Approches pour l'extraction de connaissances à partir de textes et de ressources structurées

Cours DECOL - M2 - ECA

E. Kergosien (IRSTEA - LIRMM)

eric.kergosien@lirmm.fr

Organisation du cours

- Introduction : Définitions
- Approches pour l'extraction d'informations
  - Approches manuelles versus Approches automatiques
  - Approches automatiques : statistique versus linguistique
  - Apport des ressources structurées
- Le Traitement Automatique du Langage Naturel
  - Emergence d'une chaine standardisée
  - Extraction d'entités nommées ENs
  - Les boîtes à outils pour le TALN
- Un exemple d'application : le projet Senterritoire
  - Approche théorique
  - Mise en œuvre sous Linguatream pour l'extraction des ENs
  - Exercices à venir

Références

- Cours de Xavier Tannier, Maître de Conférence à l'Université Paris Sud 11
- Cours de T. Poibeau, CNRS et Université Paris 13
- Cours de A.Rozenknop, Université Paris 13

I. Introduction

1. Définitions

a. Extraction d'informations

b. Textes non structurés semi-structurés et structurés

UNIVERSITÉ MONTPELLIER 2
SCIENCES ET TECHNIQUES

Besoin d'information

- « Disposer des bonnes informations par rapport à une question ou à un problème donné »
- Importance accrue de la veille scientifique, technologique, commerciale, culturelle, etc.
- Un besoin très grand public
- Au départ, un besoin d'accès aux informations internes, mais l'accès aux informations externes est maintenant tout aussi important

Diversité des besoins d'information (1/2)

1. La recherche d'un élément connu

   - L'utilisateur sait exactement quels éléments il recherche. Il sait reconnaître les éléments désirés s'il les voit

     Ex : recherche d'une citation bibliographique précise

   → Utilisation d'outils orientés Bases de données: SQL, Xquery, etc.

Diversité des besoins d'information (1/2)

1. La recherche d'un élément connu

   - L'utilisateur sait exactement quels éléments il recherche. Il sait reconnaître les éléments désirés s'il les voit

     Ex : recherche d'une citation bibliographique précise

   → Utilisation d'outils orientés Bases de données: SQL, Xquery, etc.

2. La recherche d'une information spécifique

   - L'utilisateur recherche une information spécifique mais ignore sous quelle forme elle se présente
   - Réponse partielle possible

     Ex : A quelle date la nouvelle mairie de Montpellier a-t-elle été inaugurée?

   → Extraction d'informations et systèmes de questions-réponses (http://perso.limsi.fr/Individu/anne/DEA/QR_cours_m2.pdf)

Diversité des besoins d'information (2/2)

3. La recherche d'un information générale
   - L'utilisateur recherche une information sur un sujet en général. Il existe de nombreuses façons de décrire le sujet
   - L'information peut ne pas être pertinente ou ne pas être reconnue

   → Recherche d'information

Diversité des besoins d'information (2/2)

3. La recherche d'un information générale
   - L'utilisateur recherche une information sur un sujet en général. Il existe de nombreuses façons de décrire le sujet
   - L'information peut ne pas être pertinente ou ne pas être reconnue

   → Recherche d'information

4. L'exploration
   - Le but n'est pas de répondre à une question en particulier, mai de parcourir l'ensemble des données pour découvrir quels types d'informations concernant un domaine ou un sujet sont présents.

   → Navigation

Positionnement

- **Recherche d'information (IR)**: identifie un ensemble de documents à partir d'un ensemble plus large (document assimilé à un « sac de mots »).

  Ex: Trouver les documents qui traitent de rachats d'entreprises

- **Extraction d'information (IE)**: extrait et structure de l'information précise contenue dans un document.

  Ex: Etablir une base de données où l'on peut retrouver les noms des entreprises informatiques cédées en 2003

- **Compréhension de texte**: représente de façon explicite toute l'information d'un document (rhétorique, intentionnalité, etc.)

  Ex: Déterminer les différentes visées stratégiques sous-jacents à travers ces ventes et acquisitions

Qu'est ce que l'extraction?

- Une tâche qui consiste à extraire de l'information structurée à partir d'un document textuel

L'université Montpellier II est séparée sur deux sites dans le nord de Montpellier.

- Relation d'acquisition:
  - Entité identifiée : université Montpellier II
  - Localisation : dans le nord de Montpellier

Qu'est ce que l'extraction?

- Une tâche qui consiste à extraire de l'information structurée à partir d'un document textuel

L'université Montpellier II est séparée sur deux sites dans le nord de Montpellier.

- Relation d'acquisition:
  - Entité identifiée : université Montpellier II
  - Localisation : dans le nord de Montpellier

→ L'extraction d'information:

- Ne cherche plus à comprendre les textes dans leur ensemble
- Vise à extraire d'un texte donné des éléments pertinents

→ Le type d'information pertinente pour une application est définie à l'avance par un modèle

Définitions

- "L'activité qui consiste à remplir automatiquement une banque de données à partir de textes écrits en langue naturelle" (T. Poibeau)
- Une approche guidée par le but :
  - Identifier les occurrences d'événements particuliers
  - En extraire les arguments impliqués
  - En donner une représentation structurée

- L'analyse s'effectue au niveau local
- Seule une partie du texte est considérée

  (10 à 20 % de texte utile pour un tâche spécifique)

Exemple d'extraction d'information non structurée (2/2)

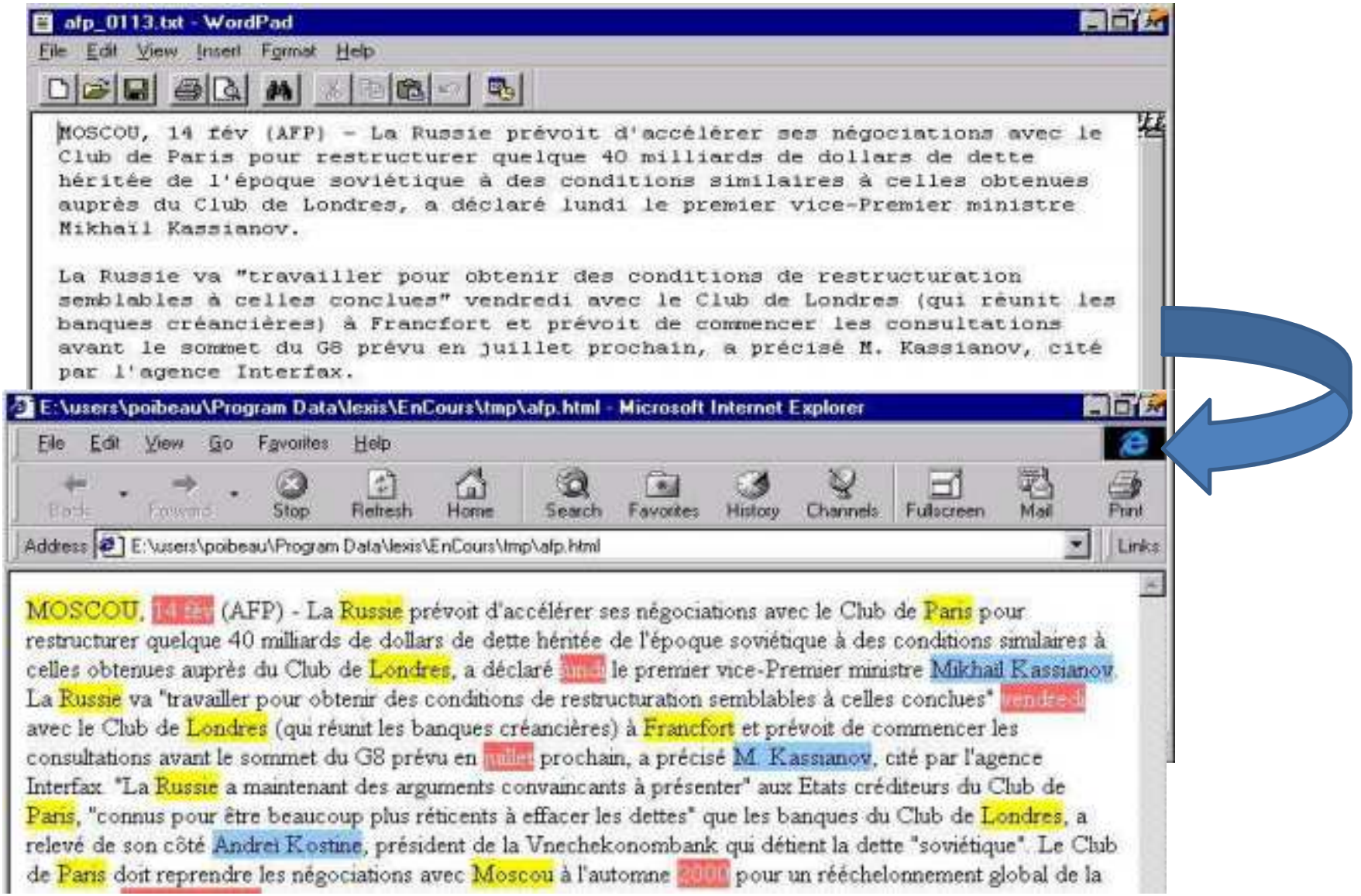
- Entrée :

  La France a battu l'Australie 33 à 6 lors d'un match de préparation de rugby le 11 novembre 2012 au stade de France.

- Cadre :

Evènement sportif
Sport: rugby Équipe 1: France équipe 2: Australie Score : 33 à 6 Date: le 11 novembre 2012 Lieu : au stade de France

Exemple d'extraction d'information non structurée (2/2)



Exemple d'extraction d'information semi-structurée

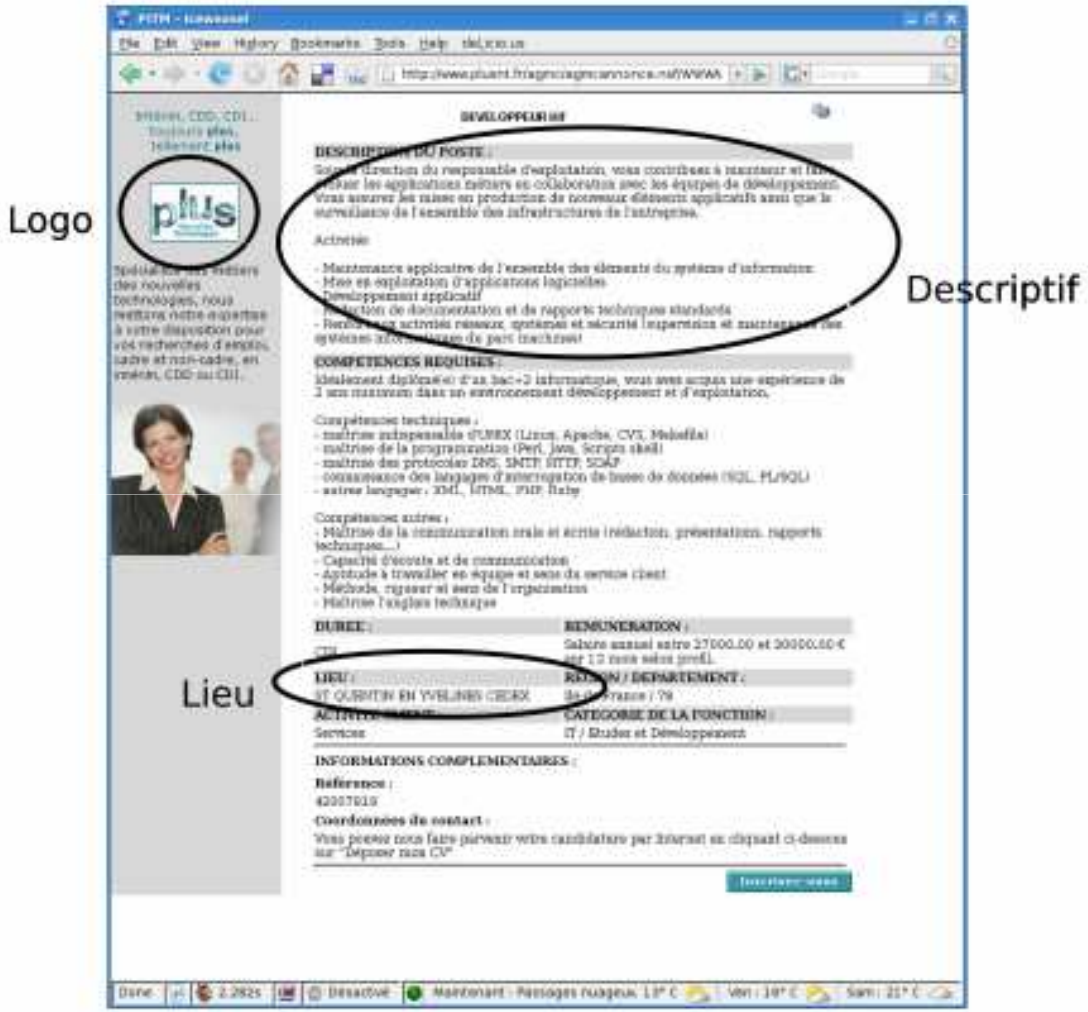


Figure: Extraction de trois champs de la page précédente

Du texte à un patron (template)

