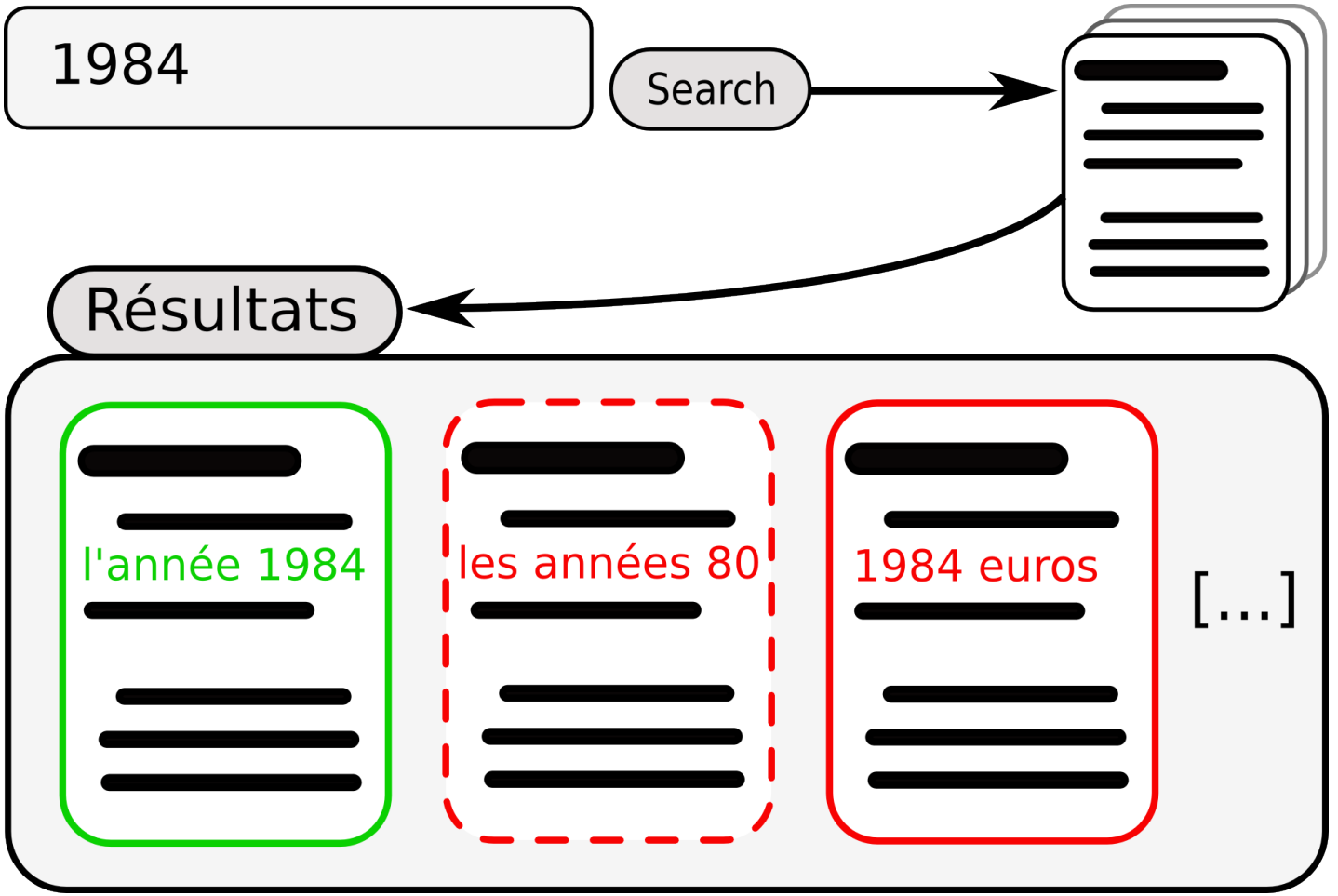
Difficultés en RI

- Sens des expressions temporelles : exemple 1



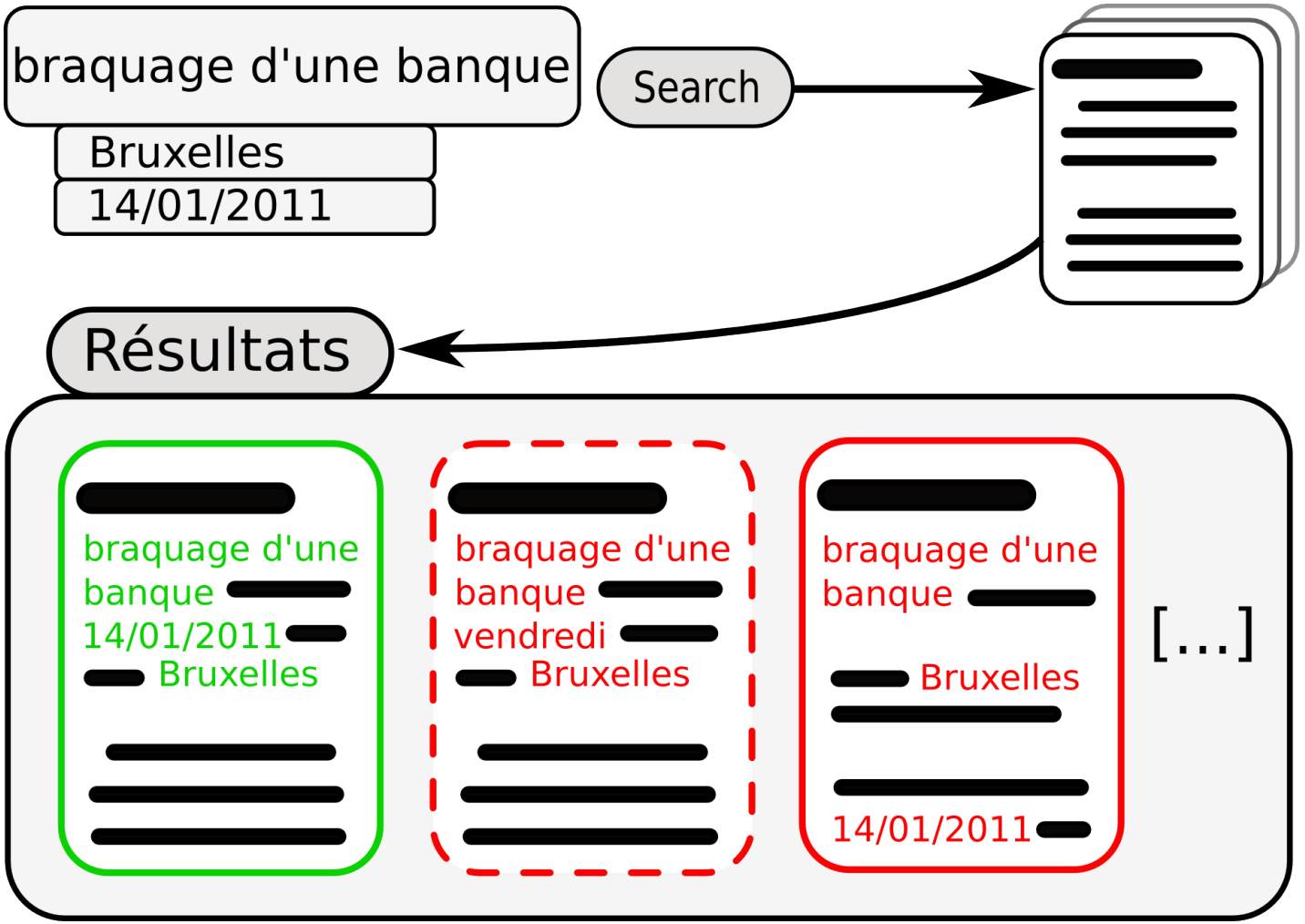
Difficultés en RI

- Sens des expressions temporelles : exemple 2



Difficultés en RI

- Informations complexes, composées



Thèse

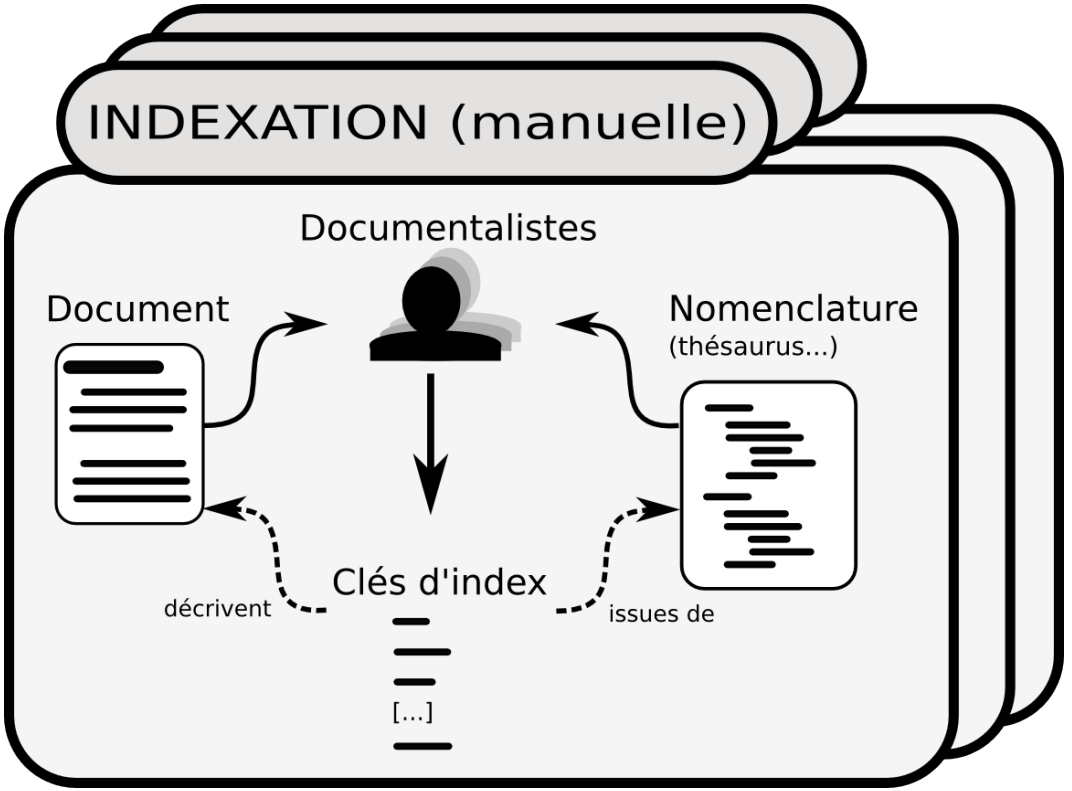
- Cette thèse a pour objectif de montrer si des techniques symboliques de TAL peuvent :
  - venir contribuer à l'**enrichissement sémantique** de la représentation des documents,
  - d'une manière qui serait susceptible d'en **améliorer l'accès**
  - c.-à-d. par le passage d'un espace de **mots** à un espace de **concepts**

Thèse

- Trois façons d'apporter des éléments de sens à la représentation des documents
  1. Indexation thématique (semi) automatique (classification supervisée)
  2. Extraction et d'interprétation d'informations temporelles
  3. Indexation thématique à dimension temporelle (indexation multidimensionnelle)

1. Indexation thématique (semi) auto.

- Classification supervisée
  - ➔ Attribution de catégories issues d'un ensemble défini a priori
  - ➔ Apporte une sémantique bien définie à chaque document ainsi indexé

