ISTEX : Des enrichissements au web de données

Alain Collignon, Pascal Cuxac

ISTEX : DES ENRICHISSEMENTS AU WEB DE DONNÉES

Alain Collignon, Pascal Cuxac

A.D.B.S. | « I2D – Information, données & documents »

2017/4 Volume 54 | pages 8 à 15

ISSN 2428-2111

MÉTHODES

ISTEX : des enrichissements au web de données

[étude] 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. Nous développerons ici les actions R&D engagées pour enrichir les données brutes ainsi qu'un nouveau processus de diffusion d'ISTEX selon les standards du web sémantique (LOD).

L'accès à des ressources documentaires riches est non seulement essentiel pour une production scientifique de rang mondial, mais il est démontré qu'il y a une corrélation forte entre la disponibilité de ces ressources et la qualité de la recherche[1].

À l'ère du Web, nous assistons au développement des données en libre accès (*OpenData*), des collections issues de bibliothèques traditionnelles qui sont maintenant accessibles librement : Gallica, Europeana, Digital Public Library of America. À ce type de bibliothèques numériques s'ajoutent les publications savantes, qui occupent une part importante des publications numériques. De récentes initiatives nationales ont également permis le développement d'importantes archives scientifiques (ISTEX en France, SwissBib en Suisse, GBV en Allemagne, Scholars Portal en Ontario).

Le web sémantique est présenté comme étant le web pour lequel les ordinateurs interprètent les métadonnées afin de mieux assister l'utilisateur dans sa recherche de l'information (Berners Lee

1. http://www.rin.ac.uk/our-work/communicating-and-disseminating-research/e-journals-their-use-value-and-impact.

et *al.*, 2001). Par conséquent, il y a délégation d'interprétation par les machines du sens que l'on donne aux ressources (Bachimont *et al.*, 2011) et cela grâce aux évolutions technologiques du web qui ont permis de passer d'un web de documents (navigation hypertextuelle) au web de données (liens entre les données elle-même, espace unifié). Ce passage a été permis grâce à l'avènement de différents standards issus du web sémantique comme RDF, OWL, SPARQL et URI[2]. Bien entendu, les bibliothèques n'ont pas été insensibles à cette évolution pour faire migrer leurs catalogues (Prongué, 2014) et nous pouvons citer data.bnf.fr[3] pour exemple. Au-delà de ces établissements, des organismes publics[4] mettent en accès libre de jeux de données, des organismes de recherche apportent leurs réflexions concernant la publication de données de recherche (Aventurier, 2013).

Le projet ISTEX[5] (initiative d'excellence en Information Scientifique et Technique) a pour objectif de permettre à la communauté ESR française (Enseignement Supérieur et Recherche) d'accéder, via un accès en ligne, à une bibliothèque numérique regroupant l'essentiel des publications scientifiques mondiales dans toutes les disciplines scientifiques, en texte intégral.

Ce réservoir de publications scientifiques est bien entendu à destination des documentalistes et chercheurs ayant un besoin documentaire mais est également une ressource unique pour tous les chercheurs gravitants autour des thématiques de la fouille de texte, du TAL (Traitement Automatique de la Langue), de la Recherche d'Information… etc. La mise en ligne de ces informations en texte intégral structuré permet de développer des fonctionnalités modernes d'extraction de connaissances basées sur les technologies de la fouille de textes.

Dans la cadre des travaux de recherche nous aborderons dans cet article uniquement les traitements visant à enrichir les données ISTEX et accessibles via l'API ISTEX. Les technologies déployées dans l'API proprement dite ne seront pas abordées ici. Dans la suite de cet article nous allons passer brièvement en revue toutes les méthodes utilisées pour enrichir les données ; elles doivent répondre à un certain nombre d'exigences dont le passage à l'échelle sur plus de 19 millions de documents (optimisation des temps de traitements, gestion de la mémoire...), l'intégration dans une même chaîne de traitement (compatibilité des programmes) et des données en sortie au format TEI.

Enfin, nous présenterons LODEX qui a pour but de publier la documentation du fonds ISTEX selon les normes du web sémantique (Linked Open Data – LOD) et ainsi faciliter davantage l'accès et la diffusion des données acquises et produites dans/et par le projet ISTEX.