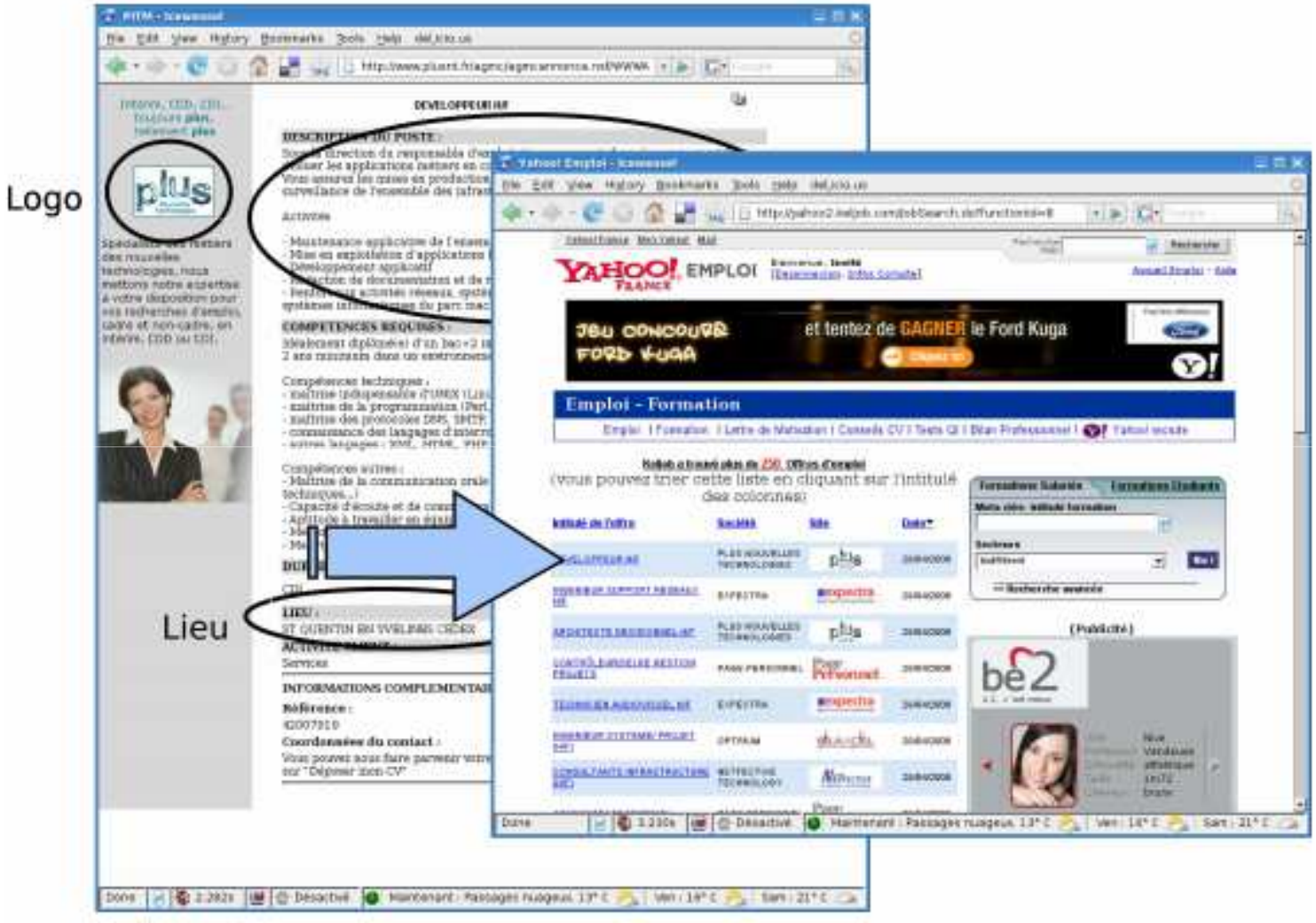


Figure: yahoo.keljob.com

Du texte à un patron (template)

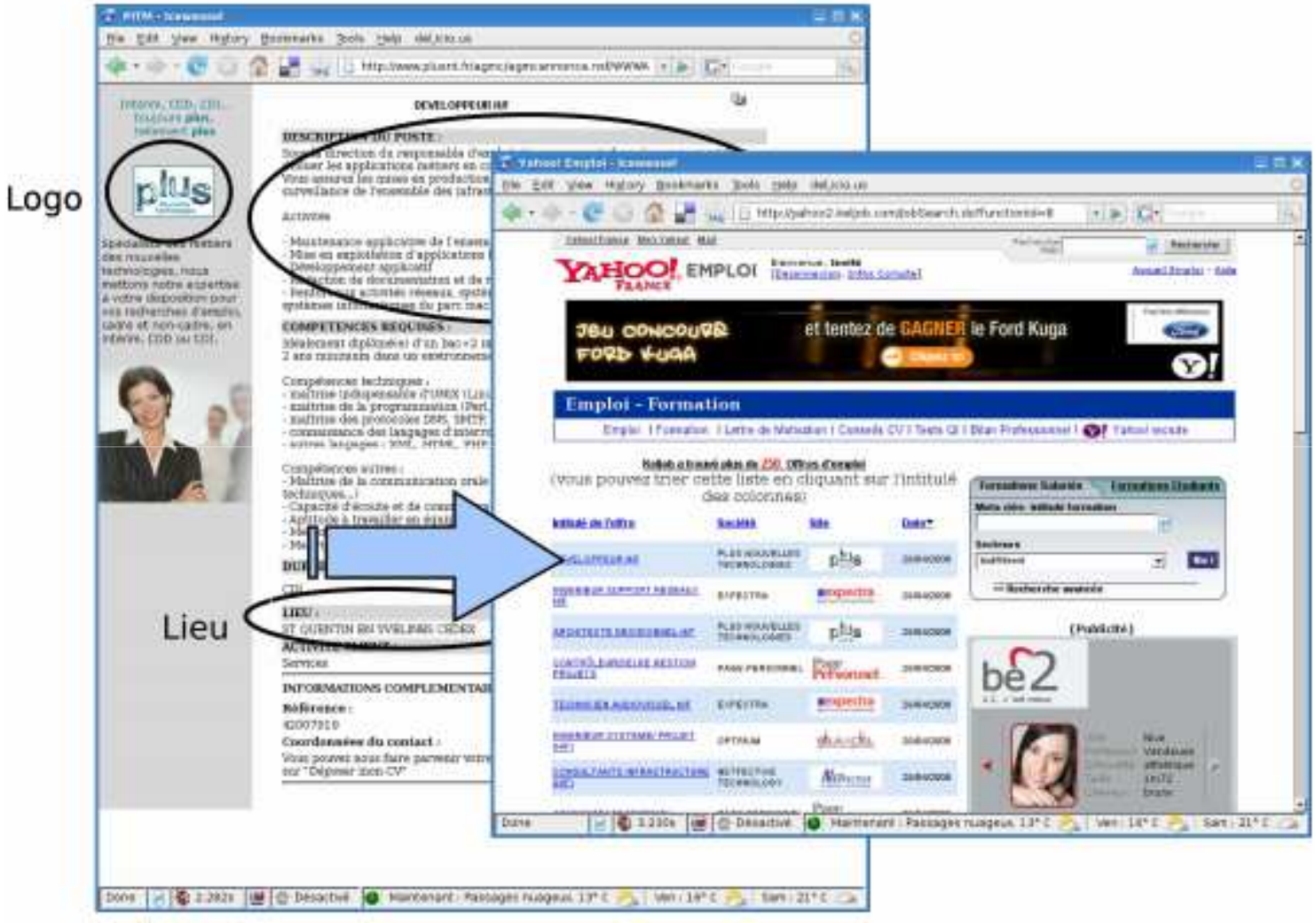


Figure: yahoo.keljob.com

Des patrons partout!

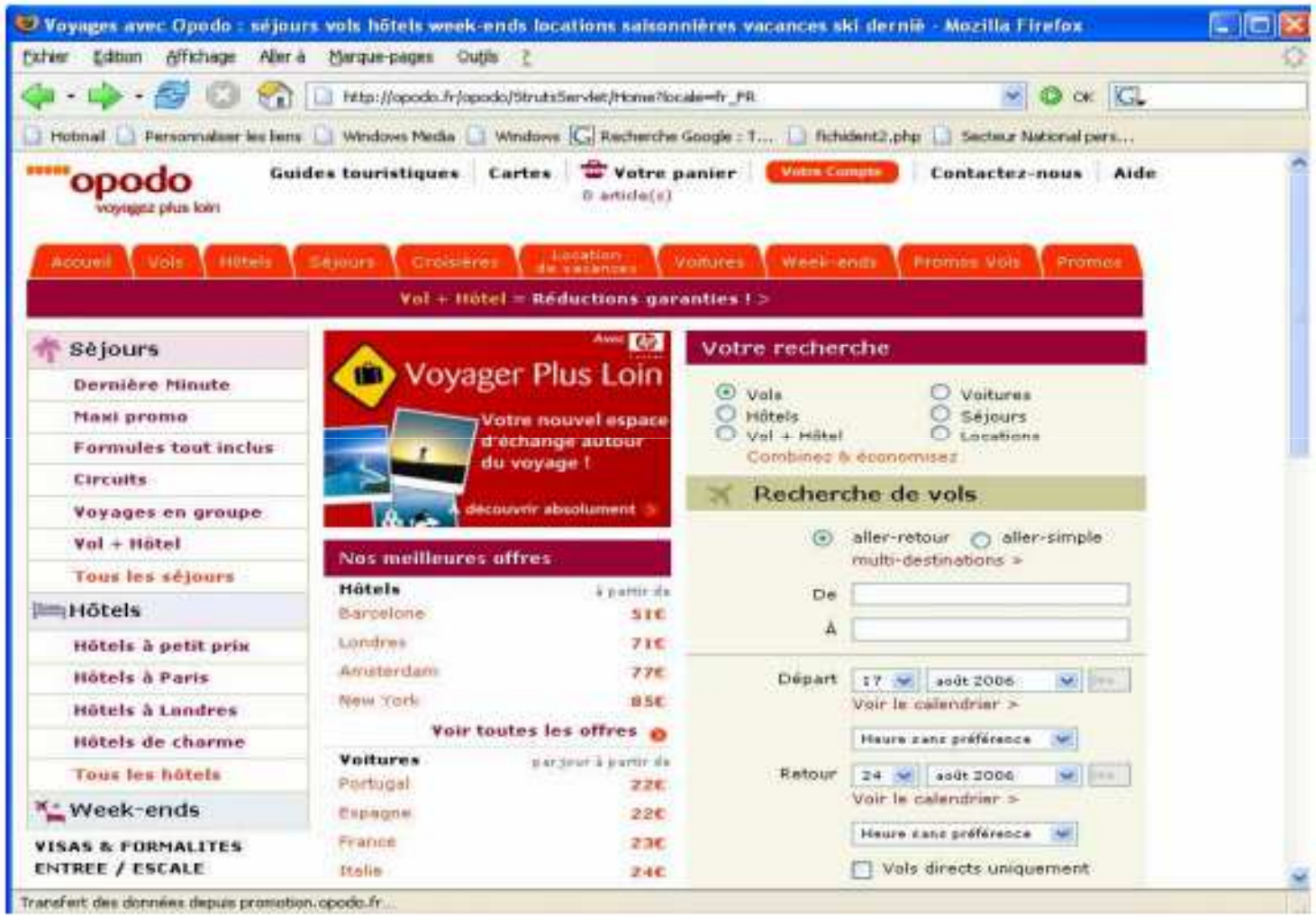
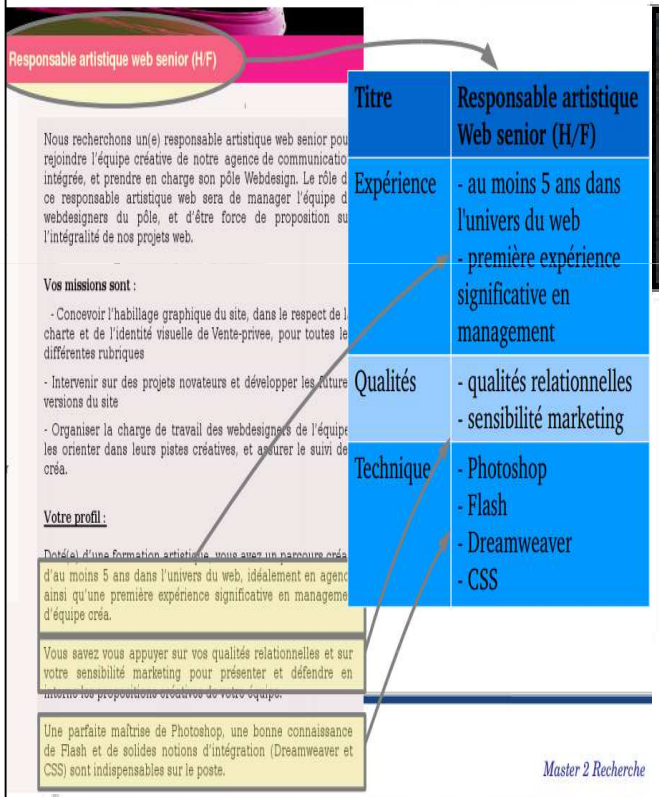


