Pour une fédération de dépôts locaux d'articles scientifiques sémantiquement reliés

Jean-Claude Moissinac*

*Telecom ParisTech 46 Rue Barrault, 75013 Paris France
jean-claude.moissinac@telecom-paristech.fr,
https://onsem.wp.mines-telecom.fr

Résumé. Le projet SemBib est une initiative au sein de Telecom ParisTech pour constituer et exploiter une base de connaissances sur nos publications scientifiques. Face à de grands entrepôts de références bibliographiques, nous considérons qu'une fédération de projets analogues à SemBib a du sens. C'est cette position que nous défendons dans ce qui suit.

1 Introduction

Des articles récents montrent un probable doublement du nombre de publications scientifiques tous les 9 ans, comme par exemple Larsen et von Ins (2010). Cette situation impose de se doter d'outils et de méthodes pour naviguer dans cette masse de connaissances, généralement produites à l'issue d'un processus d'évaluation et donc présentant toutes un certain intérêt.

On peut distinguer deux grands types d'accès : l'accès par des interfaces homme-machine plus ou moins puissantes et l'accès par logiciel. Dans ce dernier cas, on trouve deux formes d'accès : par une interface de programmation (API) spécifique ou par un accès SPARQL pour l'interrogation d'un graphe de données.

Une fédération de dépôts d'articles peut partager des outils et des méthodes -analyse de textes, représentations sémantiques...-, tout en développant des approches spécifiques à un groupe ou à une institution.

2 Contexte

De nombreuses initiatives visent à améliorer les parcours dans la masse de connaissances que constituent les publications scientifiques. Certaines s'appliquent à donner une vision analytique d'un ensemble de citations. Citons par exemple le travail de Sateli et al. (2016) pour associer des compétences à des personnes en analysant leurs publications. D'autres, par exemple, aident à trouver des documents pertinents sur un sujet donné comme Rizzo et al. (2015). Cela est souvent fait en cherchant à associer des thématiques à un article ou à un groupe d'articles, par exemple Sateli et Witte (2015).

La disponibilité d'une fédération de dépôts bibliographiques constituerait une contribution significative pour systématiser et simplifier ce type d'approches.

Dans ce qui suit, nous allons présenter quelques ressources significatives qui constituent le contexte de ce projet.

2.1 Les grands référenceurs

Plusieurs acteurs majeurs de l'informatique proposent l'accès à des ressources bibliographiques. Google Scholar est un puissant outil de recherche interactive ; cependant il ne fournit pas d'accès par programme et ne permet donc pas de bâtir des explorations personnalisées.

Microsoft Academic [1] a construit un graphe qui peut être interrogé via une API ; le modèle propose un accès gratuit jusqu'à une certaine limite de requêtes, au delà duquel l'accès devient payant.

2.2 Les éditeurs

Dans cette section, nous allons présenter quelques initiatives d'éditeurs du domaine de la publication scientifique en informatique.

Springer propose un accès SPARQL [2] à des ressources bibliographiques. Cependant, Springer ne donne que des informations de titres, de dates, de mode de publication (actes de conférence, livres). Les auteurs ne font pas partie de la base accessible. De plus, Springer n'a pas fait de choix de vocabulaires qui créent des liens avec d'autres bases : pour l'essentiel, Springer a défini son propre vocabulaire (voir notre analyse [3]).

L'IEEE est une source essentielle d'informations pour les sciences de l'information. L'IEEE propose une API REST pour interroger ses bases [4].

Elsevier propose l'API Scopus et l'API ScienceDirect [5] avec des limites sur le nombre de réponses qu'on peut obtenir.

2.3 Le stockage personnel ou institutionnel

Depuis quelques années, les éditeurs scientifiques autorisent progressivement les chercheurs à rendre accessibles leurs publications soit sur leur site personnel, soit sur le site de leur institution. Cela constitue une source distribuée massive que l'on peut accompagner par des informations sémantiques plus ou moins distribuées.

A titre d'exemples : le MIT propose une API pour ses ressources bibliographiques [6] ; Telecom ParisTech a une base bibliographique publiquement accessible par une interface utilisateur, construite sur la base d'un service REST [7].