Les objectifs d'ISTEX-RD :

La plate-forme ISTEX fournit l'ensemble de ses services sous la forme d'une API Web[6] mais également via un démonstrateur[7] qui permet de se familiariser avec les formats et la syntaxe d'interrogation. Le projet ISTEX a une vocation double (Figure 1):

- **répondre à des besoins documentaires** (recherche bibliographique, état de l'art…)
- **être un réservoir de ressources textuelles agrégées exploitable** pour des travaux de recherche en fouille de textes ou en bibliométrie par exemple.

L'axe recherche/développement autour de la plateforme ISTEX s'est concrétisé par un appel à projet « Chantiers d'usages »[8]afin de « *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 de Text and Data Mining (TDM) de qualité.* » (Pierrel, 2016). Mais il est également intégré à la plateforme à travers le projet d'enrichissements des données mené par une équipe de l'INIST-CNRS en collaboration avec le LI[9]

Alain COLLIGNON
Ingénieur de recherche au CNRS au sein du service Recherche Développement et Expérimentations de l'INIST-CNRS, s'intéresse aux aspects liés au web de données et interopérabilité. Enseignant vacataire université de Lorraine.
Alain.COLLIGNON@inist.fr

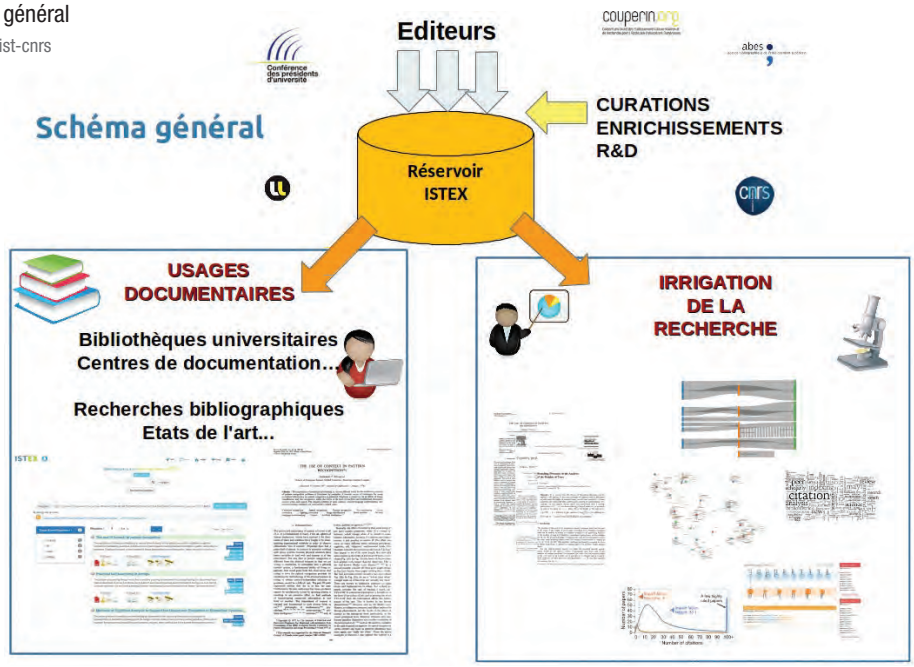
Pascal CUXAC, Docteur en Génie Géologique et Minier, CNRS comme Ingénieur de Recherche. Spécialiste des méthodes de fouilles de textes, service Recherche Développement et Expérimentations de l'INIST-CNRS (Institut de l'Information Scientifique et Technique), en charge de la R&D dans ISTEX.
Pascal.CUXAC@inist.fr

INIST-CNRS
2 allée du parc de Brabois,
CS 10310,
54519 VANDŒUVRE
LÈS NANCY CEDEX
France

2. http://fr.slideshare.net/AntidotNet/web-smantique-web-de-donnes-web-30-linked-data-quelques-repres-pour-sy-retrouver
3. http://data.bnf.fr/
4. https://opendata.paris.fr/page/home/ ; https://www.data.gouv.fr/fr/
5. http://www.istex.fr/
6. https://api.istex.fr/documentation/
7. http://demo.istex.fr/
8. http://www.istex.fr/seminaire-technique-25-et-26-avril-2016/
9. http://tln.li.univ-tours.fr/

////

Figure 1
Schéma général
Source : inist-cnrs



de Tours, le LS2N[10] de Nantes et ScienceMiner[11]. Ces travaux gravitent autour de 6 grands axes qui vont être développés par la suite :

- La catégorisation des documents,
- L'extraction et la structuration des références bibliographiques sortantes,
- La détection des entités nommées,
- L'indexation des documents,
- La structuration XML/TEI du texte plein à partir du pdf,
- L'approche « données ouvertes liées »,

Tous ses traitements aboutissent à des enrichissements directement requêtables via l'API.