Figure: yahoo.keljob.com

Largeur de la couverture désirée

Sites internet (structure)

Spécifique à un domaine (patron)



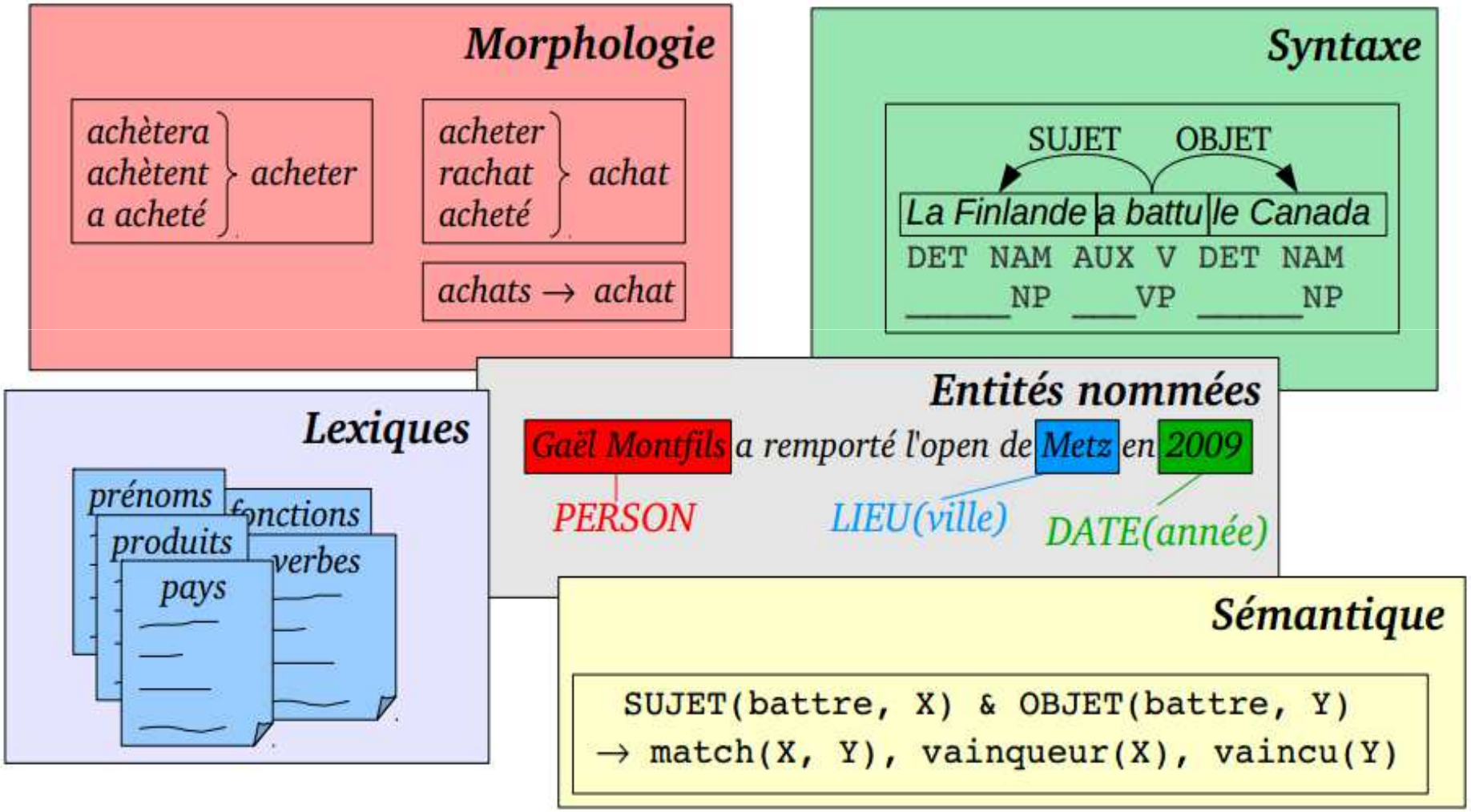
Titre	Responsable artistique Web senior (H/F)
Expérience	- au moins 5 ans dans l'univers du web - première expérience significative en management
Qualités	- qualités relationnelles - sensibilité marketing
Technique	- Photoshop - Flash - Dreamweaver - CSS

Généraliste (Langage)

Prise en compte des variantes lors de l'extraction

- Morphologiques
  - capitale de l'Europe / capitale européenne
- Lexicales
  - la reine de Hollande / des Pays-Bas
- Syntaxiques
  - Moscou compte 9 millions d'habitants / Les neufs millions d'habitants de Moscou
- Sémantiques
  - Adolf Hitler est mort / s'est suicidé

Niveaux d'analyse utilisés (1/2)



Niveaux d'analyse utilisés (2/2)

Indications de mise en forme

Liste des codes pays :

```
<ul>
  <li><b>EE</b> – <i>Estonie</i></li>
  <li><b>ET</b> – <i>Éthiopie</i></li>
  <li><b>FK</b> – <i>Maldives</i></li>
  <li><b>FR</b> – <i>France</i></li>
</ul>
```

Mécanismes d'inférence

"Federer a réalisé un superbe coup droit gagnant lors d'une balle de match en sa faveur."

```
point(P) & coup_gagnant(P, J)
→ vainqueur(J, P)
```

```
balle_de_match(B, P)
→ point(P)
```

```
match(J1, J2) & balle_de_match(B, J1) & vainqueur(J1, B)
→ vainqueur(J1), vaincu(J2)
```