1. https ://academic.microsoft.com
2. http ://lod.springer.com/sparql
3. https ://onsem.wp.mines-telecom.fr/2016/12/03/premiers-contacts-avec-lacces-sparql-de-lediteur-springer
4. http ://ieeexplore.ieee.org/gateway
5. https ://dev.elsevier.com
6. http ://libguides.mit.edu/apis
7. http ://biblio.telecom-paristech.fr/cgi-bin/consultform.cgi

2.4 Autres initiatives

ArXiv est une archive de prépublications électroniques d'articles scientifiques. Des limites quantitatives sont placées sur l'accès logiciel [8].

Sudoc est le catalogue collectif français réalisé par les centres de documentation de l'enseignement supérieur. Des services permettent des accès à ces données [9].

De nombreuses autres sources de données existents : HAL, Open Citations, DBLP, Crossref, Libgator, Semanticscholar.

2.5 Conclusion sur les sources disponibles

Chaque source a fait des choix de représentation et de mode d'accès qui en limitent la portée. Plutôt que de créer une nouvelle source qui se voudrait plus exhaustive, l'interconnexion de ces sources peut largement ouvrir les possibilités et permettre de constituer un ensemble de ressources riche ouvrant un large horizon pour une meilleure compréhension et exploitation de la production scientifique, en suivant le modèle du Linked Open Data (LOD). Nous allons présenter cette démarche à notre niveau pour concrétiser cette approche.

3 SemBib

Dans l'esprit de ce qui précède, SemBib est une expérimentation, interne à Telecom ParisTech, préparatoire à un projet pérenne.

3.1 Nos données

Actuellement, nous travaillons sur environ 4000 publications référencées dans notre base bibliographique pour les 5 dernières années. Au-delà des meta-données, seulement 1313 enregistrements contiennent une URL prévue pour donner accès à la publication proprement dite ; cela est notamment du au fait qu'encore récemment les auteurs devaient céder les droits aux éditeurs et ne disposaient pas nécessairement d'un lien direct vers le document en ligne. Sur les 1313 liens disponibles, seuls 400 environ permettent le téléchargement simple du document pour traitement ultérieur. En complément de la sollicitation directe des auteurs, les sources de données citées en section 2 contribuent à la mise en place d'automatismes pour collecter l'ensemble de nos publications.

Seulement un tiers des publications ont des mots-clés associés par les auteurs lorsqu'ils enregistrent leurs publications dans notre base. Moins de la moitié des auteurs renseignent toujours ou quelques fois des mots-clés. Seulement 39 mots-clés sont utilisés plus de 5 fois dans la base. Cette relative faiblesse de notre base nous a incité à collecter beaucoup plus de valeur -mots-clés, concepts, thématiques- directement du contenu des articles.

8. https ://arxiv.org/help/robots
9. http ://m.abes.fr/Espace-Pro-Acces-direct-a/Tous-les-Web-Services

3.2 Nos choix de représentation

Nous avons choisi d'exploiter a minima quelques vocabulaires bien identifiés pour ce type de données. Le Dublin Core est bien sûr un point de départ. Au niveau scientométrique et bibliographique, nous avons trouvé que la famille d'ontologies SPAR (Shotton et al. (2009)) constituait un ensemble solide sur lequel construire ; nous avons notamment appuyé notre choix sur l'analyse de Ruiz-Iniesta et Corcho (2014).

Au-delà des représentations des meta-données, nous avons choisi d'intégrer dans le graphe l'ensemble des 12000 mots non-creux associés aux 400 premiers articles que nous avons traités. L'idée est multiple et pour commencer :

— contribuer à la création interne de liens entre nos publications à travers le vocabulaire utilisé ;
— créer des liens avec des concepts définis à l'extérieur (thésaurus SKOS de l'ACM, concepts dans DBPedia) ; ces liens doivent constituer des passerelles vers d'autres ensembles de données, notamment ceux de la fédération de dépôts bibliographiques dont nous voulons encourager la création.