Les enrichissements

Une chaîne de traitement a été mise en place afin d'intégrer les modules d'enrichissements indépendants et faciliter l'exécution des traitements par des professionnels de l'IST. Chaque module va chercher dans le réservoir ISTEX les données nécessaires à son fonctionnement, crée un fichier d'enrichissement résultat et complète un fichier json (le doc-object) qui est la carte d'identité du document comportant le signalement de tous les traitements effectués. Une interface de gestion appelée Concerto permet facilement de lancer un ou plusieurs modules d'enrichissement sur un corpus donné, les résultats sont alors accessibles en temps réel sur l'API ISTEX au fur et à mesure de l'exécution des programmes.

Le défi relevé est de faire cohabiter dans une même chaîne des modules très différents (classification, indexation, structuration…) pouvant être exécutés individuellement ou simultanément sans alourdir de façon significative les temps de traitements. À titre d'information, 1 million de documents peuvent être traités en 9 heures environ (16 CPU, 32 Go RAM).

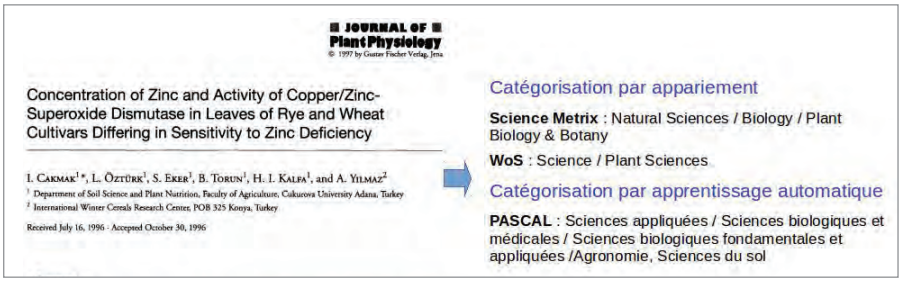
La catégorisation des documents

ISTEX contient actuellement 19,05 millions de documents couvrant tous les domaines de recherche. Un marquage de tous ces textes par une ou plusieurs catégories scientifiques est rapidement apparu nécessaire à la fois pour les décideurs, afin d'avoir des statistiques sur le fonds, et également pour les utilisateurs afin de cibler un domaine particulier lors de l'interrogation. À cette fin deux approches complémentaires ont été implémentées :

• Une catégorisation par appariement
Le principe en est simple puisqu'il s'agit de mettre en correspondance un identifiant de publication (ISSN par exemple) avec une ou plusieurs catégories attribuées à cette publication par un organisme reconnu.

10. https://ls2n.fr/equipe/taln/
11. http://science-miner.com/
12. http://ip-science.

Figure 2
Un article catégorisé par appariement et par apprentissage automatique
Source : inist-CNRS



Trois ressources ont été choisies : celle du Web of Science[12], de Science-Metrix[13] et de Scopus[14]. Mais les domaines scientifiques attribués à une revue ne sont pas toujours adaptés à catégoriser tous les articles de la même revue. C'est pour cela, et aussi parce que tous les documents ISTEX ne sont pas catégorisés par ces sources, que nous avons complété ces résultats par ceux obtenus en utilisant une méthode de classification supervisée.

• Une catégorisation par apprentissage automatique

Nous avons développé un module basé sur un Bayésien Naïf[15] avec un apprentissage sur les bases PASCAL/FRANCIS[16] du CNRS. Le module s'appuie sur un apprentissage en cascade : il commence par déterminer si le document est SHS (sciences humaines et sociales = FRANCIS) ou STM (sciences techniques et médecine = PASCAL), puis par exemple dans le cas de STM, l'étape suivante sera de déterminer si ole document appartient aux sciences de la vie ou non et ainsi de suite

La figure 2 donne un exemple d'article catégorisé à la fois par la méthode par appariement et par apprentissage automatique. La catégorisation automatique par apprentissage apporte un plus quand la catégorie associée à la revue est trop générique (revues multidisciplinaires par exemple) ou quand celle-ci n'existe pas.