I. Introduction

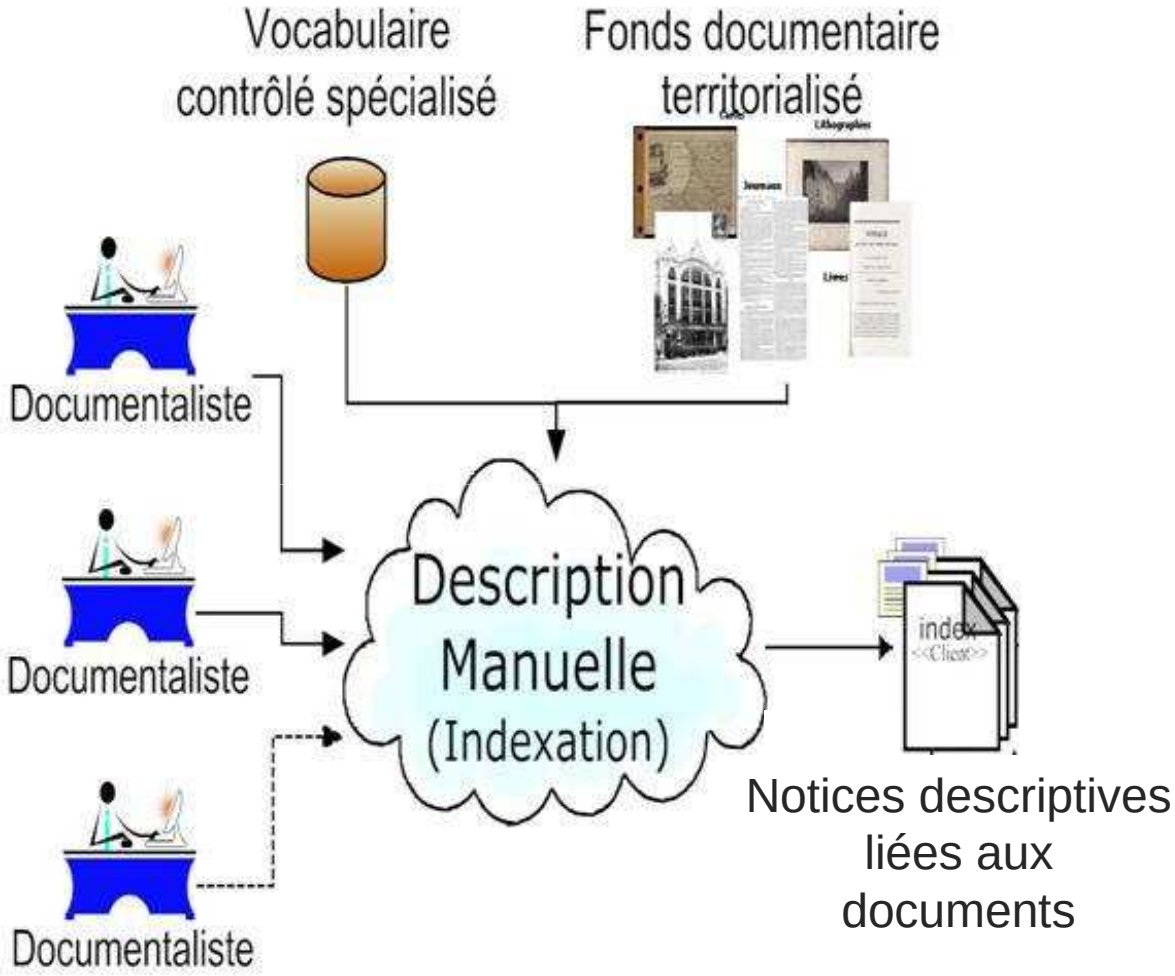
1. Définitions

2. Approches manuelles versus Approches automatiques

3. Approches auto.: statistique versus linguistique

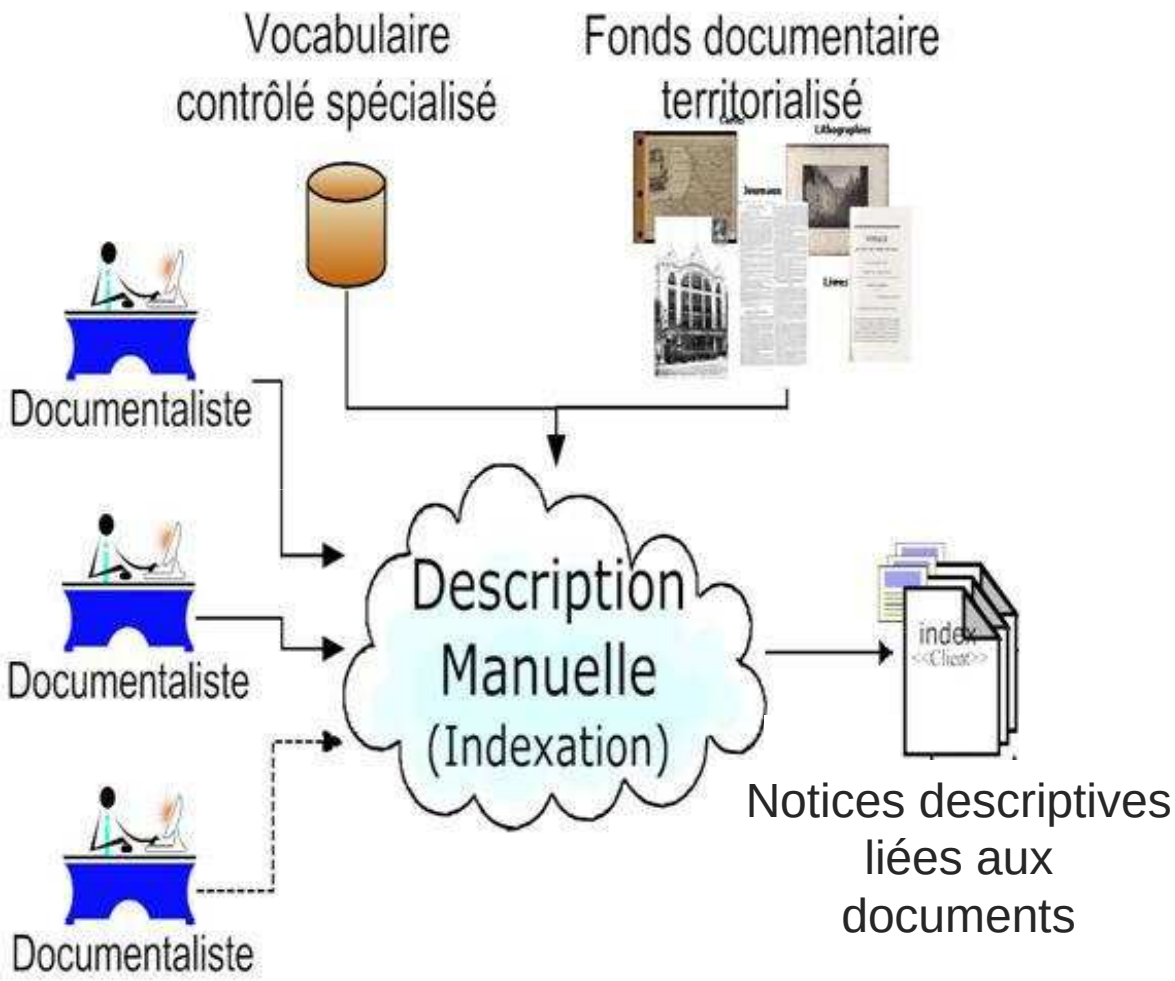
4. Apport des ressources structurées

Extraction manuelle (1/2)



Notices descriptives liées aux documents

Extraction manuelle (1/2)



- Permet une description précise et complète du documents [Chartron et al., 1989], [Anderson et al., 2001], [Savoy, 2005];
- Régulièrement : Utilisation d'un vocabulaire contrôlé (taxinomie ou thésaurus)
- Travail fastidieux et coûteux en temps

Extraction manuelle (2/2)

Lithographie

Travail d'annotation

```
<NOTICE>
<BNSA_GEOREF>Pau</BNSA_GEOREF>
<DATM>2005-12-19</DATM>
<DOC_AUTEUR>anonyme</DOC_AUTEUR>
<DOC_AUTMORAL>Labouche Frères
    (Toulouse)</DOC_AUTMORAL>
<DOC_COTE>2-045-2</DOC_COTE>
<DOC_DAT_CREAT>2005-10-04</DOC_DAT_CREAT>
  <DOC_DEE>Henri IV (roi de France ; 1553-1610) --
    Statues  -- Pau (Pyrénées-Atlantiques) -- Cartes
    postales -- 20e siècle
  </DOC_DEE>
  <DOC_DEE> Kiosques à musique -- Pau (Pyrénées-
    Atlantiques) -- Cartes postales -- 20 siècle
  </DOC_DEE>
<DOC_LANGUE>Français</DOC_LANGUE>
<DOC_REF>PHO00002843</DOC_REF>
<DOC_TITRE> Pau: Place Royale  - Statue Henri IV
</DOC_TITRE>
<DOC_TYPE>Carte postale</DOC_TYPE>
...
</NOTICE>
```

Travail d'indexation

Extraction automatique

Exemple dans un processus complet de traitement du document

Extraction automatique

- Capter les termes qui représentent le mieux le sens des documents traités pour les exploiter dans un domaine applicatif
- Le plus automatisé possible pour traiter un volume important de données dans des temps raisonnables
- Génération de bruit et de silence

I. Introduction

1. Définitions

2. Approches manuelles versus Approches automatiques

3. Approches auto.: statistique versus linguistique

4. Apport des ressources structurées

Approches statistiques (dites quantitatives)

Ce type d'approches repère les régularités statistiques d'une langue

- Les données sont classées, comptées, résumées avec des statistiques
- La fréquence des données a une grande importance
- Les données à basse fréquence sont souvent considérées comme moins importantes
- Les données sont considérées comme des échantillons, on en tire des généralisations censées être valables pour l'ensemble de la population

Approches linguistiques (dites qualitatives)

L'objectif est d'étudier le plus précisément possible un phénomène linguistique (comme le rattachement prépositionnel):

- on veut comprendre les processus linguistiques profonds sous-jacents
- on donne la même importance à toutes les unités
- on peut s'intéresser à des phénomènes très subtils, peu fréquents
- on travaille généralement sur un nombre limité d'exemples
- Les conclusions tirées d'un échantillon qualitatif ne s'appliquent pas à toute la population avec certitude, car on ne recherche pas des exemples représentatifs de la population

Statistiques versus linguistiques avec patrons

- Les arguments habituels…:
  - développement de patrons par des experts long et coûteux
  - traitement statistique a besoin de nombreux exemples annotés à la main par des experts... ce qui est long et coûteux
  - Le problème est d'autant plus important que les systèmes sont dédiés à une application donnée

Statistiques versus linguistiques avec patrons

- Les arguments habituels…:
  - développement de patrons par des experts long et coûteux
  - traitement statistique a besoin de nombreux exemples annotés à la main par des experts... ce qui est long et coûteux
  - Le problème est d'autant plus important que les systèmes sont dédiés à une application donnée
- Les voies hybrides ont vite été explorées
  - angles de vues complémentaires sur les objets linguistiques.
  - De nombreux travaux effectuent des allers-retours permanents entre théorie et expérimentation. Les modèles théoriques permettent de guider l'expérimentation. Les résultats obtenus permettent de valider, invalider ou corriger les modèles théoriques

I. Introduction

1. Définitions

2. Approches manuelles versus Approches automatiques

3. Approches auto.: statistiques versus linguistiques

4. Apport des ressources structurées

UNIVERSITÉ MONTPELLIER 2
SCIENCES ET TECHNIQUES

Apport des ressources structurées (1/3)

- **Les vocabulaire ouverts** : « indexation personnelle »

  un exemple : folksonomie (Vander Wal , 2007)

  → Problèmes : difficiles a organiser et a maintenir

Apport des ressources structurées (1/3)

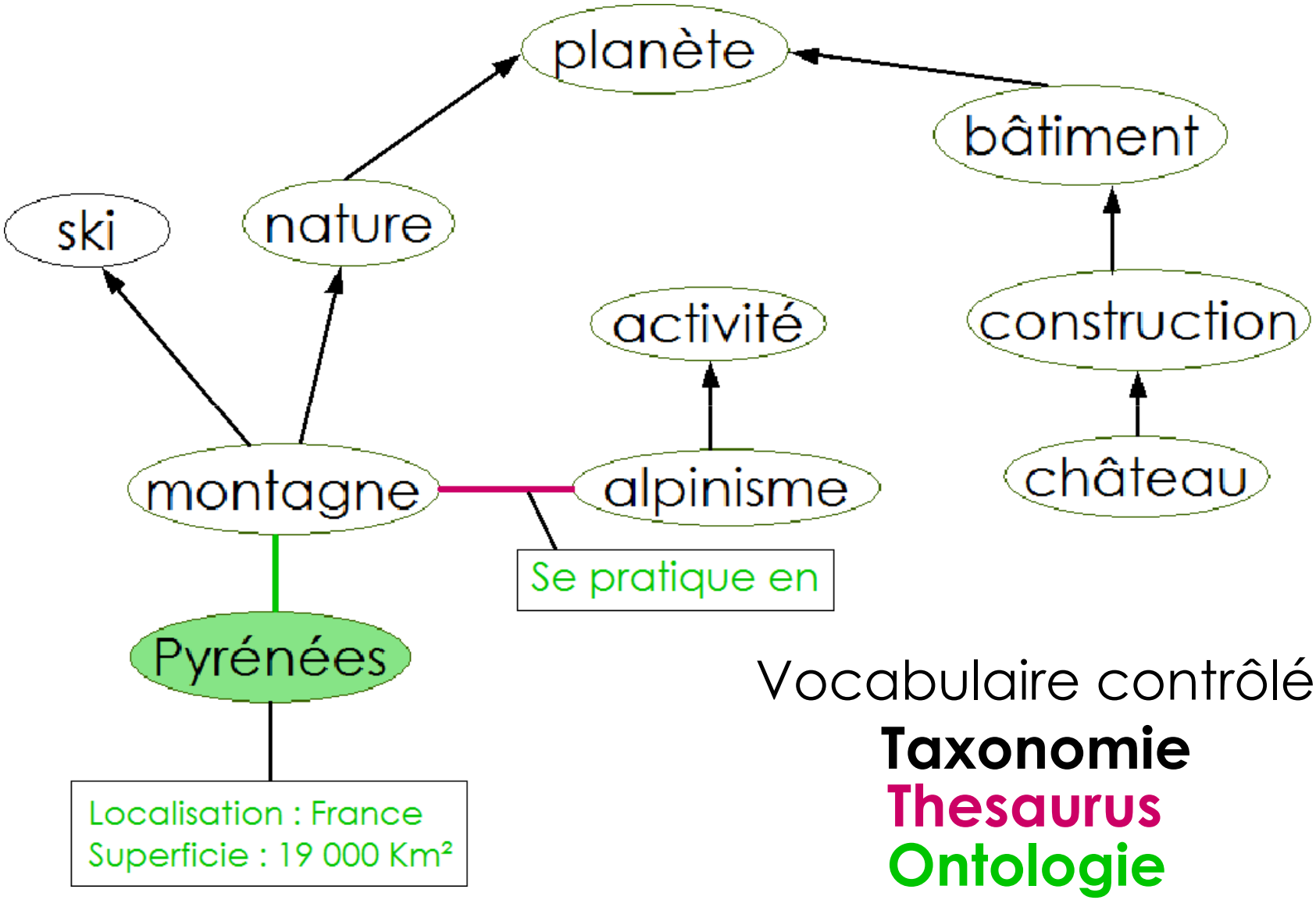
- **Les vocabulaire ouverts** : « indexation personnelle »

  un exemple : folksonomie (Vander Wal , 2007)

  → Problèmes : difficiles a organiser et a maintenir

- **Les vocabulaires contrôlés :** ensemble limite de termes défini par une communauté de pratiques afin de pouvoir annoter des contenus
  - Définis de façon précise dans les centres documentaires (normes NF Z 47-100 16 et ISO 2788):
    - Règles terminologiques (langue, singulier/pluriel) et syntaxiques (pour décrire des sujets complexes)
    - Règles sémantiques :
      - Un terme n'a qu'un seul sens précis dans le vocabulaire d'indexation donné et cet terme est le seul a avoir ce sens (contrôle de la synonymie et de la polysémie)
      - Contrôle de l'homonymie (utilisation de qualificatifs, adjectifs complémentaires, notes d'application, etc.)
      - Possibilité de structurer les termes par des relations sémantiques hiérarchiques et associatives

Apport des ressources structurées (2/3)