3.3 Les choix méthodologiques

D'un point de vue méthodologique, le dispositif s'appuie sur des web services et des automatismes d'appel. Lorsqu'un nouveau document est récolté, il est placé dans un dossier spécifique qui est régulièrement scanné. Lorsqu'un nouveau document est détecté dans ce dossier, une copie en est effectuée, ainsi que plusieurs traitements qui produisent des fichiers associés : le texte brut extrait de cette nouvelles source, puis un 'sac de mots' avec la fréquence de chaque mot et le nombre total de mots du texte (ce dernier fichier contribue ensuite aux calculs de TfIdf sur un corpus). Les méta données de chaque document et les mots associés les plus importants sont alors entrés dans notre graphe bibliographique ; celui-ci contient 68 millions de triplets pour décrire 3526 publications. Cela pose des problèmes de performances des requêtes ; nous allons devoir en analyser et traiter les causes.

Nous travaillons actuellement à consolider la description des documents par des concepts en nous appuyant sur des liens avec DBPedia suivant les principes proposés par Tiddi et al. (2015).

3.4 Proposer des visualisations

Plusieurs visualisations ont été produites à partir des résultats de requêtes SPARQL sur notre graphe. Par exemple : le graphe des co-auteurs -colorés par département d'appartenance (Figure 1), le graphe des auteurs partageant au moins deux mots-clefs (Figure 2), une table de similarité sur un groupe d'articles -par exemple, ceux d'un auteur. Chacun de ces graphes révèle des faits : mise en évidence des collaborations entre équipes, mise en évidence de sous-groupes d'auteurs présentant une cohérence...

Le processus de production est assez simple : la requête SPARQL fournit une réponse JSON qui est intégrée à une page HTML exploitant du code Javascript avec la librairie graphique d3js. Nous travaillons à l'automatisation du processus pour décliner chaque modèle de graphique sur différents ensembles de données.

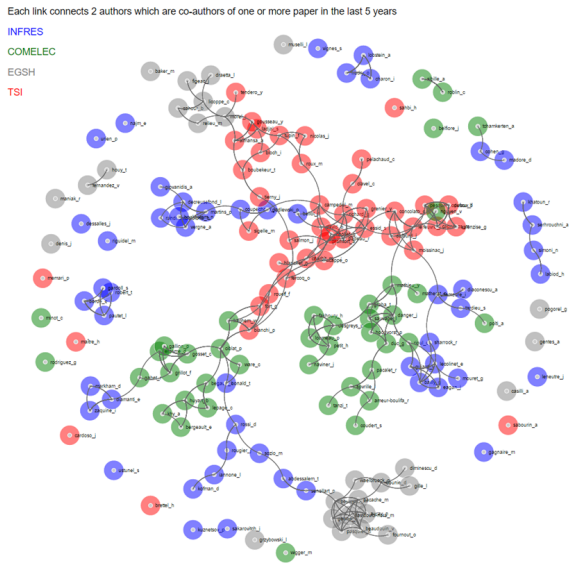

FIG. 1 – *Graphe des permanents co-auteurs*

3.5 Apports du dépôt local

Dans cette section, nous donnons quelques exemples d'apports du dépôt local.

Un problème délicat pour les bases bibliographiques est celui de l'affiliation des auteurs. Les bases en cours de constitution à grande échelle, comme celle de l'Unesco, butent sur ce problème.

A notre niveau, dans les citations d'auteurs de Telecom ParisTech, ont été recensés plusieurs dizaines d'énoncés différents de leur affiliation. Nous pouvons aisément identifier ces différents 'labels' et les qualifier comme étant une affiliation unique. Une fois ce travail fait à notre niveau, il peut se propager facilement sur une fédération de dépôts bibliographiques, à commencer par des dépôts identifiants nos co-auteurs externes. Symétriquement, nous pourrions espérer consolider l'affiliation de nos co-auteurs externes.

Le point de départ du projet Sembib a été l'accompagnement d'un effort de ré-organisation interne des activités de recherche de Telecom ParisTech, notamment en vue de notre implantation sur le campus de Paris-Saclay. Une des questions est l'obtention dynamique d'une cartographie thématique de notre recherche. Aucun des dépôts extérieurs ne nous permettait d'avoir à la fois l'exhaustivité sur nos publications et des précisions suffisantes sur les affiliations, comme l'appartenance à un groupe au sein d'un département de recherche. Ce niveau d'analyse illustre bien l'intérêt d'une représentation locale.