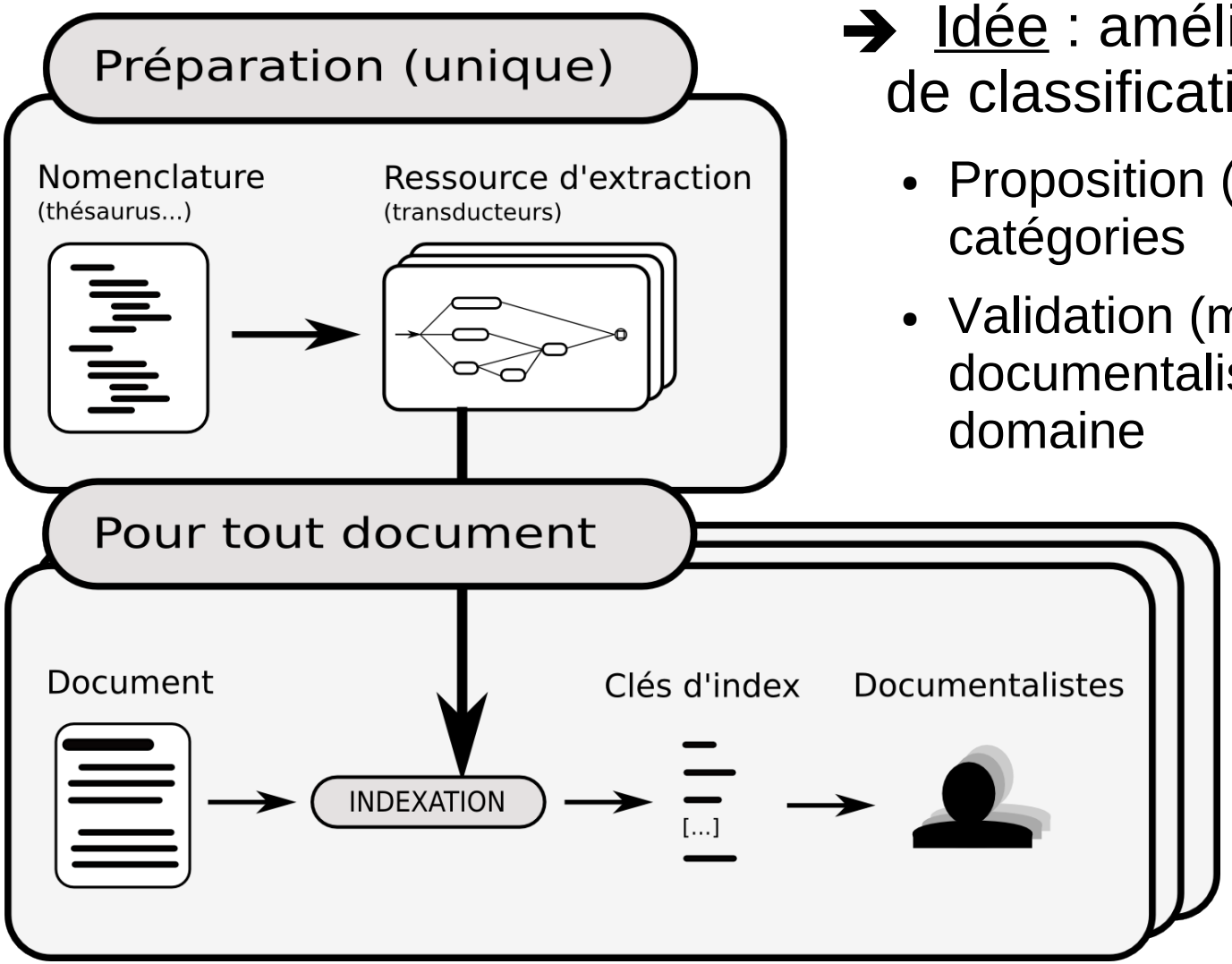


✔ Qualité de l'indexation

✗ Problème de coût et de passage à l'échelle

✗ Cohérence globale à long terme pas nécessairement assurée

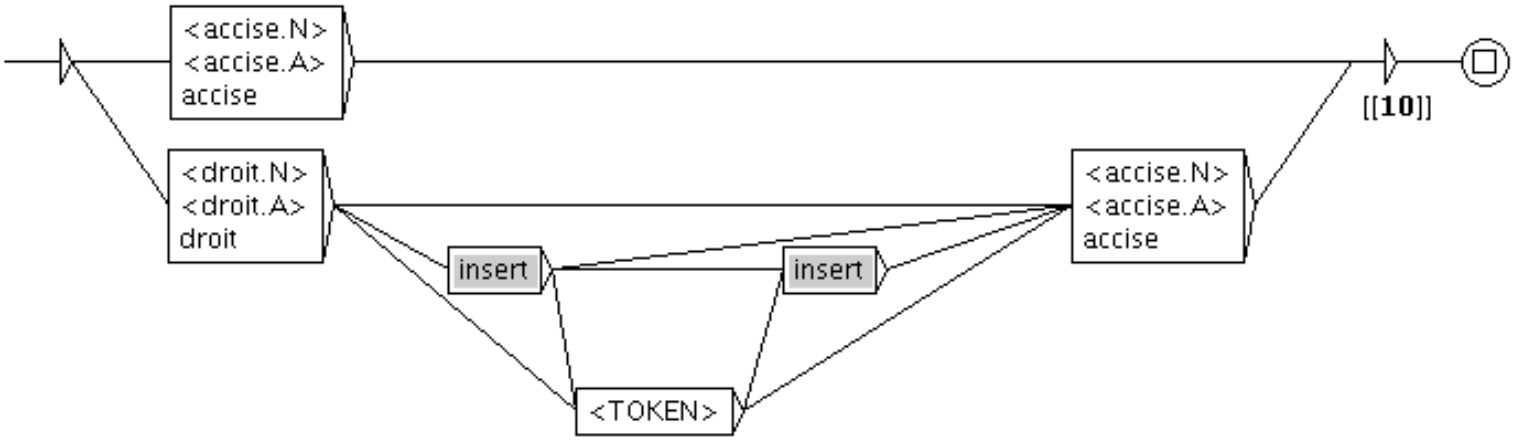
1. Indexation thématique (semi) auto.



➔ <u>Idée</u> : améliorer le processus de classification des documents

- Proposition (automatique) de catégories
- Validation (manuelle) par des documentalistes, experts du domaine

1. Indexation thématique (semi) auto.



Types de variations reconnues lors de l'extraction

- Traitement cruel, inhumain ou dégradant
  - Traitement cruel
  - Traitement inhumain
  - Traitement dégradant
- Fédération des entreprises belges (FEB)
  - Fédération des entreprises belges
  - FEB
- Traitement cruel
  - Traitements cruels
- Contrôle du chômeur
  - Contrôle (de|des) chômeurs
- Contrôle des chômeurs
  - Contrôle *** des chômeurs
    - *** = renforcé, ponctuel, régulier, etc.

1. Indexation thématique (semi) auto.

- Résultats
  - Tests sur textes parlementaires, avec thésaurus *ad-hoc*, et indexation manuelle
  - Classification « générale » (47 catégories)
    - ➔ NbCat = 5,6 – R = 85,87% – P = 32,48%
  - Classification « fine » (2.514 catégories)
    - ➔ NbCat = 10,1 – R = 62,91% – P = 30,86%
  - Les résultats peuvent être améliorés en combinaison avec d'autres méthodes (+SVM (47 catégories) : NbCat = 4,8 – R = 91,72% – P = 33,50%)
- Bilan
  - ➔ Avantage par rapport au processus manuel
    - ✔ Rapidité (coût moins élevé et meilleur passage à l'échelle)
    - ✔ Cohérence
  - ➔ Le gain sémantique amené par l'indexation en catégories devient plus abordable (par rapport au traitement manuel)