Vocabulaire contrôlé

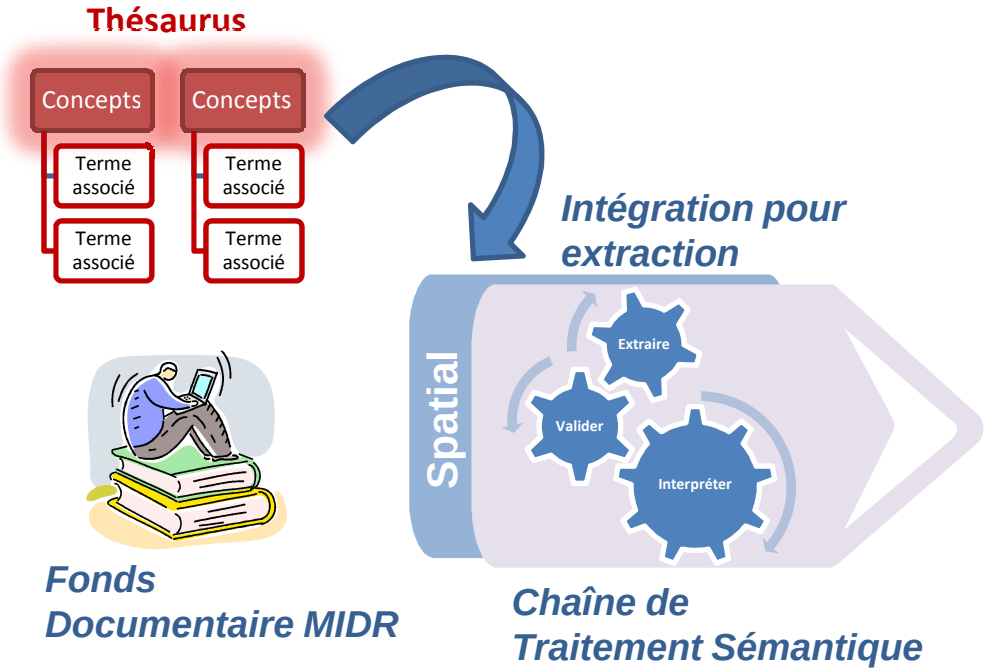
Taxonomie

Thesaurus

Ontologie

Apport des ressources structurées (3/3)

→ La connaissance experte du domaine aide à l'identification termes et des relations



Toponyme	Terme	Fréquence
Artouste	lac	2
Artouste	gave	2
Artouste	montagne	1
Artouste	pic	3

Liste termes indexés

Limite:

- Vocabulaire contrôlé construit en amont

II. Le Traitement Automatique du Langage Naturel

1. **Emergence d'une chaine standardisée**
   a. **Point de départ**
   b. **Chaine standard**

2. Extraction d'entités nommées notées (Ens)

3. Les boîtes à outils pour le TALN

UNIVERSITÉ MONTPELLIER 2
SCIENCES ET TECHNIQUES

Point de départ

- ce qu'on a dans les documents, ce sont des mots
  - approche « sac de mots » (bag of words) : on prend les mots en vrac
    - tokenisation;
    - espace, apostrophes, tirets, points...
    - approches à base de dictionnaires
    - etc.
  - on oublie la séquentialité : on considère que les mots sont indépendants

notion de déséquentialisation: les informations contenues dans la séquence des mots ou dans les mots outils sont reportées sur les unités d'information conservées :

Ex : L'homme ferme la porte de la tour

homme: nom masculin singulier déterminé
porte: nom féminin singulier déterminé
tour: nom féminin singulier déterminé
fermer: verbe présent de l'indicatif 3e pers. du singulier
SUJET(homme,fermer)
OBJET(fermer,tour)
COMPLEMENT_DE_NOM(tour,porte)

Chaine standard de TALN : un exemple

Une chaîne de traitements ressort d'un ensemble de travaux s'appliquant sur l'analyse linguistique [Abolhassani et al., 2003 ; Bilhaut, 2006 ; Lesbegueries, 2007], se décomposant selon les sous-processus d'analyse et d'extraction d'information suivants :

Chaine standard de TALN : un exemple

Une chaîne de traitements ressort d'un ensemble de travaux s'appliquant sur l'analyse linguistique [Abolhassani et al., 2003 ; Bilhaut, 2006 ; Lesbegueries, 2007], se décomposant selon les sous-processus d'analyse et d'extraction d'information suivants :

1. la lemmatisation : segmentation des mots et identification de leur lemme ;
2. l'analyse lexicale et morphologique : la reconnaissance des mots ;
3. l'analyse syntaxique : utilisation de grammaires pour trouver les relations entre les mots →permet d'identifier le rôle des termes ou des syntagmes dans la phrase ;
4. l'analyse sémantique : interprétation plus spécifique sur les syntagmes retenus : l'objectif est ici d'identifier le sens potentiel véhiculé par un mot ou un groupe de mots.

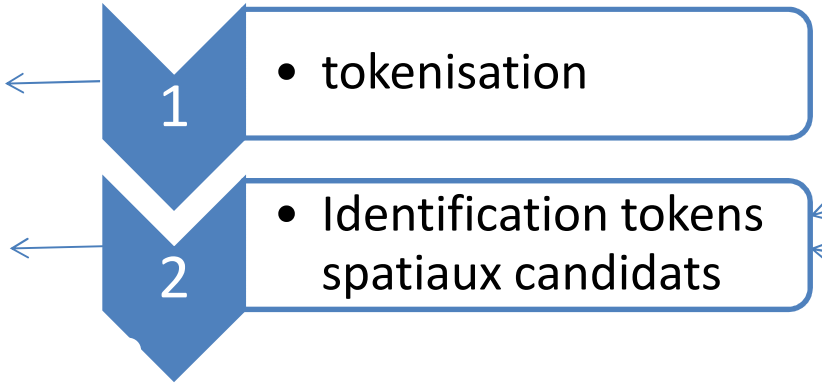
Chaine standard de TALN : un exemple

"… la ville de Montpellier…"

"… /la/ville/de/Montpellier/ …"

"/la/ville/de/Montpellier/"

type: geo
stype: S



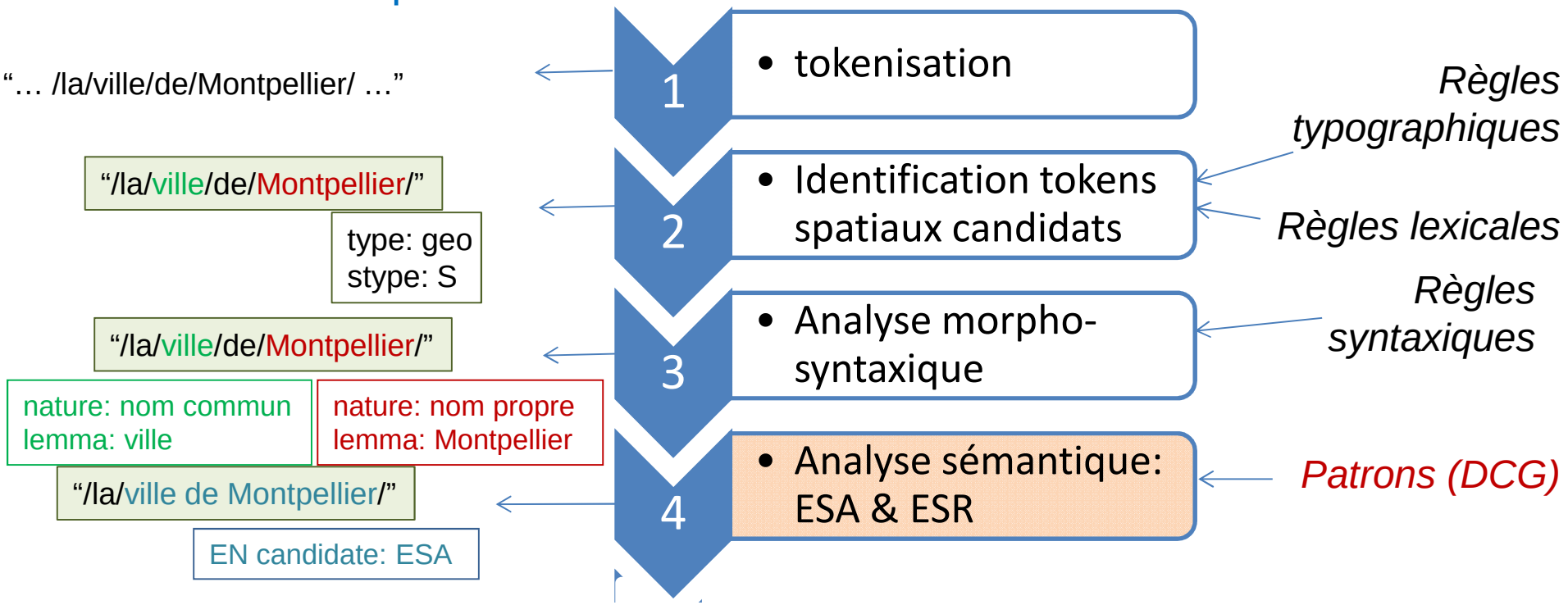
Règles typographiques

Règles lexicales

Liste Nom toponymique:
- Montpellier
- Toulouse
- Nancy
- etc.

Chaine standard de TALN : un exemple

"… la ville de Montpellier…"



II. Le Traitement Automatique du Langage Naturel

1. Emergence d'une chaine standardisée

2. **Extraction d'entités nommées (Ens)**
   a. **Définitions**
   b. **Applications**
   C. **Classes des Ens**
   d. **Identification des Ens**

3. Les boîtes à outils pour le TALN

UNIVERSITÉ MONTPELLIER 2
SCIENCES ET TECHNIQUES

Définitions (1/3)

- Entités nommées :
  - Unités lexicales particulières

  Ex : noms de personnes, noms d'organisation, noms de lieux... date, unités monétaires, pourcentages...

- Reconnaissance des entités nommées :
  - Identifier ces unités dans un texte
  - Les catégoriser
  - Éventuellement, les normaliser

Définitions (2/3)

L'ancien premier ministre socialiste Lionel Jospin a confirmé, jeudi 2 8 septembre, sur RTL, qu'il ne sera pas candidat à l'investiture socialiste pour la présidentielle de 2007.

- Identification : Lionel Jospin, jeudi 28 septembre, RTL, 2007.
- Catégorisation :

  L'ancien premier ministre socialiste <PERS>Lionel Jospin </PERS> a confirmé, <DATE>jeudi 28 septembre</DATE>, sur <ORG>RTL</ORG>, qu'il ne sera pas candidat à l'investiture socialiste pour la présidentielle de <DATE>2007 </DATE>.
- Normalisation : L. Jospin → Lionel Jospin

Définitions (3/3)

- Plus de finesse ?

<PERS><FONCTION>L'ancien premier ministre socialiste</FONCTION>Lionel Jospin</PERS> a confirmé, <DATE val="2006-09-28">jeudi 28 septembre</DATE>, sur <ORG type="radio">RTL</ORG>, qu'il ne sera pas candidat à l'investiture socialiste pour la présidentielle de <DATE val="2007">2007</DATE>.

- Le niveau dépend des capacités du systèmes mais aussi de l'application
- La reconnaissance d'entités nommées est issue de la tâche plus générale de l'extraction d'information

Applications « internes » (1/2)

- Analyse syntaxique
  - Aide à la segmentation et à la morphosyntaxe
    - HyOx, Inc.
    - Seat and Porsche had fewer registration in July 1996.
  - Aide à l'analyse syntaxique
    - He will be replaced by Eliahu Ben-Alissar, a former israeli envoy to politician.

    <LOC>Egypt</LOC> and <LOC>Jordan</LOC>.
    - He will be replaced by Eliahu Ben-Alissar, a former israeli envoy to politician.

    <LOC>Egypt</LOC> and <ORG>Likud party</ORG> politician.
  - Acquisition de dépendances "sémantiques"
    - They met in <LOC>Baghdad</LOC>

Applications « internes » (2/2)

- Coréférence
  - <PERS>John</PERS> bought a new computer. It was able to process XML.

- Traduction
  - <PERS>Jack London</PERS> was an American writer
    - ▸ Jack London était un auteur américain.
  - <LOC>London</LOC> is where I lived my best years.
    - ▸ C'est à Londres que j'ai vécu mes meilleures années.