Les références bibliographiques

Dans le réservoir ISTEX un nombre important de documents pleins textes n'est accessible qu'en format non structuré. Dès le début du projet il est apparu important de pouvoir détecter et structurer les références bibliographiques afin de les rendre interrogeables et pouvoir ainsi naviguer dans le réservoir ISTEX ou faire le lien avec d'autres ressources extérieures. Nous avons pour cela utilisé l'outil Grobid[17] développé par Science-Miner (Lopez, 2009). Partant du pdf, l'outil va utiliser des CRF (Conditional Rando Field)[18] en cascade pour découper le document et baliser les références bibliographiques. Pour chaque référence détectée nous avons un découpage en auteur, titre, année, / / / /

thomsonreuters.com/mjl/scope/scope_scie/

13. http://www.science-metrix.com/fr/classification

14. https://www.scopus.com/sources?zone=&origin=NO%20ORIGIN%20DEFINED
https://fr.wikipedia.org/wiki/Classification_na%C3%AFve_bay%C3%A9sienne

15.http://pascal-francis.inist.fr/

16. http://grobid.readthedocs.io/en/latest/

17. https://fr.wikipedia.org/wiki/Champ_al%C3%A9atoire_conditionnel

18. http://tln.li.univ-tours.fr/

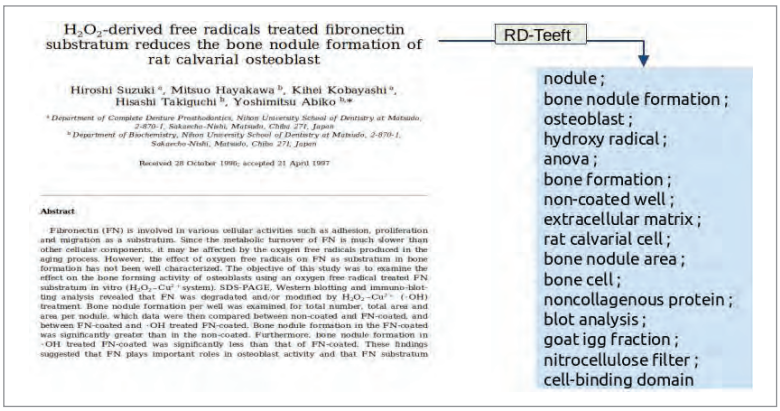


Figure 3
Exemple d'indexation Teeft.
Source : inist-CNRS

/ / / /

//// publication… etc. Cela permet par exemple de construire un graphe de citations ou de calculer une proximité entre documents.

Les entités nommées

Dans le cadre d'un partenariat avec le Laboratoire d'Informatique de l'Université François Rabelais de Tours[19] la plateforme Unitex[20] a été adaptée et complétée par un système de cascades de graphes CasSys afin de traiter de gros volumes de textes en français et en anglais. Les dix types suivants d'entités nommées ont été choisis pour être détectées et extraites : les noms de personnes, les noms de lieux, les noms d'organisations, les dates, les organismes financeurs, les projets financés, les URL, les citations, les organismes hébergeurs de ressources.

19. http://www-igm.univ-mlv.fr/~unitex/
20. http://tln.li.univ-tours.fr/Tln_Telechargement/Guide_Annotation_EN_lstex.pdf
21. http://www.tei-c.org
22. https://github.com/istex/odd-istex
23. https://raw.githubusercontent.com/Inist-CNRS/lodex/master/LICENSE

À noter que le guide d'annotation qui a permis l'évaluation de ces graphes est disponible sur le site du laboratoire de Traitement des Langues Naturelles de Tours[21].

L'indexation

L'indexation automatique de documents sélectionne des termes extraits d'un document pour donner une représentation de ce texte. Cette indexation peut être utilisée pour aider à la recherche de documents pertinents mais également dans des tâches de classification qui pourraient être appliquées à des sous corpus. Du fait de la diversité des documents, une indexation supervisée impliquant des ressources spécialisées couvrant tous les domaines n'est pas envisagée.

