SemBib, un dépôt local d'articles scientifiques sémantiquement décrits

Jean-Claude Moissinac[1]
(1) Telecom ParisTech 46 Rue Barrault, 75013 Paris France
jean-claude.moissinac@telecom-paristech.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. Nous présentons ici les avancées actuelles du projet SemBib et ses relations avec d'autres projets.

ABSTRACT
SemBib, local repository of scientific publications with semantic description

MOTS-CLÉS : RDF, web sémantique, RDFa, SPARQL, LOD, publications scientifiques.

KEYWORDS: RDF, semantic web, RDFa, SPARQL, scientific publications.

1 Introduction et 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 and Witte, 2015).

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. D'autres initiatives, comme HAL, proposent un accès de ce type.

L'inconvénient de ces ressources est qu'elles dépendent largement de la qualité des données qu'elles ont pu récolter et sur lesquelles elles ont peu de moyen d'évaluation ; l'amélioration des données concernant une institution s'avère, au mieux, complexe. A titre d'exemple, DBLP contient environ 300 articles associés à Telecom ParisTech et HAL, environ 3000, alors que notre base bibliographique en contient plus de 11000.

1. https ://academic.microsoft.com

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

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 sources locales, dont la qualité serait assurée à petite échelle, proche de la production des données, peut largement ouvrir les possibilités et permettre de constituer un ensemble riche de ressources 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. Cette démarche nous parait très complémentaire de celle portée par le projet européen OpenMinTeD [2]

2 SemBib

Dans l'esprit de ce qui précède, SemBib [3] est une expérimentation, interne à Telecom ParisTech, préparatoire à un projet pérenne.

Actuellement, nous travaillons sur un peu plus de 11000 publications référencées dans notre base bibliographique depuis 1969. Sur un sous-ensemble de 4000 publications initialement étudiées, au-delà des meta-données de base -auteurs, affiliation, date et canal de publication- seulement 1313 enregistrements contiennent une URL prévue pour donner accès à la publication proprement dite. Pour les articles qui n'étaient pas simplement disponibles via un lien, nous avons mis en place un dispositif de collecte automatisée sur le Web. Actuellement, environ 5000 articles ont pu être ainsi récupérés.

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.

Nous travaillons donc maintenant sur environ 5000 publications dont nous avons pu récupérer le texte intégral.

2.1 Nos choix de représentation

Notre choix de base consiste en l'utilisation du modèle RDF qui nous parait le plus à même de permettre une construction progressive de liens entre nos publications ainsi qu'avec des concepts et des ressources externes.

A chaque article est attribué une référence unique (URI) et les propriétés décrivant cette référence sont décrites par des ensembles de triplets concernant cette référence : titre, auteurs, mots-clés, date de publication...

Nous avons choisi d'exploiter a minima quelques vocabulaires bien identifiés pour ce type de données.

2. http ://openminted.eu/
3. http ://givingsense.eu/sembib/

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 and Corcho, 2014).

Pour l'association d'un article avec des concepts, nous avons créé des graphes RDF séparés du graphe qui contient les meta-données de base. Le principe est le suivant :

— créer un graphe de concepts associant une URI de concept à un ou plusieurs labels ; plusieurs graphes ont été créés suivant plusieurs méthodes ; par exemple, pour une méthode simpliste, on retient les mots les plus significatifs d'un corpus de documents, grâce à Tf-Idf, et on crée une URI pour chaque mot retenu,
— créer un graphe qui contient des associations entre chaque article, désigné par son URI, et un ou plusieurs concepts désignés par leur URI ; là encore, nous pouvons utiliser plusieurs méthodes pour créer plusieurs graphes d'association
— compléter les graphes de concepts par des graphes d'équivalences (propriété owl :sameAs) avec des concepts existant hors de SemBib ; nous travaillons actuellement à relier nos concepts avec des concepts de DBPedia suivant les principes proposés par (Mirizzi et al., 2012) ; ces graphes d'équivalence constituent la base de liens avec d'autres ensembles de ressources bibliographiques

Ce choix de représentation dans des graphes séparés permet d'expérimenter avec diverses méthodes, tout en construisant des liens exploitables entre les différentes données.

2.2 Apports du dépôt local et de la représentation RDF

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 (qui explique largement la faible identification de Telecom ParisTech dans DBLP et HAL). 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'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, levant les problèmes liés aux noms successifs de l'institution et permettant d'avoir des informations telles que l'appartenance à un groupe au sein d'un département de recherche. Ce niveau d'analyse illustre l'intérêt d'une représentation locale.

Ces analyses peuvent s'appuyer sur la mise en place de notre point d'accès Sparql [4], qui donne accès à l'ensemble des données, avec la possibilité d'utiliser Sparql comme langage d'interrogation.

De plus, la représentation RDF, nous a permis la mise en place d'un générateur -actuelemment dans une version préliminaire- de pages marquées sémantiquement, ce qui assure une meilleure

4. http ://givingsense.eu/sembib/sparql/

indexation dans les moteurs de recherche. Par exemple, pour l'auteur Gilles Bailly, nous avons la page `http://givingsense.eu/sembib/onto/persons/BAILLY_Gilles`.

Les données associées peuvent être obtenues pour traitement les formats ttl et json ; par exemple, pour json, on a accès à `http://givingsense.eu/sembib/onto/persons/BAILLY_Gilles.json`.

Nous voyons à travers ces exemples qu'un tel projet améliore la qualité des données décrivant notre production scientifique, permet de leur donner une meilleure visibilité et une meilleure accessibilité et fournit les bases techniques pour des traitements et analyses de ces données.

3 Conclusion

Nous avons vu l'intérêt pour exploiter des publications scientifiques de disposer d'une représentation RDF locale à des institutions ou des groupes de chercheurs. Cette représentation peut faire autorité.

Nous avons la conviction que cette approche contribuera à l'émergence d'outils puissants pour une meilleure utilisation de la production scientifique. Nous pensons que notre approche est complémentaire d'approches telles que celle portée par le projet européen OpenMinTeD.

Références

Mirizzi, R., T. D. Noia, E. D. Sciascio, and A. Ragone (2012). Using DBpedia for searching related terms in the IT domain. Technical report, Politecnico di Bari, Via Orabona, 4, 70125 Bari, Italy.

Rizzo, G., Tomassetti Federico, A. Vetrò, L. Ardito, M. Torchiano, Morisio Maurizio, and R. Troncy (2015). Semantic enrichment for recommendation of primary studies in a systematic literature review. *Digital Scholarship in the Humanities, Oxford University Press, 13 August 2015*.

Ruiz-Iniesta, A. and Ó. Corcho (2014). A review of ontologies for describing scholarly and scientific documents. In A. G. Castro, C. Lange, P. W. Lord, and 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, and 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. and 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, and A. Miles (2009). Adventures in semantic publishing : exemplar semantic enhancements of a research article. *PLoS Comput Biol 5*(4), e1000361.