- Désambiguïsation lexicale

Applications « directes »

- L'extraction d'information et la veille
  - Remplir des bases de données sur une entité ou un type d'entités donnés
  - Signaler de nouveaux documents concernant cette entité ou ce type d'entités

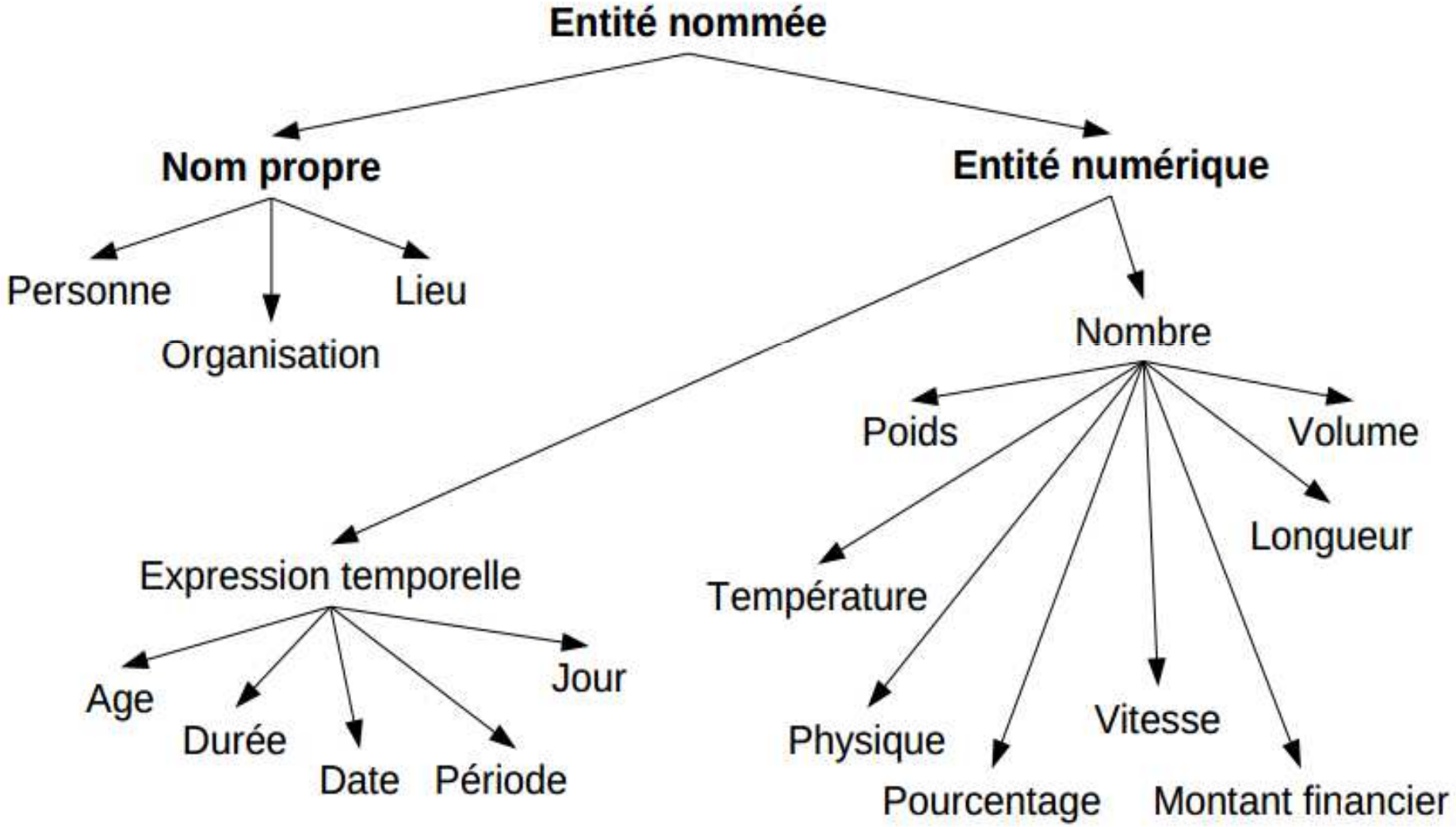
- La tâche de questions-réponses
  - Permet d'identifier le type de réponse attendu
  - cf. cours questions-réponses

- L'anonymisation

Le choix des classes

- Gouvernées au départ par les conférences MUC...
- Les valeurs sûres ("ENAMEX") :
  - ORGANISATION
  - LIEU
  - PERSONNE
- Celles qui reviennent souvent :
  - TIMEX (date, expressions temporelles)
  - NUMEX (valeur monétaire, pourcentage...)
- On peut ajouter des classes et les subdiviser à l'infini
- Certaines applications nécessitent une granularité hétérogène

Un exemple de hiérarchie (QALC, LIMSI)



Identification des Ens : preuves internes

- Majuscule (à manipuler avec précaution)
- Prénoms ou marqueurs générationnels (personnes) :
  - Lionel Jospin
  - L. Jospin
  - Benoît XVI
- Mots ou affixes de type classifiant (lieux et organisations) :
  - la Banque Populaire
  - Microsoft Inc.
  - L'avenue des Champs-Elysées
  - le Mont Valérien
- Sigles ou esperluettes (organisations) :
  - Crédit Agricole SA

Identification des Ens : preuves externes

- Contexte d'apparition des entités nommées

- Informations supplémentaires ou propriétés spécifiques (titre, grade , etc.)
  - Monsieur Jospin
  - Mme Denise
  - Général Leclerc
  - le groupe Sanofi-Aventis
  - the Coca-Cola company

- Souvent précisées lors de la première occurrence de l'EN dans le texte, d'où l'importance d'une propagation de ces informations

Identification des Ens : utilisation des lexiques

- En général, de simples listes de mots :
  - liste de prénoms
  - liste de villes, pays, fleuves...
  - liste de métiers
  - liste de marques
  - ...
- Construits manuellement ou automatiquement
- Très utiles pour les noms de lieux, d'intérêt discuté pour les autres
- Confrontés au problème classique des ressources : pas assez de mots, c'est inutile ; trop de mots, c'est ambigu

Annotation des Ens (1/2)

- Méthodes symboliques :
  - À base de règles contextuelles
  - Patrons d'extraction écrits à la main
  - Exploitation des informations :
    - Morphosyntaxiques
    - Lexicales (issues des lexiques)
  - Exemples :
    - Prénom + Mot avec une majuscule = Personne
    - Mot inconnu + "Inc." = Organisation
    - Nom Propre + '&' + Nom Propre = Organisation
    - Lieu + verbe d'action = Organisation

Annotation des Ens (2/2)

- Méthodes à base d'apprentissage :
  - Résultat : des règles logiques, un arbre de décision, un modèle numérique...
  - Nécessitent de larges corpus annotés
  - Il n'est pas toujours possible d'intervenir sur les résultats après coup

- Approches mixtes
  - Apprentissage de règles puis révision par un expert
  - Élaboration de règles par un expert puis extension automatique de la couverture

- Performances comparables mais avantages et inconvénients de chacun à prendre en compte

II. Le Traitement Automatique du Langage Naturel

1. Emergence d'une chaine standardisée

2. Extraction d'entités nommées (Ens)

3. Les boîtes à outils pour le TALN

   a. **comparatifs**

   b. **premières conclusions**

UNIVERSITÉ MONTPELLIER 2
SCIENCES ET TECHNIQUES

Comparatifs des outils complets

	UIMA	GATE	UNITEX	Linguastream	Service Web Calais	Apache Stanbol
Documentation	Bonne	Bonne	Moyenne	Moyenne	Bonne	Bonne
Architecture	Bien définie	Bien définie	Bien définie	Bien définie	Service Web	Bien définie
Technologies/Langages de programmation	Java, C++	Java	Interfaces Java, modules C	JAVA	API Java pour l'interrogation	Java
Mesures de performance	Pas d'indication	Pas d'indication	Pas d'indication	OK	Limité en taille (100000docs/jr)	Pas d'indication
Algorithmes et Techniques implémentées	Moyenne	Très riche	Riche	Moyenne (à définir)	Riche	Riche
Limitations	Aucune	Aucune	Aucune	Aucune	Volume de données	Catégories des entités nommées et langues
Types de documents traités	Tout	Texte	Texte	Texte	Texte	Texte
Correcteurs orthographique, syntaxique, grammatical	Pas de module prédéfini ; projet GramLab pour la correction orthographique en français et anglais	Pas de module prédéfini	Pas de module prédéfini	Pas de module prédéfini	Pas de module prédéfini	Pas de module prédéfini
Extraction et Catégorisation des ENs	Nécessaire d'ajouter un module NER (OpenNLP, Alchemy)	ENs prédéfinies : temporelles, Lieux, Personnes, Quantités, Compagnie, Section, Document. Possibilité de créer de nouvelles catégories d'ENs	ENAMEX (personnes, organisations, lieux, nationalités, titres, et faits), TIMEX (expressions de temps et de dates) et NUMEX (nombres, pourcentages, quantités monétaires)	Pas de module prédéfini Possibilité de créer des catégories d'entités nommées et règles adaptées	IPTC News codes: Finance, catastrophes, Education, culturel, Environnement, Sports, Politiques, intérêts humains, etc.	Lieux, Personnes et Organisation. Possibilité de prendre des vocabulaires locaux. A voir s'il est possible d'enrichir par de nouvelles catégories
Gestion des langues	Géré dans le module intégré pour le TALN	**Espagnol**, le chinois, l'arabe, le **français**, l'**allemand**, l'hindi, le cebuano, le roumain, le russe	Dictionnaires et Lexicon-grammar tables disponibles en langue française, anglaise, grecque, allemande, espagnole, italienne, etc.	Treetagger disponible en français, anglais, espagnol et allemand. Règles de marquage à redéfinir pour chaque langue	Anglais uniquement	Anglais, Allemand, Danois, Suédois, Hollandais, et Portugais. L'outil est en cours d'évolution (vérifier les langues Français, Espagnol et Italien.
Disponibilité/Licence	licence **Apache** (licence de logiciel libre et *open source*)	Logiciel libre	licence libre LGPL	Licence Open Source	Open source pour une utilisation limitée (50000 documents), payant sinon	Licence Open Source

Comparatifs des solutions légères

	OpenNLP	Stanford NER	SxPipe 2	NEXT	ABNER
Documentation	Moyenne	Moyenne	Limitée	Limitée	Limitée
Architecture	API : Absence d'architecture globale	API : Absence d'architecture globale	API : Absence d'architecture globale	Module Perl, pas d'architecture globale	API Java : pas d'architecture globale
Technologies/Langages de programmation	Java, Python	API Java	API Java	Perl	Java
Mesures de performance	Pas d'indication	Pas d'indication	Pas d'indication	Pas d'indication	Bonne uniquement dans le domaine biomédical
Algorithmes et Techniques implémentées	Riche mais non intégré	Moyen	Riche	Moyen	Moyen
Limitations	Aucune	Langue	Langue	Langue	Extracteur d'entités nommées uniquement et dans domaine d'application biomédical
Types de documents	Texte	Texte	Texte	Texte	Texte
Correcteurs orthographique, syntaxique, grammatical	Pas de module prédéfini	Pas de module prédéfini	module unifié de correction orthographique et de reconnaissance de formes composées	Pas de module prédéfini	Pas de module général pour la correction
Extraction et Catégorisation des ENs	2 modules : dates, lieux, personnes, expressions de temps, organisations, quantités, valeurs monétaires, pourcentages, etc. Gestion des langues	Personnes, Lieux et Organisations	Module NP pour l'extraction et la catégorisation des ENs (lieux, des organisations, des personnes, des entreprises, des produits et des oeuvres (titres de livres, de films, etc.)	Version initiale MUC/MET (MUC7) : Personne, Organisations, Lieux, Date, expression temporelle, Monnaies et Pourcentage. Possible de définir de nouveaux types d'entités nommées et de les appliquer à des entités nommées identifiées.	Pas de module général pour l'extraction et la catégorisation d'ENs
Gestion des langues	modèles de langues sont définies pour l'anglais, l'allemand, le suédois, le danois et le portugais. Des possibilités existent pour le français	Anglais	Français, polonais, anglais, espagnol, slovaque et slovène	Analyseur morphologiques japonais (ChaSen JUMAN et sur les analyseurs syntaxiques CaboCha et KNP. **Pas de module pour traiter le français, l'espagnol, l'allemand.**	Anglais
Disponibilité/Licence	licence LGPL	Licence libre pour la recherche et commerciale sinon	Licence LGPL	logiciel libre	Licence libre

SENTERRITOIRE : Identification automatique de sentiments liés aux territoires

Journées JiGOT – 11/2012

E. Kergosien (IRSTEA - LIRMM)

eric.kergosien@lirmm.fr

Maguelonne Teisseire (IRSTEA – LIRMM)

Sandra Bringay (LIRMM)

Mathieu Roche (LIRMM)

