Le TDM pour tous grâce à des web services au sein de LODEX, outil libre de visualisation

Valérie Bonvallot*, François Parmentier*, Lucile Bourguignon*, Isabelle Clauss*, Stéphanie Gregorio*

*Inist-CNRS 2, rue Jean Zay CS 10310 54519 Vandoeuvre-lès-Nancy, France
valerie.bonvallot@inist.fr
francois.parmentier@inist.fr

Résumé. Pour faciliter l'accès aux techniques de fouille de données notamment pour les non spécialistes, le service TDM (Text and Data Mining) de l'Inist-CNRS développe des web services autour du traitement de l'information scientifique et technique. Ces services peuvent être appelés en ligne de commande ou au sein de LODEX, outil libre de visualisation. La démonstration montre comment, à partir des informations présentes dans une notice bibliographique et plus particulièrement à partir d'une adresse d'auteur, l'identifiant RNSR (Répertoire national des structures de recherche) est attribué automatiquement au document initial et comment cette nouvelle donnée est exploitée au sein de LODEX. Ainsi, programme ou algorithme développé par des enseignants chercheurs pourrait être adapté pour devenir un web service et être utilisé par le plus grand nombre.

1 Introduction

L'exploitation de la bibliométrie et de la fouille de données ne cesse de se développer avec l'accroissement de la quantité d'information sur internet et l'augmentation de la puissance de calcul des ordinateurs à portée de main.
Cet article se focalise sur les données bibliographiques et donc les données structurées, domaine privilégié de l'Institut de l'information scientifique et technique (Inist) du CNRS. Cela ne l'empêche pas de traiter du texte libre en analysant notamment les titres et les résumés.

Depuis la publication des deux plans nationaux de la science ouverte [1], documentalistes, bibliothécaires, professionnels de l'information doivent répondre aux demandes croissantes de tableaux de bord de la part de leur direction pour mettre en évidence, entre autres, des taux d'*open access* en fonction des disciplines, des instituts ou tout autre indicateur.

1. Le premier plan national : https://www.enseignementsup-recherche.gouv.fr/fr/le-plan-national-pour-la-science-ouverte-les-resultats-de-la-recherche-scientifique-ouverts-tous-49241
Le second plan national : https://www.enseignementsup-recherche.gouv.fr/fr/le-plan-national-pour-la-science-ouverte-2021-2024-vers-une-generalisation-de-la-science-ouverte-en-48525

Il existe certes des applications « presse bouton » mais leur utilisation dépend d'un abonnement (SciVal avec les données Scopus d'Elsevier [2], Incites [3] avec les données du WoS de Clarivate Analytics). En plus du coût, les données ne sont pas toujours homogénéisées ou le sont en partie grâce au travail de curateurs de certains établissements [4]. De plus, tous les domaines scientifiques ne sont pas toujours représentés, notamment les sciences humaines et socialesMaddi et De La Laurencie (2018).

Les professionnels de l'information n'ont pas toujours des facilités technologiques, voire informatiques, ou n'ont pas la chance de pouvoir travailler avec des informaticiens, pour ne pas dépendre de ces applications onéreuses et pour développer leurs propres outils.

On constate également la mise en ligne croissante, via GitHub ou GitBucket, de programmes permettant de traiter des données. Or leur mise en oeuvre souvent complexe n'incite pas les non informaticiens à les utiliser. Des plate-formes comme CortextBaneyx et al. (2009) Barbier (2021), GargantextChavalarias et al. (2021) Lobbé et al. (2021) et WekaFrank et al. (2010) sont aussi disponibles mais elles nécessitent souvent un niveau de connaissance des méthodes de TDM pour choisir parmi les algorithmes proposés et les paramétrer. [5]

Devant cette situation, en 2021, le service TDM [6] de l'Inist-CNRS a commencé à mettre en ligne des web services qui permettent à tout à chacun de faire du TDM. Les non spécialistes ont ainsi à portée de main des outils simples d'utilisation pour réaliser de l'extraction de connaissance nécessaire à leurs études. Dans cet article, nous présenterons dans un premier temps la philosophie de ces web services, le fonctionnement de l'un d'entre eux et dans un second temps, nous verrons comment ils peuvent être utilisés indépendamment de toute plate-forme ainsi qu'au sein d'une application libre de visualisation de données (LODEX).

2 Web services : spécificités et exemple

2.1 Spécificités

2.1.1 Aucune installation

L'Inist-CNRS propose une série de services web. Ces derniers donnent accès à des traitements spécialisés en IST (information scientifique et technique), sans avoir à installer de programme spécifique, ni son environnement particulier (comme par exemple un langage de programmation et son gestionnaire de bibliothèques). Cette facilité d'utilisation est liée au fait que les outils sont installés sur les serveurs de l'Inist-CNRS. Ainsi tout est transparent pour l'utilisateur.

