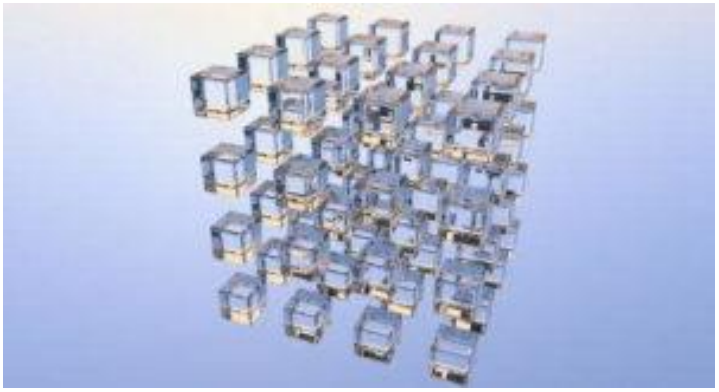
Le data mining

Eliane Caillou

LE DATA MINING

GALERIE 06/12/2019 ÉLIANE CAILLOU



Cette semaine, je me suis intéressée à la fouille de textes à partir d'un corpus donné.

Ce service proposé par l'INIST-CNRS m'a paru pratique et utile pour les chercheurs.

Ca semble compliqué mais il faut juste essayer et des formations existent dans le but d'aider à la prise en main des outils en question.

L'INIST, Institut National de l'Information Scientifique et Technique propose des formations au **Data Mining** ou **DTM** à partir d'un corpus donné, extrait d'ISTEX.
Istex, quezaco encore ? Un nouvel acronyme pour bibliothécaire…

ISTEX est une plateforme sur laquelle le chercheur et l'étudiant peuvent retrouver un ensemble de ressources documentaires acquises pour l'enseignement supérieur sous forme de Licences nationales.

Ces ressources rétrospectives d'*e-books* et de ***revues éléctroniques*** ont été négociées auprès des éditeurs par le MESRI pour toute la communauté universitaire et sont accessibles via le portail ISTEX. Petit à petit, en fonction des négociations avec les éditeurs, le portail s'étoffe de nouvelles ressources acquises dans le cadre de ces **Licences nationales**.

ISTEX prend plus largement place dans le vaste projet BSN (bibliothèque scientifique numérique) en 10 points avec 10 équipes projet, qui avaient pour but de rénover l'accès à l'information scientifique et technique à travers divers éléments à réorganiser et moderniser en fonction des pratiques actuelles. Ex : Les pratiques anciennes telles le PEB (prêt entre bibliothèque) devenaient obsolètes et ne répondaient plus aux besoins de la communauté universitaire.

BSN a finalement évolué vers le CoSO (comité pour la science ouverte). Ce dernier est composé d'experts chargés d'accompagner les initiatives pour la science ouverte. L'objectif est de mettre à disposition de tous et rapidement les résultats des données de la Recherche.

Istex répond à cet objectif. L'Agence bibliographique de l'enseignement supérieur, le CNRS, l'Université de Lorraine et Couperin ont participé à l'élaboration d'Istex. Il s'agit d'une initiative d'excellence faisant partie des investissements d'avenir du MESRI.

Cette plateforme donne accès à du texte intégral et permet, grâce à des services qui se développent aujourd'hui, de faire de la fouille de texte ou data mining

---

ISTEX mode d'emploi:

Tout d'abord, après avoir défini son champs d'étude, aller sur Istex-dl , plateforme qui a permis de faciliter l'extraction de corpus d'Istex
Ex de recherche sur ISTEX: " *Space and Planetary Science*"

- choisir ensuite la langue voulue
- le format d'extraction voulu
- puis télécharger.

Une fois le corpus téléchargé, utiliser **Lodex** à télécharger sur Gitub. C'est un outil open source.

Une fois téléchargé et installé, voici des tutos pour l'utiliser.

Vous trouverez aussi le wiki de Lodex en anglais pour de plus amples renseignements ainsi que la page de Lodex pour l'utilisateur, réalisée par l'INIST-CNRS
A partir de là, vous pouvez par exemple transformer un jeu de données en site web.

L'outil peut générer des graphiques, des statistiques, des cartes avec vos données. Il est adapté au web sémantique et génère automatiquement des URI, adresses web pérennes par opposition aux URL basées sur la localisation de vos données.

Via le format ARK, il génère aussi des identifiants perennes
Sur le site de l'INIST-CNRS, vous avez toutes les infos ainsi qu'un ensemble de tutoriels.

AUTRES OUTILS : GARGANTEX et CILLEX
Ces outils sont conçus pour vous fournir :

- **GARGANTEX** Des cartes évolutives et dynamiques au fur et à mesure que vous travaillez :
- **CILLEX** Des cartes en fonction des métadonnées récupérées lors d'une requête dans l'API ISTEX. Ces cartes aident à la sélection de métadonnées pertinentes liées à un choix de recherche.

Bon, voilà . On a fait le tour. C'est rapide mais l'essentiel y est. A tester maintenant pour ceux qui veulent. Voir cet exemple. Sympa non les graphiques !