Objectif général

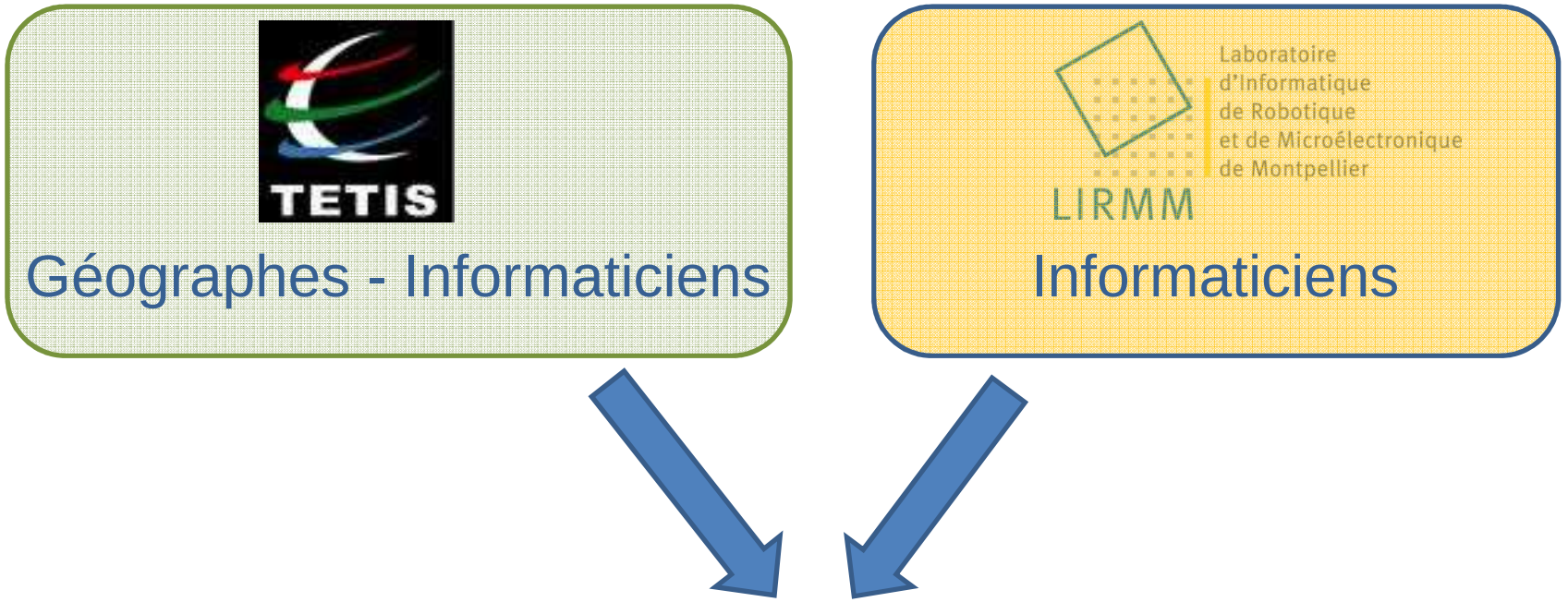
Documents d'actualités

Objectif général

Documents d'actualités

La perception de l'aménagement d'un territoire

Objectif général



→ proposer un environnement décisionnel basé sur une analyse automatique des textes liés à l'aménagement du territoire

I. Introduction

1. Problématiques
   a. Identifier les informations propres à un territoire : restriction à l'entité spatiale;
   b. Comment extraire les entités spatiales (ES) d'un texte?

2. Démarche générale

UNIVERSITÉ MONTPELLIER 2
SCIENCES ET TECHNIQUES

Restriction de la notion de territoire

- Définition complexe sujette à discussion (Di Meo, 1998)

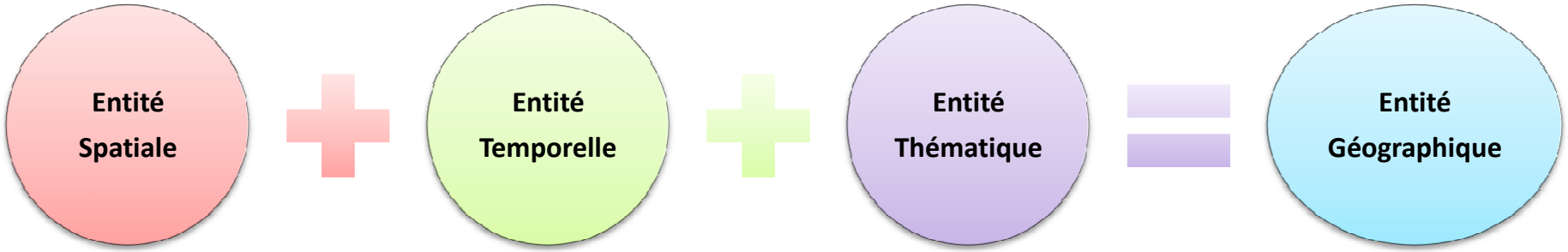
  Appropriation à la fois économique, idéologique et politique (sociale donc) de l'espace par des groupes qui se donnent une représentation particulière d'eux-mêmes, de leur histoire, de leur singularité

Restriction de la notion de territoire

- Définition complexe sujette à discussion (Di Meo, 1998)

Appropriation à la fois économique, idéologique et politique (sociale donc) de l'espace par des groupes qui se donnent une représentation particulière d'eux-mêmes, de leur histoire, de leur singularité

ENTITÉ GÉOGRAPHIQUE VS ENTITÉ SPATIALE (Usery, 2000)



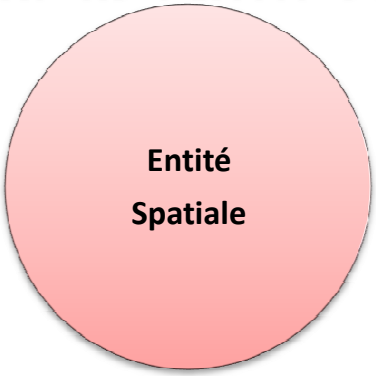
Les instruments de musique dans les environs de Montpellier au XIXe siècle

Restriction de la notion de territoire

- Définition complexe sujette à discussion (Di Meo, 1998)

Appropriation à la fois économique, idéologique et politique (sociale donc) de l'espace par des groupes qui se donnent une représentation particulière d'eux-mêmes, de leur histoire, de leur singularité

ENTITÉ GÉOGRAPHIQUE VS ENTITÉ SPATIALE (Usery, 2000)



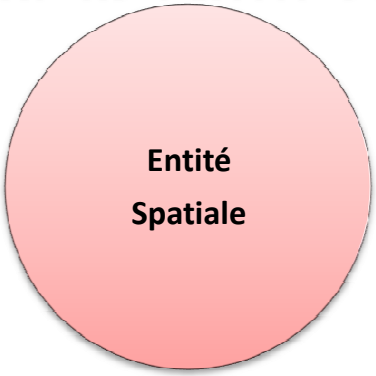
Les instruments de musique dans les environs de Montpellier au XIXe siècle

Restriction de la notion de territoire

- Définition complexe sujette à discussion (Di Meo, 1998)

Appropriation à la fois économique, idéologique et politique (sociale donc) de l'espace par des groupes qui se donnent une représentation particulière d'eux-mêmes, de leur histoire, de leur singularité

ENTITÉ GÉOGRAPHIQUE VS ENTITÉ SPATIALE (Usery, 2000)



Les instruments de musique dans les environs de Montpellier au XIXe siècle

PROBLÉMATIQUE

Quelle est la forme de l'entité spatiale dans les textes et comment l'extraire?

Notions d'opinion & sentiment

- Opinion : négatif; neutre ; positif

Notions d'opinion & sentiment

- Opinion : négatif ; neutre ; positif

- Sentiment: lexiques de sentiments

Notions d'opinion & sentiment

- Opinion : négatif; neutre ; positif

- Sentiment: lexiques de sentiments

PROBLÉMATIQUE

Comment identifier et extraire l'opinion (puis le sentiment) dans les textes?

Démarche générale

Documents d'actualités

1ère Phase

Extraction des entités spatiales

Démarche générale

Documents d'actualités

2ème Phase

Identification des opinions associées

1ère Phase

Extraction des entités spatiales

Démarche générale

Documents d'actualités

2ème Phase

Identification des opinions associées

1ère Phase

Extraction des entités spatiales

La perception de l'aménagement d'un territoire

II. Notre approche

1. Extraction d'ES : choix du modèle Pivot (Lesbegueries, 2007)
2. Le TAL pour l' extraction d'ES : la chaine PIV
3. Définitions de patrons pour l'extractions d'ES et d'Organisations
4. Expérimentations

UNIVERSITÉ MONTPELLIER 2
SCIENCES ET TECHNIQUES

Extraction d'ES

- Extraction automatique des descripteurs géospatiaux et de leur lien sémantique dans les données textuelles

  - Proposition d'une chaine de traitements linguistiques (Stage Recherche Sabiha Tahrat, Janvier-juin 2012))

    - chaine TAL standard (Abolhassani et al., 2003 ; Bilhaut, 2006)

    - Patrons pour l'extraction d'ES définis sur la base des travaux de (Lesbegueries, 2007) dans le cadre d'un projet nommé PIV

Extraction d'ES : choix du modèle Pivot

- Modéliser les descripteurs géospatiaux et leurs liens sémantique dans les données textuelles

Choix du modèle Pivot (Lesbegueries 2007): définit deux types d'entité spatiales

- Entité spatiale absolue(**ESA**): <(Indicateur spatiale)*, Entité Nommée>
- Entité spatiale relative(**ESR**): <(Relation)+, ESA>; <(Relation)+, ESR>;

Extraction d'ES : choix du modèle Pivot

- Modéliser les descripteurs géospatiaux et leurs liens sémantique dans les données textuelles

Choix du modèle Pivot (Lesbegueries 2007): définit deux types d'entité spatiales

- Entité spatiale absolue(**ESA**): <(Indicateur spatiale)*, Entité Nommée>
- Entité spatiale relative(**ESR**): <(Relation)+, ESA>; <(Relation)+, ESR>;

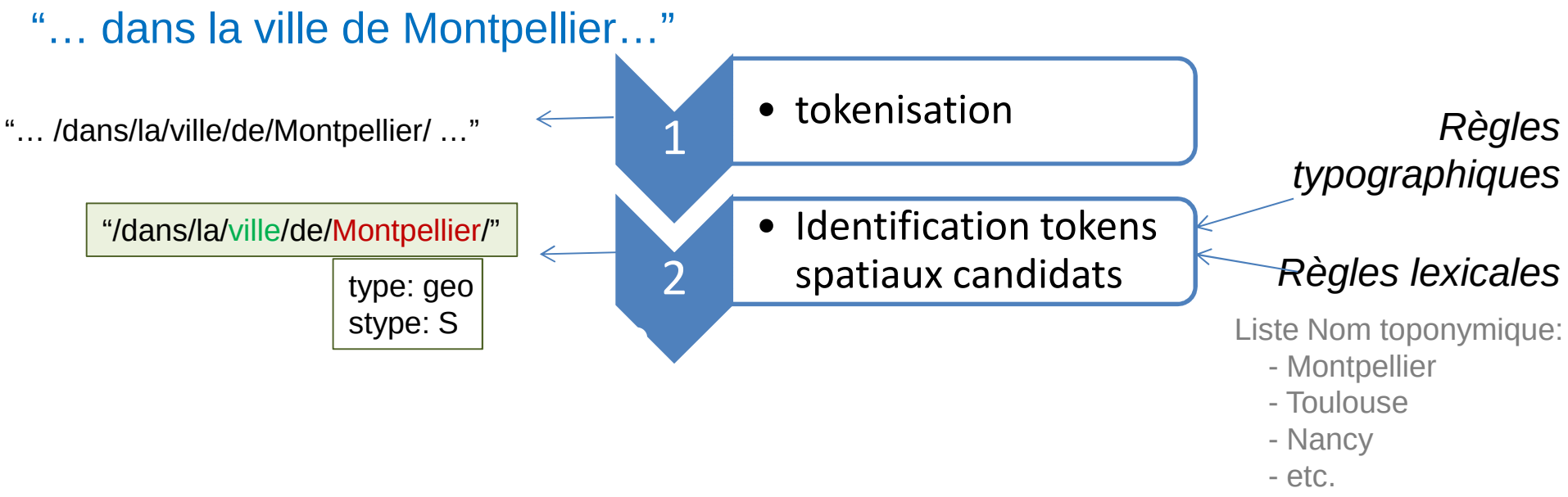
Instances

"southern city of Montpellier"
« le sud de la ville de Montpellier »
Relation: **orientation**
ASF site