2. Elsevier : https://www.elsevier.com/fr-fr/solutions/scival
3. Clarivate : https://clarivate.libguides.com/incites_ba
4. Les données WoS ont un champ « Organization_enhanced » (renommé « Affiliation » dans la nouvelle interface) qui utilise une forme d'écriture préférentielle https://support.clarivate.com/ScientificandAcademicResearch/s/article/Web-of-Science-Core-Collection-How-to-create-Organization-Enhanced-Preferred-name-and-update-variants?language=en_US
5. Cortext https://www.cortext.net ; Gargantex https://gargantext.org ; Weka https://waikato.github.io/weka-wiki/text_categorization_with_weka/
6. Service TDM https://www.inist.fr/services/analyser/text-data-mining/

2.1.2 Un seul traitement par web service sans paramétrage

Un web service effectue la plupart du temps un seul traitement ou des traitements proches. Nous avons porté notre choix sur des web services minimalistes pour différentes raisons :

— un repérage facile des fonctionnalités ;
— des besoins différents des utilisateurs ;
— une souplesse et modularité de leur utilisation : l'utilisateur peut lancer successivement plusieurs web services ;
— l'utilisation par le plus grand nombre.

Pour l'instant les services sont codés en python ou en NodeJS.

2.2 Quels web services ?

2.2.1 Présentation générale

Plusieurs services sont mis à disposition [7]. Ils peuvent être regroupés par domaine, en fonction des données à enrichir :

— les affiliations : traitements sur les affiliations des auteurs dans des notices bibliographiques (que ce soit la structuration des adresses via libpostal, bibliothèque de références pour le traitement des adresses postales, l'attribution d'identifiants RNSR [8] (Référentiel national des structures de recherche) à partir de l'affiliation, l'attribution des instituts CNRS à partir des identifiants RNSR repérés) ;
— la création d'identifiants pérennes : ARK [9] ;
— les données bibliographiques : enrichissement des informations bibliographiques à l'aide des services comme CrossRef [10] (attribuer l'éditeur à partir de la racine DOI) et Unpaywall [11] (récupérer les informations liées à l'*open access*) ;
— le texte : indexation via extraction de termes d'un texte à partir de terminologies [12], attribution d'un domaine scientifique grâce à des méthodes de classification supervisées.

Cette liste a naturellement vocation à s'étoffer en fonction des besoins repérés et/ou exprimés. Les services seront détaillés sur un site web en cours de réalisation.

Nous illustrons ici leur utilisation à travers l'exemple du web service dédié à l'attribution d'un identifiant RNSR à partir d'une affiliation, d'une adresse d'auteurs.

2.2.2 Attribution d'identifiant(s) RNSR

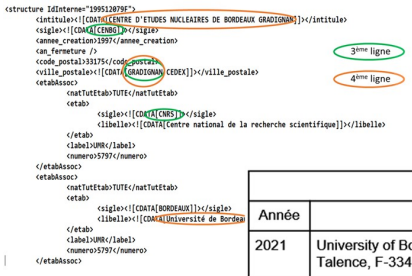
Malgré les consignes données aux chercheurs pour rédiger l'adresse, l'affiliation qui apparaîtra dans l'article, les formes d'écriture sont multiples. Pour repérer les publications d'une entité (équipe, laboratoire, institut ou fédération), l'homogénéisation des noms est un passage obligé. En France, il existe le RNSR (Répertoire national des structures de recherche), qui attribue à une structure un identifiant. Un des web services permet ainsi de retrouver, à partir d'une adresse d'affiliation d'un chercheur et d'une année de publication, l'identifiant RNSR de sa

7. Web Services : https://gitbucket.inist.fr/tdm/web-services
8. RNSR : https://appliweb.dgri.education.fr/rnsr/index.jsp?INIT=OK
9. ARK : https://arks.org/
10. CrossRef : https://www.crossref.org/
11. Unpaywall : https://unpaywall.org/
12. Loterre (Linked open terminology resources) plateforme d'exposition et de partage de terminologies scientifiques multidisciplinaires et multilingues (https://www.loterre.fr/)

structure de recherche. Il a été constitué à partir de règles « métier » [13] en repérant la présence de différentes informations. Un identifiant est attribué si dans l'adresse sont retrouvés : (code postal OU ville) ET tutelle ET ((label ET numéro) OU sigle OU libellé). Ces règles ont été éprouvées dans le cadre des campagnes de repérage des affiliations CNRS en partenariat avec le Service appui au pilotage scientifique de l'Inist-CNRS et le Service d'appui à la politique et à la prospective scientifique de la Direction d'appui aux partenariats publics du CNRS.

