TextMine, 21 octobre 2022

Réflexions en cours pour dynamiser la recherche en TAL et TDM autour des données ISTEX

Mathieu Constant
Université de Lorraine, CNRS, ATILF

ISTEX (www.istex.fr)

- Réservoir d’archives scientifiques normalisées au service de la recherche française pour un usage documentaire et TDM
  - 27 millions de documents
  - toutes les disciplines scientifiques
  - s’étalant sur 700 ans

- Une plateforme open-source, fondée sur une API, pour
  - Rechercher des documents
  - Définir et télécharger vos propres sous-corpus
  - Analyser votre corpus à partir des métadonnées et enrichissements
  - Accéder à quelques corpus prêts à l’emploi

ISTEX (2)

- Une volonté nationale
  - Née d’un programme d’investissement d’avenir (2012 – 2018, 60 millions d’euros)
  - Inscrite sur la feuille de route des infrastructures de recherche du MESRI depuis 2022

- Organisation
  - Comité de pilotage : CNRS, Consortium Couperin, Agence bibliographique de l’enseignement supérieur, Université de Lorraine, France Universités
  - Hébergement et développements : INIST (CNRS)

ISTEX (3)

- Acquisition des ressources
  - une politique massive d’achat centralisé d’archives scientifiques et de collections rétropectives
  - via le PIA ISTEX (2012 – 2018) et le GIS Collex-Persée (depuis 2019)

- Qui peut accéder à la base ?
  - Accès limité à l’ensemble des personnels de l’enseignement supérieur et de la recherche français

La base ISTEX

- Des collections rétrospectives multilingues et multidisciplinaires de la littérature scientifique mondiale
- Jeu de métadonnées modélisées dans les normes du web sémantique
- Ressources librement accessibles, sauf les textes intégraux réservés aux membres de l’enseignement supérieur et de la recherche français

La base ISTEX – son contenu

- 28 éditeurs (70 % des documents viennent d'Elsewier, Wiley et Springer)
- Multidisciplinaire : sciences physiques (9,2M), sciences de la santé (7,4M), sciences de la vie (6,4M), sciences sociales (4,1M), général (0,4M)
- Multilingue : 50 langues, mais 90 % des documents sont en anglais
- 95 % sont des publications issues de périodiques

Projet d’infrastructure de recherche
Feuille de route nationale 2022

- **Axe 1 : Développer le corpus**
  - plus d’open access
  - partenariats européens

- **Axe 2 : consolider et développer l’offre de service**
  - Enrichissement des métadonnées
  - Développement d’outils avancés avec les communautés TAL et TDM
  - Offre de service TDM pour tous

Groupe de travail pour dynamiser la recherche autour des données ISTEX

- Deux pistes
  - Une compétition autour des données ISTEX avec des tâches intéressantes
  - Un projet scientifique en TAL et fouille de textes

- Condition: constituer un jeu de données ouvert (textes intégraux + métadonnées)
  - A minima, un sous-corpus de taille suffisante pour mener des expériences de TAL et/ou fouille de textes (apprentissage de modèles et leur évaluation)
  - Accessible à tous (communauté internationale, entreprises, ...)
  - Première piste : sous-corpus des articles open-access (en cours d'exploration)

Groupe de travail pour dynamiser la recherche autour des données ISTEX

- Quelques tâches envisagées pour la compétition
  - Prédiction de métadonnées : résumé, mots-clés, année de publication, …
  - Analyse des textes intégraux : entités nommées, relations sémantiques, structure discursive, ...

- Quelques thématiques générales de recherche
  - Simplification
  - Multilinguisme
  - Temporalité

Conclusion

- ISTEX : un réservoir unique d’archives scientifiques normalisé
- Textes intégraux ouverts uniquement aux personnels de l’ESR français
- Objectif : développer un terrain de jeu ouvert pour les chercheurs en TAL et TDM
- Suggestions, commentaires, questions ?