Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Vers un corpus optimal pour la fouille de textes : stratégie de constitution de corpus spécialisés à partir d'ISTEX

Camille de Salabert[1] Sabine Barreaux[1]
(1) Inist-CNRS, 2, rue Jean Zay, CS 10310, 54519 Vandœuvre-lès-Nancy, France
camille.de-salabert@inist.fr, sabine.barreaux@inist.fr

RÉSUMÉ

Préalable indispensable à de nombreuses activités de TAL et de fouille de textes, l'élaboration d'un corpus peut nécessiter plusieurs phases de traitement pour améliorer sa qualité et ainsi obtenir les meilleurs résultats d'analyse automatique. Les post-traitements appliqués à un tel corpus, notamment pour garantir la pertinence de son contenu et l'homogénéité de son format, pourront s'avérer d'autant plus coûteux et fastidieux que la construction du corpus de travail aura été imprécise. Cette démonstration se proposera de tirer parti de la plateforme ISTEX et de ses services associés pour constituer, au travers d'un cycle itératif, un corpus homogène de publications scientifiquement pertinentes pour une utilisation simplifiée par des outils de fouille.

ABSTRACT

Towards an optimal corpus for text mining: specialized corpus building strategy from ISTEX.
The development of a corpus is an essential prerequisite for many NLP and text mining activities. This may require several processing phases to improve its quality and thus obtain better automatic analysis results. Post-processing applied to such a corpus in particular to guarantee the relevance of its content and the homogeneity of its format may be more costly and tedious if the construction of the working corpus is more basic. This demonstration will take advantage of the ISTEX platform and its associated services to build a homogeneous corpus of scientifically relevant publications for a simplified use by text mining tools through an iterative cycle.

MOTS-CLÉS : Fouille de textes, Corpus thématique, ISTEX, Affinement, Méthodologie.
KEYWORDS: Text mining, Thematic corpora, ISTEX, Refinement, Methodology.

Les algorithmes de fouille de textes promettent une accélération sans précédent de la recherche scientifique par le traitement automatique de corpus massifs de publications scientifiques qu'ils autorisent. Mais l'analyse textuelle peut se révéler compliquée quand il s'agit de traiter des corpus de grande dimension, multithématiques et multilingues, dont les textes, bien souvent non structurés et aux formats hétérogènes, sont rédigés dans des langues de spécialité. Dès lors, disposer d'un corpus tout à la fois adapté à l'objectif visé et approprié par rapport au cadre méthodologique et à l'outil considéré apparaît indispensable.

Si les guides de bonne pratique à l'intention de la communauté des linguistes s'intéressent à la construction de corpus équilibrés, représentatifs, échantillonnés ou adaptés à l'application recherchée (Wynne, 2005), peu de détails sont donnés en revanche quant à la phase de constitution

Volume 4 : Démonstrations et résumés d'articles internationaux, pages 66–69.
HAL : hal-02768520.

de corpus dans les travaux plus généraux de traitement automatique des langues (TAL) ou de fouille de textes et de données (TDM). La démarche mise en œuvre pour créer un corpus dépend fortement du but poursuivi et des sources utilisées, et se résume dans bien des cas à recourir à une équation basique, suivie d'étapes de nettoyage ou de reformatage, *a fortiori* dans le cas de textes intégraux (Nguyen, 2019). Dans cette démonstration, nous nous attacherons à montrer, au travers d'un cas d'usage pouvant intéresser la communauté scientifique, comment l'utilisation de la bibliothèque scientifique numérique nationale ISTEX[1] et de ses services additionnels permet de créer, en adéquation avec la finalité souhaitée, un corpus raffiné au travers de cycles successifs d'exploration des résultats et d'ajustement de la requête exploitant les points forts d'ISTEX. Cette démonstration permettra *in fine* de parvenir à un corpus qui minimise les étapes de post-traitement habituellement opérées, telles que l'élimination du bruit et du silence, le reformatage des données multisources pour les rendre homogènes ou le repérage de documents au format inadéquat pour l'application ciblée.