Ces règles strictes permettent d'attribuer de manière quasi certaine le bon identifiant. Des méthodes de classifications supervisées ont été développées et pourront venir augmenter le nombre de web services consacrés aux traitements des affiliations et compléter ainsi les enrichissements liés aux identifiants RNSR.

La figure 1 illustre qu'à chaque ligne d'adresse, et donc à différentes formes d'écriture, associée à une année de publication, le web service attribue un identifiant RNSR grâce à un alignement de l'adresse sur les données du RNSR. Il peut parfois en attribuer plusieurs si sur la même ligne sont renseignés différents laboratoires, ou un laboratoire et une fédération.



En entrée		En sortie
Année	Adresse	Identifiant
2021	University of Bordeaux, IMS, CNRS UMR5218, Talence, F-33405, France	200711887V
2014	IMS Lab, Université de Bordeaux, Talence, 33405, France	200711887V
2021	CNRS, IN2P3, CENBG, F33170 Gradignan, France	199512079F
2014	Université de Bordeaux, Centre d'Etudes Nucléaires de Bordeaux Gradignan, Chemin du Solarium, BP120, Gradignan, F-33175, France	199512079F

FIG. 1: Informations en entrée et récupérées en sortie.

3 Web services : exploitation et étude de cas

Ces services peuvent être utilisés en ligne de commande ou dans des programmes. Un exemple sera rapidement présenté. En effet, le but ici est de montrer un usage simple de ces services sans avoir de connaissances informatiques : ce cas sera exposé à travers l'outil LODEX, outil libre de visualisation développé par l'Inist-CNRS.

3.1 Par lignes de commande

Comme évoqué plus haut, plusieurs moyens pour faire appel à ce web service existent. Le plus direct, quand on a accès à un terminal sous Linux, est sans doute d'utiliser la commande curl. Les données en entrée doivent respecter un format JSON comme indiqué sur la Figure 2.

13. Pour plus de détails : https://github.com/Inist-CNRS/ezs/blob/master/packages/conditor/README.md#règles-certaines

La réponse du service sera dans le même format, le champ id servant à lier une réponse à sa « question ».

```
cat <<EOF | curl -X POST --data-binary @-
"https://affiliations-tools.services.inist.fr/v1/rnsr/json?indent=true"
[
{ "id":1, "value": { "year": "2021", "address": "University of Bordeaux,
IMS, CNRS UMR5218, Talence, F-33405, France" } },
{ "id":2, "value": { "year": "2021", "address": "CENBG, CNRS/IN2P3,
Chemin du Solarium B. P. 120, Gradignan, F-33175, France" } },
{ "id":7, "value": { "address": "Inist-CNRS, UPS76, 2 rue Jean Zay, 54519
Vandoeuvre-lès-Nancy" } }
]
EOF
[{
    "id": 1,
    "value": [
        "200711887V"
    ]
},
{
    "id": 2,
    "value": [
        "199512079F"
    ]
},
{
    "id": 7,
    "value": [
        "198822446E"
    ]
```

FIG. 2: Structure JSON du format d'entrée pour le service RNSR, appel du webservice et réponse au format JSON.

3.2 Avec LODEX

3.2.1 Qu'est-ce que LODEX ?

Développé à l'Inist-CNRS avec des technologies JavaScript, LODEXGregorio et al. (2019) est un outil *open source* de visualisation de données structurées (code source accessible sur GitHub [14] et sous licence CeCILL). Il permet de :

— créer des tableaux de bord et des représentations graphiques à façon (Figure 3) ;
— naviguer dans le corpus importé grâce à des systèmes à facettes (Figure 4) ;
— enrichir les données via soit des transformeurs [15] proposés par LODEX, soit des web services. Si les transformeurs modifient les données après leur import grâce à des scripts, l'appel au(x) web service(s) se fait actuellement lors du chargement et du reformatage des données.

3.2.2 Web services et LODEX

Suite à une demande de la Direction des données ouvertes de la recherche du CNRS (DDOR [16]), le service TDM a développé différents web services dont ceux permettant d'attribuer un identifiant RNSR, puis d'assigner l'institut CNRS correspondant.