Au niveau d'une institution, il est possible de construire -par apprentissage, par exemple- un vocabulaire et un ensembles de connaissances, de modèles, qui vont contribuer à enrichir et préciser les représentations des publications de l'institution.

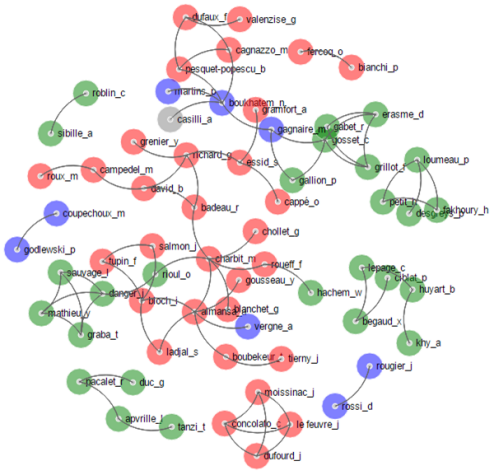

FIG. 2 – *Graphe des permanents partageant des mots-clés*

4 Conclusion

Nous avons vu l'intérêt pour exploiter des publications scientifiques de disposer d'une représentation locale à des institutions ou des groupes de chercheurs. Cette représentation peut faire autorité. Nous avons aussi vu qu'il y a de nombreux travaux qui suggèrent l'intérêt de parcourir des ensembles de publications. Les technologies du LOD apparaissent alors clairement comme un support à l'interconnexion de grands ensembles de données bibliographiques.

Il est nécessaire d'avancer sur les méthodes possibles pour faciliter ces interconnexions d'une part par la promotion de vocabulaires communs, de méthodes -par exemple basées sur des solutions Open Source-, mais aussi par des solutions distribuées de découverte des autres jeux de données -par exemple, chaque jeu de données pouvant référencer quelques autres jeux de données dont il a connaissance. Nous souhaitons que cet article constitue une première étape pour le partage de cette approche.

Nous avons la conviction que cette approche contribuera à l'émergence d'outils puissants pour une meilleure utilisation de la production scientifique.

Références

Larsen, P. O. et M. von Ins (2010). The rate of growth in scientific publication and the decline in coverage provided by science citation index. *Scientometrics 84*(3), 575–603.

Rizzo, G., Tomassetti Federico, A. Vetrò, L. Ardito, M. Torchiano, Morisio Maurizio, et R. Troncy (2015). Semantic enrichment for recommendation of primary studies in a syste-

matic literature review. *Digital Scholarship in the Humanities, Oxford University Press, 13 August 2015*.

Ruiz-Iniesta, A. et Ó. Corcho (2014). A review of ontologies for describing scholarly and scientific documents. In A. G. Castro, C. Lange, P. W. Lord, et R. Stevens (Eds.), *Proceedings of the 4th Workshop on Semantic Publishing co-located with the 11th Extended Semantic Web Conference (ESWC 2014), Anissaras, Greece, May 25th, 2014.*, Volume 1155 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Sateli, B., F. Löffler, B. König-Ries, et R. Witte (2016). Semantic user profiles : Learning scholars' competences by analyzing their publications. In *Semantics, Analytics, Visualisation : Enhancing Scholarly Data (SAVE-SD 2016)*. Springer : Springer.

Sateli, B. et R. Witte (2015). Semantic representation of scientific literature : bringing claims, contributions and named entities onto the linked open data cloud. *PeerJ Computer Science 1*, e37.

Shotton, D., K. Portwin, G. Klyne, et A. Miles (2009). Adventures in semantic publishing : exemplar semantic enhancements of a research article. *PLoS Comput Biol 5*(4), e1000361.

Tiddi, I., M. d'Aquin, et E. Motta (2015). Using linked data traversal to label academic communities. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, New York, NY, USA, pp. 1029–1034. ACM.

Summary

.