1 Cas d'usage

À l'occasion du 250e anniversaire de la naissance du musicien allemand Beethoven, la création d'un corpus en musicologie de publications relatives au compositeur nous offre l'opportunité d'ajouter un corpus thématique en sciences humaines et sociales aux corpus spécialisés déjà diffusés sur le site web d'exposition des données ISTEX : data.istex[2]. L'objectif final envisagé est de valoriser ce corpus sur le web sémantique au moyen d'alignements avec les données DOREMUS. DOREMUS[3] – DOing REusable MUSical data – est en effet un graphe de connaissances d'œuvres musicales interconnectées décrivant et contextualisant les catalogues musicaux de la BnF, de la Philharmonie de Paris et de Radio France qui sont ainsi partagés sur le web de données. Une étape préliminaire de reconnaissance d'entités nommées, entités propres à cet art telles que les noms de musiciens, les œuvres musicales ou les instruments de musique, ou génériques comme les dates et les lieux, sera nécessaire pour notre corpus. Elle pourrait être matière à comparer des outils de détection en fournissant un corpus d'apprentissage.

2 ISTEX

Source de données pour la fouille de textes

ISTEX, archive riche de plus 23 millions de publications numériques dans toutes les disciplines et en plus de 50 langues, depuis le XIVe siècle jusqu'à aujourd'hui, constitue une source de choix pour y rechercher les documents qui nous intéressent et leur appliquer des outils de fouille de textes, ce droit étant concédé par la licence d'utilisation de la plateforme ISTEX.

Financée par l'ANR et négociée dans le cadre des licences nationales, l'initiative d'excellence en information scientifique et technique ISTEX a de fait pour ambition de bâtir le socle de la bibliothèque scientifique numérique nationale et d'offrir, au travers de sa plateforme, l'accès aux collections rétrospectives de la littérature scientifique, ainsi que l'opportunité de s'en servir comme matériau de fouille à des fins de recherche scientifique.

1 https://www.istex.fr
2 https://www.data.istex.fr
3 http://data.doremus.org

Atouts de la ressource

Disponibles dans différents formats, fournis par les éditeurs ou produits par ISTEX, métadonnées et texte intégral des publications sont également proposés en format standard (respectivement MODS et TEI), facilitant considérablement l'exploitation de la diversité des formats reçus des 28 éditeurs présents dans l'archive.

A cette homogénéisation, sont ajoutés des enrichissements, disponibles dans un format TEI standoff, résultats du traitement de l'ensemble des collections par des outils de fouille de textes développés ou adaptés pour ISTEX (Cuxac, 2017). Ces enrichissements, de typologie plurielle, donnent notamment accès au contenu scientifique des ressources :

- catégories scientifiques attribuées respectivement à 75% et 44% des documents, soit par appariement entre un identifiant tel qu'un ISSN et une ou plusieurs catégories affectées à la publication concernée par le Web of Science, Science-Metrix ou Scopus, soit par apprentissage automatique sur les bases de données bibliographiques PASCAL et FRANCIS du CNRS ;
- entités nommées détectées dans 68% des textes en anglais ou français, grâce à une cascade de graphes mise au point pour le logiciel Unitex/CasSys (Maurel, 2019), ces entités nommées concernant les noms de personnes, lieux, organismes, indicateurs temporels, URL, etc. ;
- termes représentatifs du contenu du texte intégral, avec leur fréquence et leur spécificité, extraits de 73% des publications en anglais par TEEFT, outil d'indexation non supervisée qui dispense de la constitution problématique de ressources de référence spécialisées, nécessitées par une archive multidisciplinaire telle qu'ISTEX ;
- références bibliographiques structurées dans 58% des ressources au moyen de GROBID, outil de segmentation et de structuration fonctionnant par apprentissage automatique.

Forte de ces millions de publications numériques à valeur ajoutée, l'archive ISTEX forme une ressource majeure pour les travaux de fouilles de textes à l'usage de la communauté de l'enseignement supérieur et de la recherche.