Une instance sous LODEX des publications CNRS a été mise en place avec des données issues de CorHAL [17]. Pour y parvenir, un *loader* (module de chargement des données) a été spécialement créé. Il permet de :

14. Code source : https://github.com/Inist-CNRS/lodex ; https://lodex.gitbook.io/lodex-user-documentation/
15. Transformeurs : https://lodex.gitbook.io/lodex-user-documentation/administration/modele/transformers
16. DDOR : https://www.science-ouverte.cnrs.fr/ddor-cnrs-direction-des-donnees-ouvertes-de-la-recherche/
17. CorHAL, continuité de Conditor, collecte, à partir de sources en *open access*, les métadonnées de publications scientifiques françaises et propose l'import automatique du texte intégral dans HAL par le chercheur. Les notices issues de plusieurs sources sont reformatées en un format TEI pivot et enrichies ; les doublons

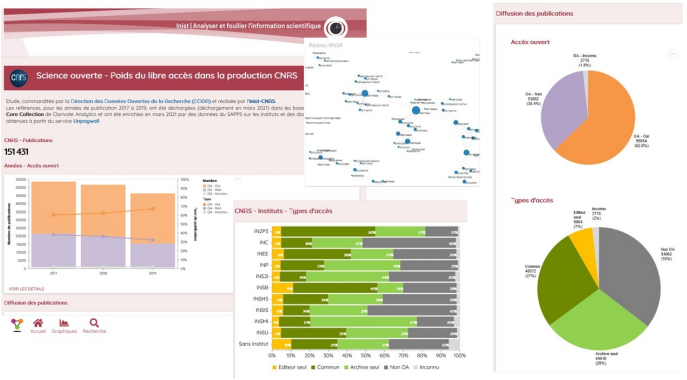


FIG. 3: Représentations graphiques sous Lodex.

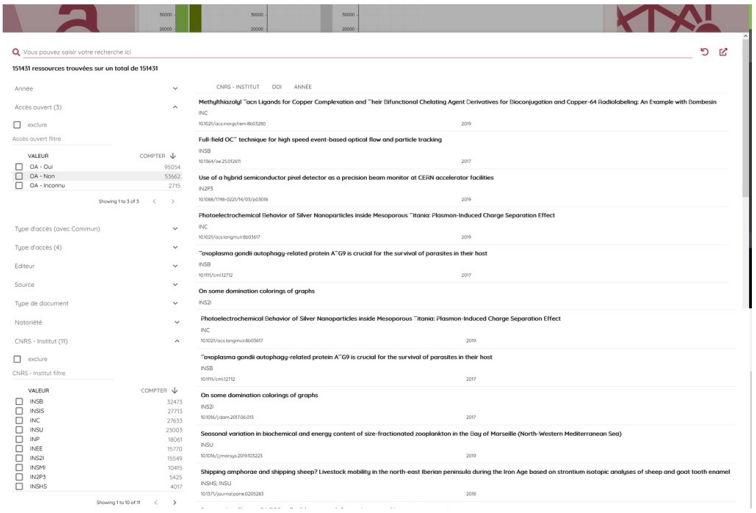


FIG. 4: Recherche par facettes sous Lodex.

— reformater le JSON des notices unifiées CorHAL (Figure 5), sous forme de tableau pour traiter les données (Figure 6) ;
— faire appel à plusieurs web services dont les 2 cités ci-dessus. Une fois l'appel effectué, l'exécution se fait automatiquement lors du chargement des données pour produire des données enrichies.

Une intégration plus facile des web services au sein de LODEX est en cours qui évitera à terme la saisie de l'appel au sein du *loader*. Après import des données, les web services nécessaires seront sélectionnés via un menu déroulant.

sont repérés et permettent la création d'une notice unifiée proposant les informations de chaque notice doublon. Pour plus de précisions : https://wiki.conditor.fr/conditor/index.php/Accueil, https://github.com/conditor-project, https://www.inist.fr/wp-content/uploads/2021/09/Dossier-Poster-CorHAL-VF-13.09.21.png

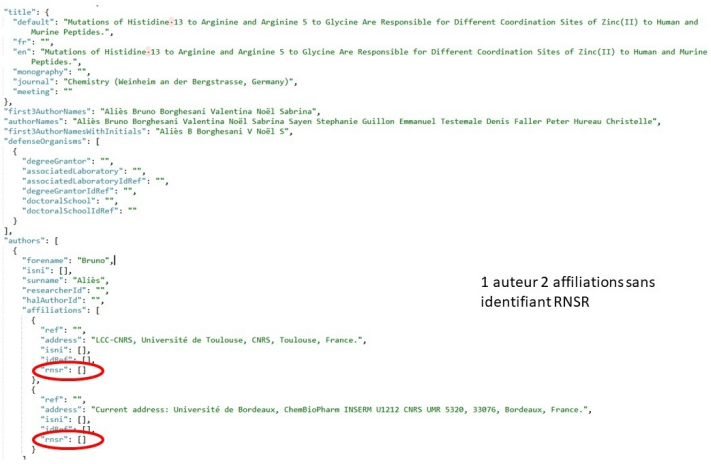


FIG. 5: Notice JSON (sans RNSR).

