Indexer les entités nommées dans une volumineuse bibliothèque numérique scientifique multidisciplinaire

Anubhav GUPTA*† et Denis MAUREL†

† *LI, Université François-Rabelais, France*

* *DIST, CNRS, France*

Contenu

- Istex
- Définition du service proposé
- Les graphes et l'évaluation
- Passage à l'échelle

Université François-Rabelais Tours

Laboratoire d'informatique
EA 6300

Une volumineuse bibliothèque numérique scientifique multidisciplinaire

ISTEX

Istex

- Istex
  - Une immense base de connaissance, de 18,2 millions d'articles scientifiques
  - Avec une couverture multidisciplinaire
    - Sciences "dures" : biologie, médecine, physique, chimie, astronomie, mathématiques…
    - Sciences humaines et sociales : sociologie, histoire, géographie, alimentation…

Istex

- Istex
  - Un défi : interroger cette base
    - Méthodes documentaires
      - titre, auteurs, résumés, mots clés…
      - Recherche trop restreinte
    - Moteurs de recherche
      - ensemble du texte
      - Recherche trop large

Istex

- Istex
  - Des projets à valeurs ajoutées pour améliorer l'interrogation de la base
  - Parmi eux, Istex-Entités nommées, pour une interrogation sur les noms propres et les dates contenues dans les articles
    - "Washington - nom de personne"
    - "2005 - date"

Istex-Entités nommées

DÉFINITION DU SERVICE PROPOSÉ

Quelques remarques

1. Les noms d'université, de centre de recherche, de laboratoire ne figurent pas dans les mots-clés, même si les affiliations des auteurs sont dans les signatures
2. De même pour les noms de projets qui apparaissent parfois en note ou en remerciements

Quelques remarques

3. Le lieu où est réalisé une expérience n'est pas forcément l'adresse du laboratoire
4. Les dates des expériences ne correspondent pas à celle de parution de l'article

Quelques remarques

5. Les noms de chercheur cités ont une importance, alors que souvent la bibliographie indique plusieurs personnes comme signataires d'un article
6. En SHS, des lieux, des institutions, des personnes (avec leur titre ou profession), des dates sont cités, indépendamment du rattachement des auteurs

Notre choix

- Entités choisies
  - personnes
  - lieux
    - administratifs et géographiques
  - organisations
    - dont les financeurs de projet
    - et les hébergeurs de ressources
  - temps
    - années, décennies, siècles, millénaires
  - références

Notre choix

- Présentation des résultats
  - Plusieurs projets travaillent sur les mêmes textes
    - Impossible d'annoter directement le texte lui-même
  - Pour chaque texte, les informations ont donc été placées dans des fichiers associés
    - Au format TEI-standOff

Notre choix

```
<standOff>
    <teiHeader>
            …
    </teiHeader>
    …
    <listAnnotation type=placeName xml:lang="en">
        <annotationBlock corresp="text" xmls="https://wwww.tei-c.org/ns/1.0">
            <placeName change="#Unitex-3.2.0-alpha" resp="istex-rd"
scheme="http://placename-entity.lod.istex.fr">
                <term>Montréal</term>
                    <fs type="statistics">
                        <f name="frequency">
                            <numeric>1</numeric>
                        </f>
                    </fs>
                </placename>
        <annotationBlock>
    </listAnnotation>
        …
</standOff>
```