2. Extraction temporelle

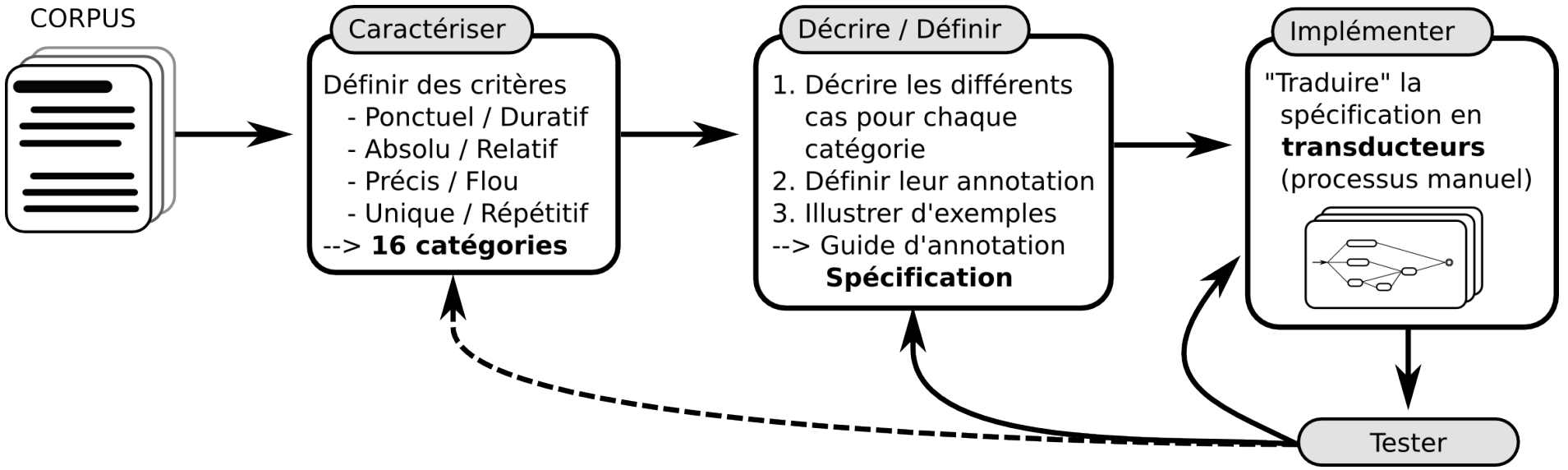
- En RI, l'information temporelle peut être exploitée pour
  - Trouver les documents pertinents
  - Améliorer le *ranking*
  - Filtrer les résultats en exprimant des contraintes additionnelles
  - Proposer de nouveaux modes de présentation des résultats
- Donner aux expressions temporelles une valeur, un sens
  a) Reconnaissance d'expressions porteuses d'une information temporelle
  b) Interprétation de ces éléments et calcul d'une valeur normalisée

2. Extraction temporelle

- Centrée sur les expressions adverbiales
  - ➔ moyen très fréquent de véhiculer une information temporelle dans un texte
- Méthode symbolique
  - ➔ bien adaptée pour décrire le plus complètement et précisément possible ce type d'expression
- Ressource d'extraction
  - ➔ Transducteurs (patrons lexico-syntaxiques) créés à la main
- D'autres éléments interviennent dans le modèle temporel complet (verbes, structure syntaxique...)