RD-TEEFT est un outil développé en interne ; il traite les documents en texte plein en anglais pour produire une liste de termes extraits et leur spécificité. Teeft est basé sur les bibliothèques Topia (méthode POS) et NLTK. Sur l'exemple de la figure 3 nous pouvons constater que la méthode est capable d'extraire un nombre limité et représentatif de termes à partir d'un article de plusieurs pages.

Notre étude est de mettre en ligne les jeux de données pour répondre aux demandes des documentalistes et des chercheurs, en utilisant la structuration sémantique

La structuration XML/TEI des documents

Pour la plupart des documents existent différents formats : mods (métadonnées), pdf, texte, xml (TEI). Pour l'instant le Xml-TEI d'un document ne structure pas le corps du document qui se retrouve en format texte dans une balise <body>. Avoir des documents entièrement au format Xml-TEI aurait des intérêts multiples : faciliter les manipulations de textes pour l'utilisateur, permettre un « nettoyage » des documents avant divers traitements (par exemple, ôter les figures, tableaux, en-têtes… afin d'avoir une extraction d'entités nommées, ou une indexation plus performante), utiliser la structure même du document pour améliorer certains algorithmes. Nous avons abordé cette tâche en utilisant l'outil Grobid qui était déjà en production pour la structuration des références bibliographiques. Cela demande cependant un travail important pour pouvoir faire un bon apprentissage du CRF, nécessitant de baliser manuellement plusieurs centaines de documents de différents éditeurs avant de pouvoir tester notre approche sur un corpus réel.

Le reversement des données

À travers l'API ISTEX, les données sont déchargeables aux formats pdf, texte et xml-TEI. Tous les enrichissements produits et reversés doivent pouvoir être visibles par l'utilisateur final voire être interrogeables. Nous avons décidés d'enrichir le document xml-TEI : pour chaque document les enrichissements produits sont placés dans une balise <standOffs> après les métadonnées du document et conforme aux standards de la TEI (Text Encoding Initiative)[22]. La TEI est un format XML de description de textes. Elle permet de décrire la structuration du texte tel qu'il a été conçu et non son rendu final.

Afin de pouvoir traiter et valider ces enrichissements nous avons produit un schéma ODD-ISTEX[23] qui est mis à disposition de tout partenaire désireux de créer ses propres enrichissements au format TEI. Le choix du « *standOff* » TEI a été dicté par le souhait de ne pas modifier les textes publiés : les enrichissements viennent donc compléter les textes comme des métadonnées.

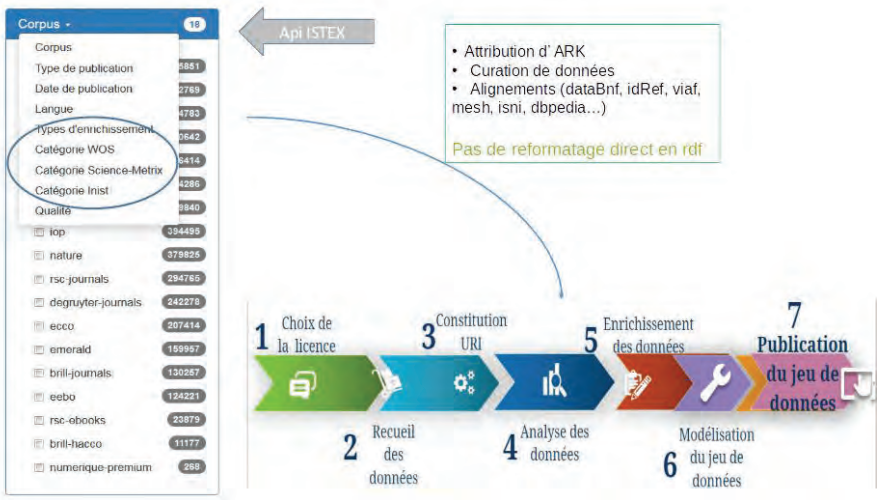
Exposition des enrichissements dans le web de données

Objectif

Le challenge principal de notre étude est de mettre en ligne les jeux de données précédemment produits dans le respect des normes et standards du W3C. Cela afin de répondre aux demandes des documentalistes et des chercheurs, en utilisant la structuration sémantique comme un moyen pour répondre à plusieurs besoins :