Istex-Entités nommées

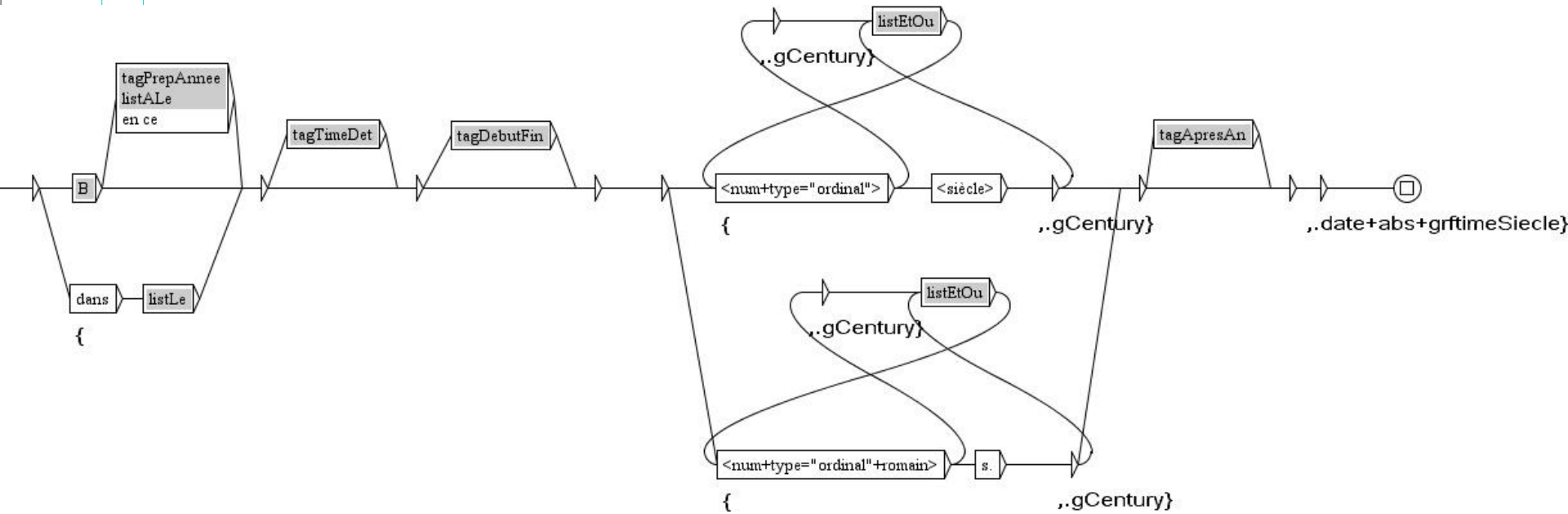
LES GRAPHES ET L'ÉVALUATION

Unitex

- Un logiciel libre d'analyse lexicale
  - Qui allie un système informatique performant
  - Des ressources lexicales
  - et une interface conviviale (des graphes)

Laboratoire d'informatique
EA 6300

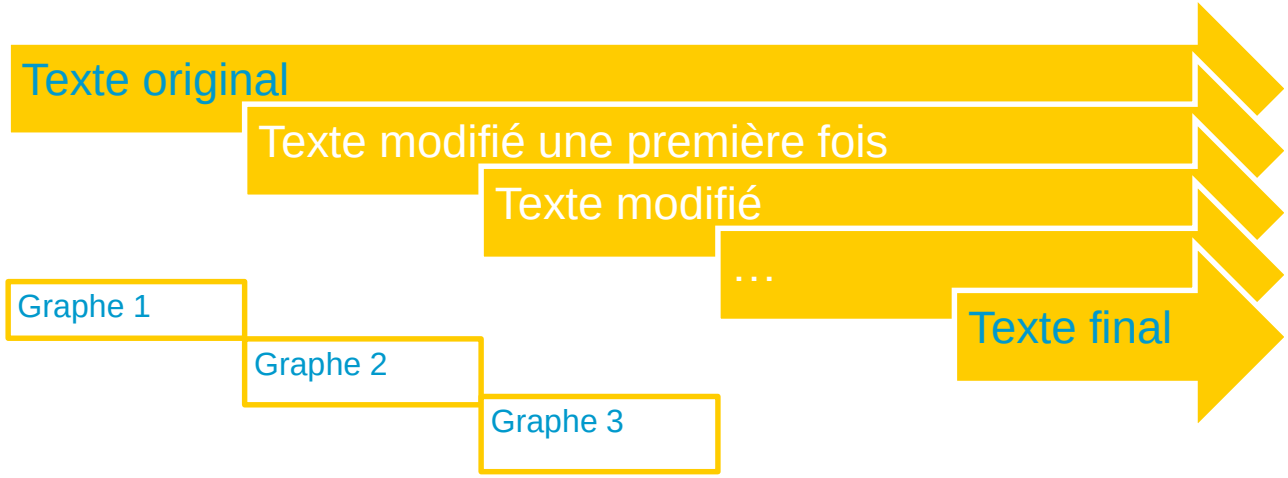
Unitex



À la fin du XXème siècle

Unitex

- Ces graphes sont passés les uns après les autres, en cascade



Unitex

- La cascade hiérarchise les graphes (l'ordre de passage est important)
  - Complémentarité
    - rue du 11 novembre 1918
    - Centre Georges Pompidou
  - Concurrence
    - Il est arrivé le 29 février de l'année 2008
    - Il est arrivé le 29 février de l'année 2008
    - Il est arrivé le 29 février de l'année 2008