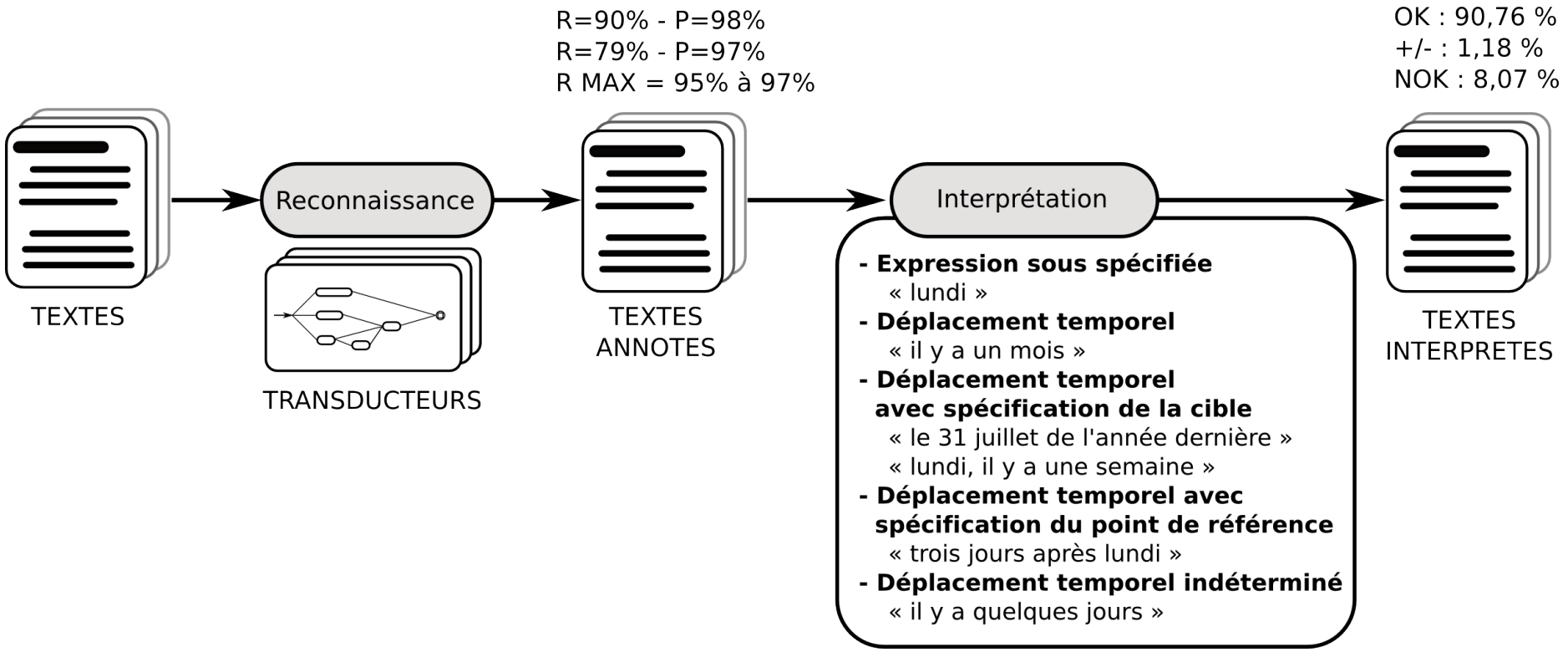
2. Extraction temporelle

- Création de la ressource d'extraction



2. Extraction temporelle

- Reconnaissance et interprétation



3. Indexation à dimension temporelle

- Utiliser les résultats des deux premières parties pour gérer les informations complexes en RI
- Constat dans les systèmes de RI
  - La recherche d'informations complexes (p. ex. incluant une dimension temporelle) est rarement possible
  - Peu de support **réel** des aspects temporels
    - ➔ La date prise en compte est souvent la date de création / diffusion du document

3. Indexation à dimension temporelle

- Idée
  - Indexation classique (thématique) peut évoluer vers une indexation multidimensionnelle

    [*attaque à main armée* ; *banque*]

    [(*attaque à main armée*, *17/03/2005*, *Bruxelles*) ; (*banque*,_,_)]
  - Lier les indices thématiques et les expressions temporelles au niveau de la proposition
- Tests et évaluation
  - Liens thématico-temporels : précision de 98,97%
  - Apport informationnel aux catégories de l'indexation :
    - Catégories avec dimension temporelle : 27,32 %
    - Uniquement celles différentes de la date de l'article : 18,90%
    - Remarque : toutes les catégories ne doivent pas nécessairement avoir une dimension temporelle!

3. Indexation à dimension temporelle

- Exemple

<text id='F051230A_0098.txt' date='20051230-15:59'>
<title>VTT - F. Meirhaeghe engagé pour 3 ans chez Landbouwkrediet (1LEAD) . </title>
§ ANDERLECHT **30/12** (BELGA) = Le coureur **belge** de mountainbike Filip Meirhaeghe a signé **jeudi** un **contrat** de trois ans avec la formation cycliste Landbouwkrediet-Colnago, a **communiqué à la presse**, **vendredi**, la direction de l'équipe dirigée par Gérard Bulens.
§ Meirhaeghe, suspendu **depuis l'été 2004 jusqu'au 14 janvier 2005** pour avoir utilisé de l'EPO, a été engagé pour cette période de 3 ans afin de pouvoir se préparer au mieux pour les **Jeux OLympiques** de Pékin **en 2008**. Le biker flandrien sera alors âgé de 37 ans.
</text>

Indice thématique	Expr. temporelle	Val. temporelle
belge	30/12	30/12/2005
	jeudi	29/12/2005
contrat	30/12	30/12/2005
	jeudi	29/12/2005
Communiqué à la presse	vendredi	30/12/2005
Jeux OLympiques	En 2008	2008 (fuzzy=1)