- **proposer une documentation structurée et interopérable** du fonds ISTEX pour les utilisateurs de portail documentaire comme pour les chercheurs,
- **mettre à disposition des équipes de recherche des jeux de données** très spécifiques permettant d'alimenter leurs travaux de recherche sur du *machine learning* ou du *data alignement*,
- **valoriser les jeux de données** produits par des travaux de recherche,
- **rendre compatible le fonds ISTEX avec des entrepôts de**

Figure 4
Vue analytique du traitement des données.
Source : inist-CNRS



données présents dans le web sémantique,
- **faciliter d'avantage les travaux de recherche** dédiés à la fouille de textes (bibliométrie, scientométrie, etc.).

À terme le but est de proposer une nouvelle vue des documents ISTEX au travers de jeux de données liés, alignés et interopérables.

Les jeux de données

Les jeux de données sont là pour venir compléter, enrichir, consolider et lier toutes les informations présentes dans la plateforme Istex. Nous souhaitons proposer un graphe de jeux de données structurées reliées à des ressources extérieures ou à des référentiels d'autorité. *In fine*, ce lacis de données conduira toujours à un retour vers les documents plein texte présents dans ISTEX. C'est une autre façon pour diffuser et exploiter les ressources acquises.

Des jeux de données sont constitués principalement à partir d'informations extraites automatiquement : entités nommés, catégories scientifiques Sciences-Metrix / WOS, catégories Pascal (*cf.* paragraphe 3). De plus, pour augmenter le graphe, nous leurs avons adjoint des données récupérées à partir des informations induites et produites par les documentalistes (types de publication, regroupement des langues, etc.).

La méthodologie

La démarche adoptée a permis à ce que l'ensemble de ces données suit le traitement élaboré dans le cadre d'un processus empirique permettant d'aboutir à la mise en ligne de données liées et ouvertes conformément au processus élaboré par Fabry *et al.* (2017). Ainsi, nous confrontons les notions théoriques du web sémantique appliquées en milieu documentaire (Bermès *et al.* 2013) avec notre réalité de terrain. En particulier nous nous sommes attardés sur le caractère hétérogène des ressources et son incidence sur le protocole à mettre en œuvre. L'originalité de ce travail est de publier des données autres que des données bibliographiques et ceci sans reformatage.

Comme il est illustré dans la figure 4, l'ensemble des différentes étapes a été réalisé dans l'outil LODEX.

L'outil LODEX

L'avènement des bibliothèques numériques a motivé les professionnels à faire évoluer leurs pratiques concernant le traitement de leurs données. Dorénavant, la curation, la modélisation, la normalisation, le modèle RDF sont au cœur des préoccupations des data-managers. Ceci a eu pour incidence, l'émergence d'outils dédiés à ces activités comme par exemple LODRefine et Catmandu (Harlow, 2015). Plus près de nos préoccupations, le logiciel (Content System Management – CubicWeb) dédié aux techniques du web sémantique est utilisé dans le développement de l'application data.bnf.fr (Le Bœuf, 2013).

Le logiciel CubicWeb, présente de nombreuses fonctionnalités ////

//// pouvant nous être utiles, cependant l'usage de ce *framework* nécessite l'appui technique de la société Logilab, par conséquent, nous nous sommes orientés dans le développement d'une solution logicielle en interne.

Il a été développé avec les technologies en javascript, et plus particulièrement en Ecmascript 6. L'outil LODEX est un logiciel libre dont le code source est accessible sur github (https://github.com/Inist-CNRS/lodex) et sous licence CECILL[24].

24.https://raw.githubusercontent.com/Inist-CNRS/lodex/master/LICENSE

Son back office permet de réaliser toutes les fonctionnalités nécessaires au traitement ou « stylage » d'un jeu de données : création de l'URI, curation, et la publication du jeu de données. Il présente la particularité de permettre entre autre de donner du sens aux ressources grâce à la fonctionnalité format (par exemple une URL est affichée comme telle donc cliquable ou en affectant une propriété ou une classe à chaque ressource).

Après curation, « sémantisation », le jeu de données est publié *via* le front office (https://data.istex.fr/). Différents exports compatible aux techniques du web sémantique sont possibles (Turtle pour sa lisibilité ; N-Quads et N-Triple pour leur simplicité et JSON pour son application).