Nos cascades

- Anglais
  - cascade de 55 graphes
  - tests réalisés sur 49 documents contenants 5 414 entités nommées
- Français
  - cascade de 130 graphes
  - tests réalisés sur 40 documents contenants 4 695 entités nommées

Évaluation

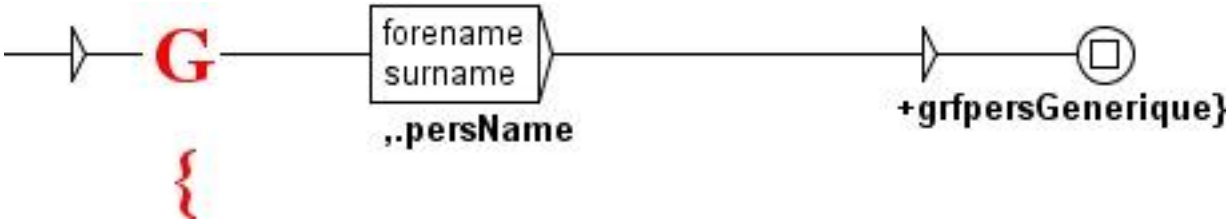
	Anglais	Français
Rappel	55,7%	71,5%
Précision	91,5%	87,1%
Précision du typage	85,6%	84,8%

Istex-Entités nommées

PASSAGE À L'ÉCHELLE

Améliorations logicielles

- Graphes de généralisation d'étiquetage
  - La directrice […], Margaret Somerville, continue [...]
  - Somerville : C'est complexe, mais cela se fait [...].
  - Somerville : L'éthique a besoin de faits précis [...]



Améliorations logicielles

- Passage à l'échelle
  - Optimisation du code Unitex
    - écriture d'un seul fichier en sortie
    - amélioration de la vitesse de traitement
    - scripts de lancement sur plusieurs cœurs de la plateforme Istex

Tâches logicielles

- Passage à l'échelle
  - 998 828 documents
    - temps d'exécution: 2 904 minutes
    - soit une moyenne de 0,17 s par document
  - 10 000 000 documents

Laboratoire d'informatique
EA 6300

Merci !

Les entités nommées

- Les conférences américaines Muc
  - Sur la recherche d'information
    - attentats
    - rachats d'entreprise…
  - Les entités nommées
    - noms propres
    - dates
    - monnaies…

Les entités nommées

- Les conférences américaines Muc
  - Exemple :
    - Migaud : "Il faut trouver de l'ordre de 33 milliards d'euros pour 2013". Premier président de la Cour des comptes et ancien député PS, Didier Migaud a remis...

Extrait du site du journal Le Monde, consulté le 2 juillet 2012

Les entités nommées

- Les conférences américaines Muc
  - Exemple :
    - <person>Migaud</person> : "Il faut trouver de l'ordre de <money>33 milliards d'euros</money> pour <date>2013</date>". Premier président de la <organization>Cour des comptes</organization> et ancien député PS, <person>Didier Migaud</person> a remis...

Extrait du site du journal Le Monde, consulté le 2 juillet 2012

Notre choix

- Dix entités choisies (balises TEI)
  - personnes
    - <persName>
  - lieux
    - administratifs: <placeName>
    - géographiques: <geogName>

Notre choix

- Dix entités choisies (balises TEI)
  - organisations
    - <orgName>
    - financeurs / projets <orgName type="funder">
    - hébergeurs de ressources <orgName type="provider">

Notre choix

- Dix entités choisies (balises TEI)
  - temps
    - années, décennies, siècles, millénaires: <date>
  - références
    - URL: <ref type="url">
    - citations: <ref type="bibl">
    - dans le corps du texte: <bibl>