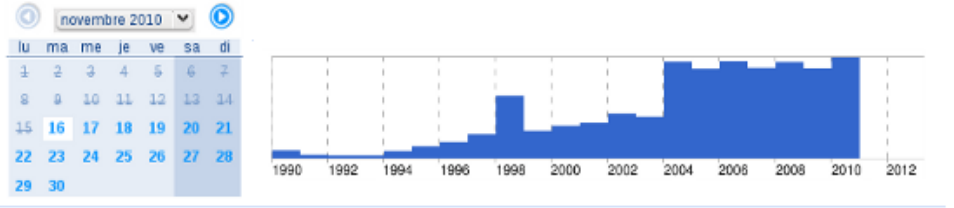
Perspectives

CENTAL

Rechercher

Environ 199.000 résultats (0,05 secondes) Recherche avancée

- 04 VIE POLITIQUE
- 08 RELATIONS INTERNATIONALES
- 10 COMMUNAUTÉS EUROPÉENNES
- 12 DROIT
- 16 VIE ÉCONOMIQUE
  - 1606 politique économique
  - 1611 croissance économique
  - 1616 région et politique régionale
  - 1621 structure économique
  - 1626 comptabilité nationale
  - 1631 analyse économique
- 20 ÉCHANGES ÉCONOMIQUES ET COMMERCIAUX
- 24 FINANCES
- 28 QUESTIONS SOCIALES
- 32 ÉDUCATION ET COMMUNICATION
- 36 SCIENCES
  - 3606 sciences naturelles et appliquées
  - 3611 sciences humaines
- 40 ENTREPRISE ET CONCURRENCE
- 44 EMPLOI ET TRAVAIL
- 48 TRANSPORTS
- 52 ENVIRONNEMENT



Traitement automatique du langage naturel - Wikipédia
Le **Traitement automatique du langage** naturel (abr. TALN) ou Traitement automatique des langues (abr. **TAL**) est une discipline à la frontière de la **...**
Histoire - TAL statistique - Les applications TAL - Voir aussi
fr.wikipedia.org/.../**Traitement_automatique_du_langage**_naturel - En cache

Catégorie:**Traitement automatique du langage** naturel - Wikipédia
Une page de Wikipédia, l'encyclopédie libre. Aller à : Navigation, rechercher. Article principal : **Traitement automatique du langage** naturel. **...**
fr.wikipedia.org/.../Catégorie:**Traitement_automatique_du_langage**_naturel -
En cache - Pages similaires

UCL - Centre de **traitement automatique du langage**
Présentation de l'équipe de recherche, de l'actualité et des projets du laboratoire de Cédrick Fairon.
www.uclouvain.be/cental - En cache - Pages similaires

UCL - Option en **traitement automatique du langage** [15.0]
Traitement automatique du langage naturel, Cédrick Fairon, 22.5h, 5 crédits, 1q. Au choix LFLTR2630, Méthodologie du traitement informatique des données **...**
www.uclouvain.be/prog-2010-loptrom2m_tl.html - En cache
Plus de résultats de uclouvain.be

Association pour le **traitement automatique** des langues
2 nov. 2010 **...** Rechercher. Actualités. **TAL** 50:3 paru. Le numéro Apprentissage automatique pour le **TAL** (2009) est paru. [Espace privé] - RSS. Sur le Web **...**
www.atala.org/ - En cache - Pages similaires

Patrice Mellot Consultant - **Traitement automatique du langage ...**
développement de solutions de **traitement automatique du langage** naturel.

Conclusions

- Les techniques symboliques de TAL peuvent contribuer à enrichir et améliorer la représentation sémantique des documents
  - De manière automatique ou semi automatique
  - À partir de ressources linguistiques construites automatiquement ou manuellement
  - Pour indexer du contenu thématique ou relatif à d'autres dimensions, telle que l'information temporelle
- Les systèmes de RI actuels sont susceptibles de tirer parti de ces apports sémantiques
- D'autres applications peuvent également en profiter

ACCÈS SÉMANTIQUE AUX BASES DE DONNÉES DOCUMENTAIRES

Techniques symboliques de traitement automatique du langage pour l'indexation thématique et l'extraction d'information temporelle

Thèse – Défense publique
Louvain-la-Neuve
31-01-2011