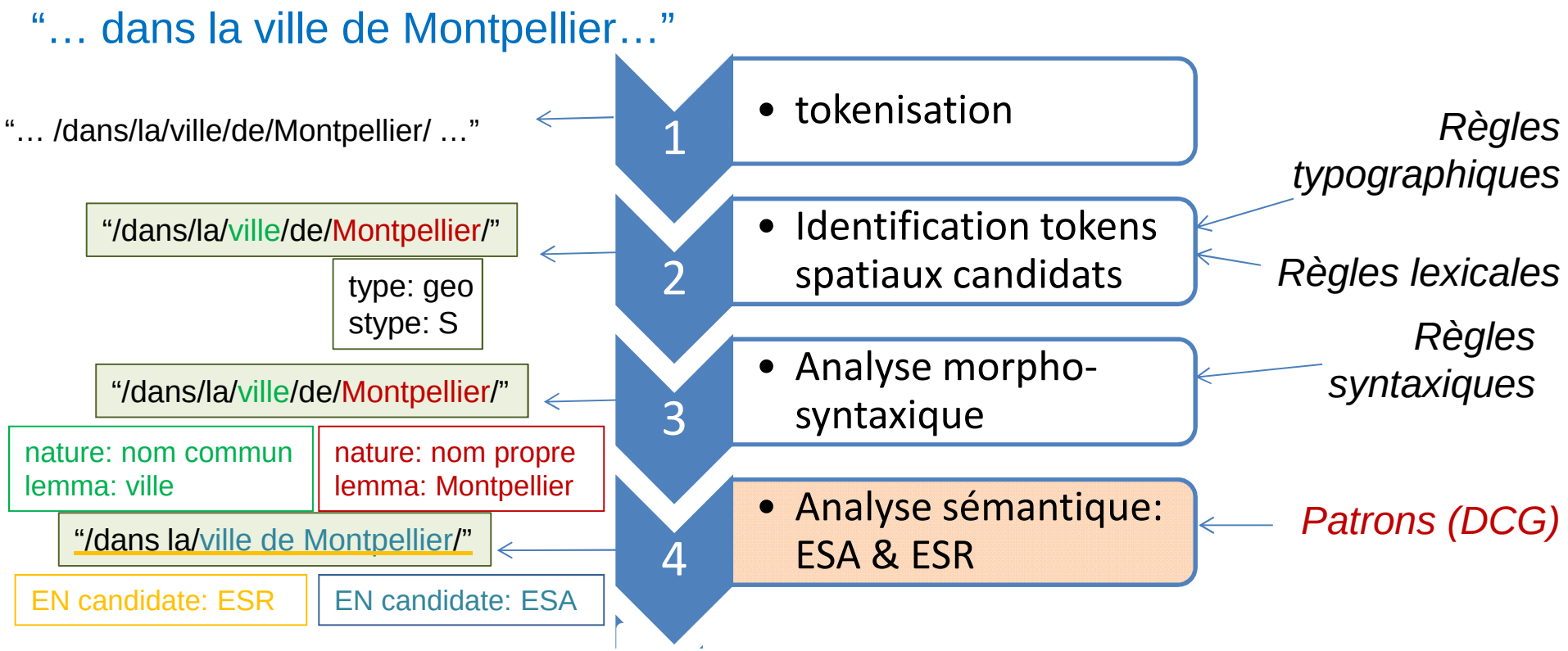
"around Montpellier"
« environs de Montpellier »
Relation: **adjacence**
ASF

"north near Montpellier"
«au nord des environs de Montpellier »
RSF
Relation: **orientation**

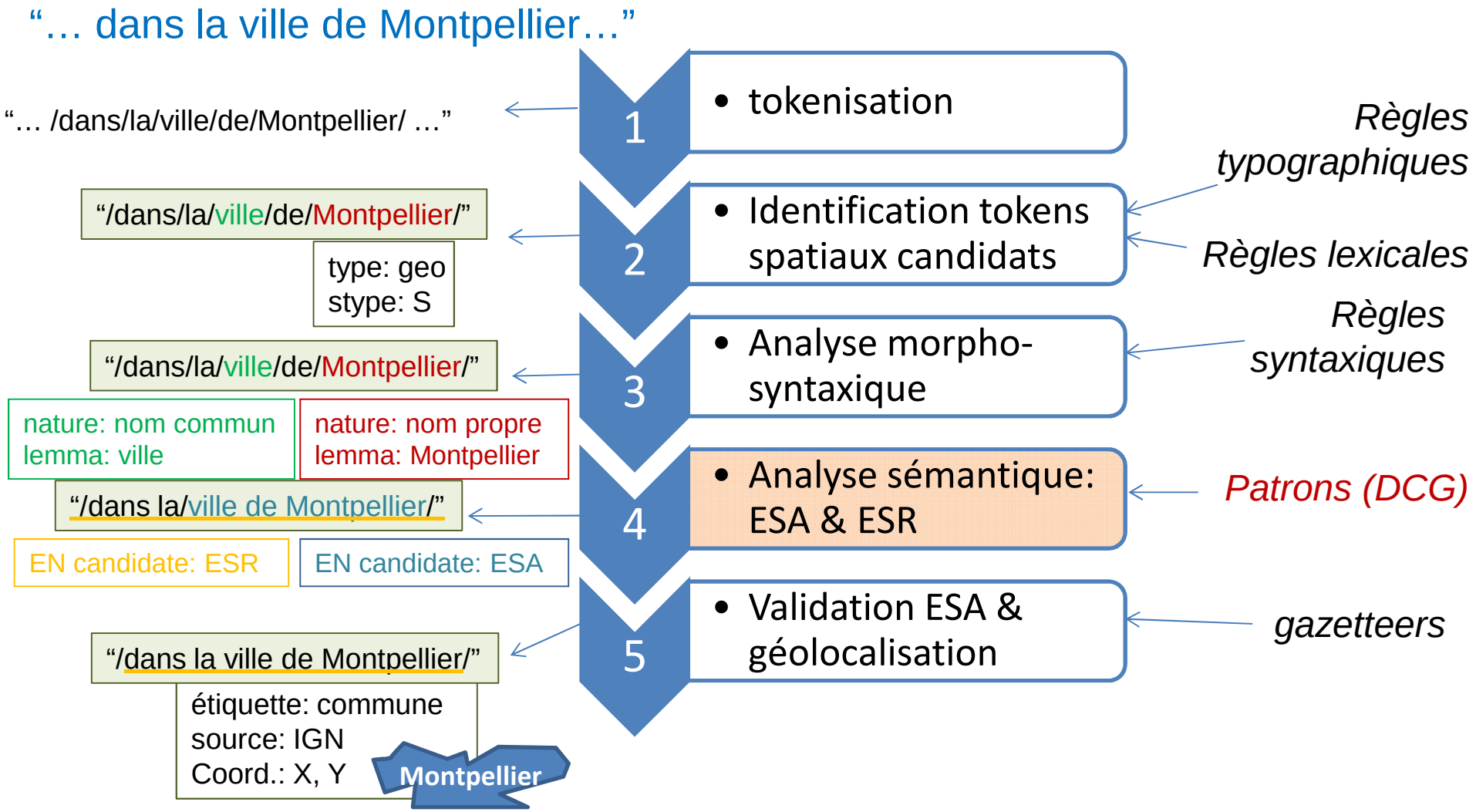
Extraction d'ES : la chaine PIV



Extraction d'ES : la chaine PIV



Extraction d'ES : la chaine PIV



Expérimentations sur la chaine PIV

qualité → $\text{Rappel} = \frac{\text{Nombre de réponses correctes extraites}}{\text{Nombre de réponses correctes existantes}}$

exhaustivité → $\text{Précision} = \frac{\text{Nombre de réponses correctes extraites}}{\text{Nombre de toutes les réponses existantes}}$

$$\text{F-mesure} = 2 \times \frac{\text{Précision} \times \text{Rappel}}{\text{Précision} + \text{Rappel}}$$

Corpus
20 articles journalistiques
8141 mots

Annotation
Automatique & Manuelle

Mesure	ESA	ESR	ORG
Précision	53%	84%	92%
Rappel	94%	66%	35%
F-Mesure	67%	74%	50%

Le sens dépend du contexte

Extraction d'ES : implémentation

- Cette chaine d'extraction a été implémentée et validée sous la plateforme LINGUASTREAM [http://www.linguastream.org/].

  - Collaborations avec M. Loglisci (http://www.di.uniba.it/~loglisci/),

  Perspectives

  - Améliorations à apporter dans la définition des patrons linguistiques pour la prise en compte des composantes phénomènes et temporelles

II. Orientations actuelles & Perspectives

1. Définition et formalisation de la notion de Territoire
2. Enrichissement de la chaine TAL pour l'extraction d'informations temporelles et thématiques
3. Extraction des sentiments liés à un territoire
4. Viewer Senterritoire pour les décideurs

UNIVERSITÉ MONTPELLIER 2
SCIENCES ET TECHNIQUES

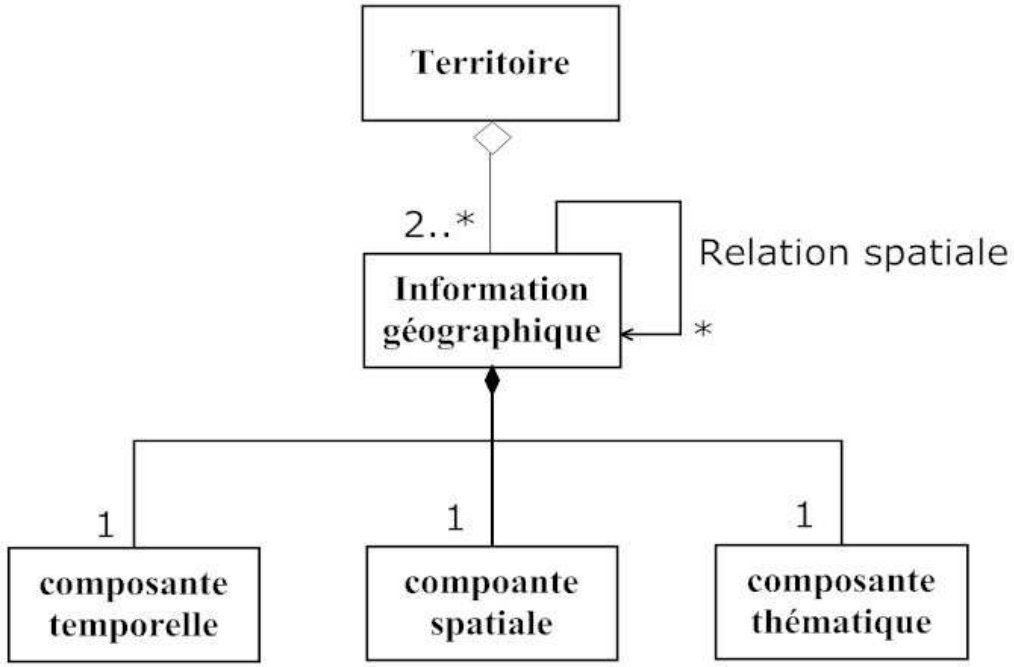
Définition et formalisation de la notion de Territoire

- S'appuyer sur travaux des géographes

(Piolle, 1992 ; Di Méo et al., 2005 ; Guichard, 2007, etc.)

→ *Continuer un travail initié avec les géographes pour proposer une définition que l'on puisse formaliser dans le cadre de Senterritoire:*

Territoire = un ensemble de lieux, de relations spatiales et temporelles mis en évidence par un ensemble de faits



Extraire les informations associées à un territoire

- Possibilité d'automatiser l'extraction d'informations relatant un territoire dans les documents texte :
  - prise en compte des entités thématiques et temporelles
  - Mise en relation des entités géographiques

Extraire les opinions/sentiments associées à un territoire

- Définition du concept de sentiment lié aux données textuelles territoriales
  - *Collaborations avec l'équipe TAMALE (The Text Analysis and Machine Learning Group) de l'Université d'Ottawa (Canada) qui travaillent sur ce sujet : projet 'Mining Public Opinion in Tweets and Other Social Media'*

- Compréhension des perceptions d'un même territoire par les acteurs est difficile
  - Méthode de fouille de textes pour extraire les opinions dans les textes (par apprentissage à partir des corpus *DEFT 2007* et *Bassin de Thau*)
  - Méthode de fouille de textes pour extraire les sentiments dans les textes :
    - déterminer les opinions positives et/ou négatives présentes dans les données textuelles [Roche and Poncelet 2009]

Viewer Senterritoire pour les décideurs

Documents d'actualités

Merci

pour votre attention

Eric KERGOSIEN