ISTEX

L'excellence documentaire pour tous

ISTEX, des services offerts aux membres de l'ESR

Jean-Marie Pierrel, Université de Lorraine

Rappel des objectifs d'ISTEX

Création d'une **plateforme nationale** intégrant des collections rétrospectives de la littérature scientifique dans toutes les disciplines :

- **Des acquisitions,** sous forme de licence nationale, de ressources documentaires d'une ampleur inégalée
- **L'agrégation des ressources** au sein d'une plateforme apportant une plus-value basée sur le traitement des données en texte intégral
- **Une plateforme interopérable** avec celles des établissements et organismes du paysage français de l'ESR
- **Une offre de services et usages complémentaires :** traitement des données, extraction de données, fouille de textes, production de synthèses documentaires, de corpus terminologiques …

Un réservoir unique d'archives documentaires scientifiques

- Le programme d'acquisition de ressources concerne **des collections rétrospectives de revues et de livres électroniques** ;
- **L'acquisition du plein texte de ces ressources** permet la mise en œuvre de fouille de textes sur l'ensemble des ressources
  - venant de divers éditeurs,
  - représentant aujourd'hui plus de 21 millions d'articles

Des services liés à l'exploitation de pleins textes

- **Au delà de la recherche documentaire ou bibliographique** de base, ISTEX propose donc des services fondés sur l'**exploitation du plein texte**
  - **Enrichissement des données et des métadonnées**
  - **Extractions de sous-corpus**
  - **Classification incrémentale et cartographie des résultats de recherche**
  - **Mise en œuvre d'outils de fouille de textes (TDM)**

Enrichissement des données et des métadonnées (1)

- **Standardisation** du codages des données : de normes spécifiques à chaque éditeur à un codage unifié **XML/TEI permettant des traitements unifiés sur l'ensemble du corpus**
- **Balisage** fin des principaux champs **des références bibliographiques** de chaque article
  - Pour faciliter une hyper-navigation d'un article vers un article cité
  - Pour rechercher les articles citant un article donné

Enrichissement des données et des métadonnées (2)

- **Détection d'entités nommées dans le plein texte**
  - Noms de personnes, lieux, organisations, financeurs, hébergeur, date, adresse web, etc.
  - Réinjection des résultats dans les métadonnées des articles
- **Détection en plein texte de termes** et de leurs variantes
  - Vers un enrichissement et une mise à jour des métadonnées au vu des évolutions de la science

Classification incrémentale et cartographie des résultats de recherche

- Vu la taille des corpus traités, une requête peut conduire à un volume de résultats difficilement appréhendable par l'utilisateur
- **Deux outils** permettent à l'utilisateur d'avoir une vision plus précise de ces résultats
  - Une **classification incrémentale** de ces résultats
  - Une **cartographie** de ces résultats

Des possibilités d'extraction de sous-corpus en vue d'étude de Text Mining

- **Sélection de sous corpus** par exploitation des métadonnées
  - Par genre, thème, dates, objet d'étude, type de support de publication etc.
- **Possibilité de téléchargement** de ces sous-corpus en vue de traitement locaux de *Text Mining*

ISTEX support de projets de fouille de textes

- Mise en place de projets d'**exploitation du plein texte pour**
  - Démontrer l'**intérêt d'ISTEX dans une optique de Text Mining**
    - Projets de services à valeur ajoutée
    - Chantiers d'usage
  - **Créer une dynamique de recherche développement** autour de la plateforme ISTEX qui puisse servir de déclencheur à des activités plus larges d'appropriation par les chercheurs des contenus d'ISTEX pour développer des recherches en fouille de textes

Pour quels usages ?

- **Interrogation** en texte intégral
  - sur les objets numériques indexés dans leur totalité.
- **Production de synthèses** documentaires
  - par analyse de sous corpus individualisés pour l'occasion, et auxquels sont appliquées des méthodes de text mining.
- **Représentation et visualisation de données**
  - basées sur des technologies de cartographie de la connaissance.
- **Utilisation à des fins de recherche**
  - Par exemple en ingénierie de la langue, génomique, histoire des sciences….

Exemples d'exploitation

- **Sélection de sous corpus d'articles**
  - citant tel auteur, tel article
  - issus de travaux de tel projet (projet Européen, projet ANR, …),
  - s'appuyant sur telle donnée (ou exploitant tel corpus), etc.
- **Caractérisation de l'évolution des recherches ou connaissances**
  - dans un domaine particulier
  - au cours d'une période temporelle donnée.
- **Ré-indexation terminologique** d'articles scientifiques
  - un nouveau concept n'est pas détectable dans les mots clés proposés au sein des métadonnées, mais uniquement par l'analyse du plein texte.
  - Exemples : apparition récente des termes « actif toxique », « nuage informatique » ou « Cloud computing », etc.)

Apport d'ISTEX face à l'existant

- Il y a certes des possibilités de fouille de textes sur les plateformes des éditeurs
- MAIS **ISTEX permet**
  - **de faire de la fouille de textes transverses** sur l'ensemble des ressources acquises et donc pas uniquement celle d'un seul éditeur
  - **de développer son propre système de fouille de textes** sur des sous-corpus d'ISTEX

Merci de votre attention.