Conclusions et perspectives

Cet article présente un cas d'application de méthodes de fouilles de textes et de TAL sur une bibliothèque numérique volumineuse afin d'enrichir les données et de faciliter l'interrogation et l'analyse.

Nous avons mis l'accent sur l'utilisation du réservoir ISTEX en tant qu'archive pour la recherche documentaire, mais aussi réservoir de documents pleins textes dans toutes les disciplines scientifiques pour des développements d'applications et d'outils de TDM. Nous avons illustré quelques développements qui demandent encore, pour certaines méthodes, à être consolidés. À travers quatre axes de travail (structuration des documents ; indexation automatique ; reconnaissance d'entités nommées ; catégorisation des documents) nous avons répondu aux trois principaux challenges rencontrés, c'est à dire :

- **Mise au point et intégration des outils :** entraînement,

Bibliographie

P. Aventurier.
Données ouvertes de la recherche : nouvelles pratiques de publication et de partage. Cours Enssib (Cours), 2013.
http://prodinra.inra.fr/record/217160

B. Bachimont, F. Gandon, G. Poupeau, B. Vatant, R. Troncy, S. Pouyllau.
Enjeux et technologies : des données au sens. *Documentaliste-Sciences de l'information*, 2011, vol. 48, n° 4, p. 24-41

E. Bermès, A. Isaac, G. Poupeau.
Le Web sémantique en bibliothèque. 2013, Electre-Ed. du Cercle de la Librairie, Paris. Collection : Bibliothèques – 171 pages.

T. Berners Lee, J. Hendler, O. Lassila.
The semantic web. *Scientific American*, 2011, p. 29-37.

C. Fabry, C. Roussel, A. Collignon, E. Moreau, F. Parmentier, N. Thouvenin.
Publier des données liées et ouvertes en sept étapes. *I2D – Information, données & documents*, 2017, n° 1 (Volume 54), p. 12-14.

C. Harlow. Data Munging tools in preparation for RDF: Catmandu and LODRefine.
Code {4}lib journal, 2015, vol. 30, p. 1-12.

A. Joulin, E. Grave, P. Bojanowski, T. Mikolov.
Bag of tricks for efficient text classification.
arXiv:1607.01759 [cs], 2016.
http://arxiv.org/abs/1607.01759.

P. Le Bœuf.
Customized OPACs on semantic Web : the OpenCat prototype. Actes IFLA Satellite Meeting 2013, Singapore, 2013.

P. Lopez.
GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications. *Research and Advanced Technology for Digital Libraries*, 2009, 13th European Conference, ECDL 2009, Corfu, Greece

J. Pierrel.
Séminaire chantiers d'usage. 2016, Communication orale.

N. Prongué.
Modélisation et transformation des métadonnées de RERO en Linked Open Data. Haute école de gestion de Genève, 2014.
http://doc.rero.ch/record/232839.

configuration, adaptation, mise en production,
- **Passage à l'échelle :** actuellement 19,05 millions de documents à traiter,
- **Reversement des données :** modélisation, ré-intégration, mise à disposition.

L'objectif principal de notre approche est de pouvoir générer de nouveaux enrichissements par différentes techniques à partir du fonds ISTEX et de développer une méthodologie didactique afin de les valoriser via le web de données ou Linked Open Data. Ce réseau sémantique a pour but de centrer les utilisateurs non plus sur le document mais la donnée elle-même.

L'originalité de nos travaux, au-delà de la mise en évidence de riches et nombreux enrichissements, est de les rendre visible donc les valoriser en respectant les normes du web sémantiques ce qui présente l'avantage d'ouvrir le fonds ISTEX au Web.

Nos travaux ultérieurs consistent à agréger de manière cohérente toutes les données publiées via l'outil LODEX en respectant une ontologie spécifique. Cette agrégation doit amener à un SPARQL endpoint contenant un graphe global des données ISTEX. L'objectif ultime étant d'insérer ISTEX dans le graphe global géant du Web pour faciliter les collaborations entre institutions mais aussi pour introduire la possibilité de faire raisonner les données ISTEX.

Remerciements

Ces travaux ont été financés par le projet ISTEX avec le soutien de l'Agence Nationale pour la Recherche dans le cadre du programme d'Investissements pour le Futur de référence ANR-10-IDEX-0004-12. ■