Des services complémentaires

Toute une gamme de services a été conçue autour de l'API ISTEX afin de simplifier et de développer son utilisation, en particulier deux applications, l'une de téléchargement de corpus volumineux – ISTEX-DL –, l'autre d'exploration des corpus extraits, ainsi que d'exposition sur le web sémantique des corpus finalisés – LODEX :

- l'application ISTEX-DL[4] – ou ISTEX Download –, interface de type formulaire intuitive et conviviale, permet facilement de télécharger jusqu'à 100 000 documents répondant à une requête, classique ou issue d'une liste d'identifiants. Le choix des formats des textes intégraux, tout comme celui des métadonnées ou des enrichissements correspondants s'opère en quelques clics, de même que la limitation du volume souhaité (comprenant l'option de sélection aléatoire d'un sous-ensemble) et la modulation du type et du niveau de compression du corpus ;
- LODEX[5] – ou Linked Open Data ISTEX –, logiciel open source associant sémantisation et visualisation de données, est dédié quant à lui à la valorisation de données structurées (Gregorio, 2019). Cet outil transforme un corpus en site web dynamique, offrant d'une part une navigation dans ce corpus selon différents angles de vue au travers de graphiques

4 https://dl.istex.fr
5 https://lodex.inist.fr

filtrables par des facettes, et d'autre part une description et un accès aux documents du corpus. En outre, l'exposition sur le web de données autorise l'alignement avec des données similaires ou connexes, et permet le référencement et la réutilisation des corpus spécialisés.

3 Stratégie de constitution de corpus

Notre démonstration se proposera de débuter par une requête simple relative au musicien Beethoven qui ira se complexifiant, grâce à la visualisation des résultats de l'interrogation de l'API ISTEX via son démonstrateur[6], suivie de l'extraction à l'aide d'ISTEX-DL puis de l'analyse dans LODEX des états d'avancement successifs du corpus (états intermédiaire et affiné du corpus disponibles aux adresses suivantes : https://beethoven-ludwigv0.corpus.istex.fr/ et https://beethoven-ludwigv1.corpus.istex.fr). L'exploration à ces différentes étapes du contenu scientifique du corpus et de ses métadonnées nous permettra de proche en proche d'en déduire les correctifs à apporter à notre équation d'interrogation et nous donnera des clés pour enrichir la requête, tirant profit des qualités de la ressource ISTEX qui assure en amont les traitements de reformatage ou de complétion des données, de différents enrichissements tels que Unitex et TEEFT afin de limiter silence comme bruit – corollaire des personnages célèbres ! – et de permettre une sélection fine de documents pertinents, et des indicateurs de qualité disponibles sur la plateforme ISTEX au service du TDM pour ajuster aussi finement que possible le corpus à l'usage ciblé.

Ainsi, la méthodologie itérative de requêtage, d'extraction et d'exploration du corpus, permet de constituer un corpus de musicologie quasiment prêt à l'emploi pour l'extraction d'entités nommées, en réduisant le nombre des post-traitements de curation habituellement requis pour une exploitation valide.

Références

CUXAC P. & THOUVENIN N. (2017). Archives numériques et fouille de textes : le projet ISTEX. *In Atelier TextMine, EGC 2017 (Extraction et Gestion des Connaissances), Grenoble, France, January 24-27, 2017.*

GREGORIO S., COLLIGNON A., PARMENTIER F. & THOUVENIN N. (2019). LODEX : des données structurées au web sémantique. *In Atelier Web des Données, EGC 2019 (Extraction et Gestion des Connaissances), Metz, France, January 21-25, 2019.*

MAUREL D., MORALE E., THOUVENIN N., RINGOT P. & TURRI A. (2019). Istex: A Database of Twenty Million Scientific Papers with a Mining Tool Which Uses Named Entities. *Information*. DOI : 10.3390/info10050178, HAL: hal-02152978, version 1.

NGUYEN N., GABUD R. & ANANIADOU S. (2019). COPIOUS: A gold standard corpus of named entities towards extracting species occurrence from biodiversity literature. *Biodiversity Data Journal,* 7, e29626. DOI : 10.3897/BDJ.7.e29626

WYNNE M., Éd. (2005). *Developing Linguistic Corpora: a Guide to Good Practice*. Oxbow Books

[6] https://demo.istex.fr