FIG. 6: Tableau LODEX après reformatage et enrichissement.

4 Conclusion

Le besoin croissant pour les non spécialistes du TDM d'utiliser ces techniques dans le monde de l'information scientifique et technique explique la mise à disposition, par l'Inist-CNRS, de web services. Ces services offrent de nombreuses possibilités et une souplesse d'utilisation avec la spécificité d'un service pour un traitement dédié, et par le fait que les web services sont indépendants de tout langage et système d'exploitation.

Actuellement les web services réalisés pour fonctionner dans LODEX traitent les fichiers document par document. Cela ne permet pas de faire certains traitements qui nécessitent d'analyser un corpus dans son ensemble (clustering ou topic modeling par exemple). La perspective d'évolution est le développement et l'adaptation des méthodes de traitement de flux de données afin de pouvoir offrir à terme ce genre de service.
Les prochains web services seront soit des services génériques, soit des services plus spécifiques répondant à des demandes précises d'utilisateurs. Le déploiement de ces services étant facilité par un environnement approprié mis en place à l'Inist-CNRS, il est facile d'adapter un programme pour le rendre « compatible LODEX », la principale modification étant de prendre

en compte les formats d'entrée/sortie imposés en JSON. Potentiellement tout type de programme de TAL (Traitement automatique des langues) ou de fouille de textes est susceptible d'être utilisé en tant que web service LODEX à condition qu'il respecte le principe imposé, à savoir « un service - un traitement sans paramétrage par l'utilisateur ».

Références

Baneyx, a., P. Breucker, et M. M. Barbier (2009). Création d'une plateforme logicielle pour la maitrise de grands corpus textuels au sein de l'Institut francilien " recherche, innovation et société " (https ://hal.inrae.fr/hal-02817282). In *Atelier de formation aux outils d'analyse textuelle et de visualisation de données, Cortext, ifris*, Champs sur marne, France, pp. 18. Saisissez le nom du laboratoire, du service ou du département., Ville service.

Barbier, M. (2021). Delineating scientific domains in the literature using the CorTexT platform (https ://hal.archives-ouvertes.fr/hal-03367721). In *50° Convegno Nazionale Societ'a Italiana di Agronomia*, Udine, Italy.

Chavalarias, D., Q. Lobbe, et A. Delanoë (2021). Draw me Science - multi-level and multi-scale reconstruction of knowledge dynamics with phylomemies (https ://hal.archives-ouvertes.fr/hal-03180347). working paper or preprint.

Frank, E., M. Hall, G. Holmes, R. Kirkby, B. Pfahringer, I. H. Witten, et L. Trigg (2010). *Weka-A Machine Learning Workbench for Data Mining (https ://doi.org/10.1007/978-0-387-09823-4*$_6$*6), pp.* $1269 - -1277. Boston, MA : SpringerUS.$

Gregorio, S., A. Collignon, F. Parmentier, et N. Thouvenin (2019). LODEX : des données structurées au web sémantique (https ://hal.archives-ouvertes.fr/hal-01990444). In Atelier Web des Données de la 19ème Conférence sur l'Extraction et la Gestion des Connaissances (EGC 2019)*, Metz, France.*

Lobbé, Q., A. Delanoë, et D. Chavalarias (2021). Exploring, browsing and interacting with multi-level and multi-scale dynamics of knowledge (https ://hal.archives-ouvertes.fr/hal-03181233). working paper or preprint.

Maddi, A. et A. De La Laurencie (2018). La dynamique des SHS françaises dans le Web of Science : un manque de représentativité ou de visibilité internationale ? (https ://hal.archives-ouvertes.fr/hal-01922266). working paper or preprint.

Summary

To facilitate access to data mining techniques, particularly for non-specialists, the TDM (Text and Data Mining) service of Inist-CNRS is developing web services around the processing of scientific and technical information. These services can be called from the command line or within LODEX, a free viewing tool. The demonstration shows how, from the information in a bibliographic record and more particularly from an author's address, the RNSR identifier (National Directory of Research Structures) is automatically assigned to the initial document and how this new data is operated within LODEX. Thus, a program or algorithm developed by teacher-researchers could be adapted to become a web service and be used by as many people as possible.