ISTEX, un projet national d'archives documentaires : au-delà de l'accès au texte intégral, l'enrichissement des données par méthodes de fouille de textes.

Pascal Cuxac Alain Collignon

pascal.cuxac@inist.fr alain.collignon@inist.fr

INIST - CNRS
2, allée du parc de Brabois
CS 10310
54519 Vandœuvre lès Nancy Cedex

Le projet ISTEX (initiative d'excellence en Information Scientifique et Technique) a pour objectif de permettre à la communauté scientifique française d'accéder, à une bibliothèque numérique pluridisciplinaire en texte intégral regroupant l'essentiel des publications scientifiques mondiales. Ces dernières sont accessibles à tous les chercheurs, notamment ceux gravitants autour des thématiques de la fouille de texte, du TAL, de la recherche d'Information, etc. Cela se concrétise par des actions R&D à la fois pour enrichir les données brutes et aussi pour développer de nouveaux algorithmes de fouille et d'analyse de textes.

A travers quatre axes d'enrichissement (structuration des documents ; indexation automatique ; reconnaissance d'entités nommées ; catégorisation des documents) nous avons répondu aux trois principaux challenges rencontrés :

- Mise au point et intégration d'outils : entraînement, adaptation, mise en production,
- Passage à l'échelle : 20 millions de documents à traiter,
- Reversement des données.

Le résultat d'une ou toute partie de ces travaux a permis de proposer un nouveau processus de diffusion d'ISTEX en construisant des triplets de données alignées et interopérables selon les standards du web sémantique (LOD).

Nous construisons maintenant une plateforme dédiée à la fouille de textes directement connectée aux données ISTEX. Les outils mis à disposition, peuvent être développés en collaboration avec tout laboratoire désireux de faire partager une application.