Jeu de données SemBib: représentation sémantique des données bibliographiques de Télécom ParisTech

Jean-Claude Moissinac

Jeu de données SemBib: représentation sémantique des données bibliographiques de Télécom ParisTech

Jean-Claude Moissinac*

*LTCI, Télécom ParisTech, Université Paris-Saclay, 75013, Paris, France
jean-claude moissinac@telecom-paristech.fr,
https://moissinac.wp.imt.fr/

Résumé. Nous allons présenter ici le jeu de données SemBib, représentation sémantique des données bibliographiques de Télécom ParisTech. Ce travail est mené dans le cadre du projet SemBib, au sein de Telecom ParisTech.

1 Introduction

Nous allons présenter ici le jeu de données SemBib, représentation sémantique des données bibliographiques de Télécom ParisTech. Ce jeu de données est en libre accès.

SemBib vise à constituer et exploiter une base de connaissances sur nos publications scientifiques avec plusieurs objectifs : pouvoir mieux exploiter notre production scientifique ; améliorer la qualité de notre base bibliographique ; pouvoir nous interconnecter avec d'autres bases ; disposer d'un jeu de données riche pour nos travaux d'étudiants, mais, aussi, nous l'espérons, pour d'autres expérimentations sur les données bibliographiques. SemBib recense actuellement des metadonnées pour 11311 publications et dispose des textes intégraux de 4939 publications (ces derniers ne sont pas encore accessibles librement).

2 Contexte

Un grand intérêt a été porté ces dernières années pour la création d'outils facilitant l'exploration d'un nombre en croissance rapide de publications scientifiques. Nous avons connaissance de multiples initiatives pour constituer des outils autour de bases bibliographiques : CrossRef, DBLP, HAL, Google ScholarMicrosoft Academic Graph...

Les cas de DBLP et HAL illustrent l'intérêt que nous portons à la constitution d'une base propre. Pour Télécom ParisTech, DBLP ne connait qu'environ 300 publications et HAL environ 3000, alors que nous en avons plus de 11000.

Une analyse de nos publications nous a montré que 53 formes ont été utilisées au cours du temps par nos chercheurs pour indiquer leur affiliation. Le rapprochement entre ces désignations n'est pas chose facile pour un organisme externe. Nous pouvons aisément faire ces rapprochements. Ce type d'observation peut être décliné pour les noms d'auteurs et les canaux de publication. Notre approche de consolidation d'une base interne nous parait être une étape nécessaire pour être bien représenté dans des bases externes comme HAL.

TAB. 1 – *Exemple de représentation de publication (en turtle, préfixe non déclarés pour simplicité)*

```
<http ://givingsense.eu/sembib/onto/tpt/biblio/14557>
        rdf :type               ns0 :ResearchPaper ;
        ns2 :firstAuthor        <http ://givingsense.eu/sembib/onto/persons/Angelini> ;
        ns2 :state              "published" ;
        ns3 :entrytype          "article" ;
        ns3 :fromDpt            <http ://givingsense.eu/sembib/onto/tpt/TSI> ;
        ns3 :fromGroup          <http ://givingsense.eu/sembib/onto/tpt/TII> ;
        ns3 :ref                "YH :TMI-14" ;
        ns4 :creator            <http ://givingsense.eu/sembib/onto/persons/Hoffman> ,
                                <http ://givingsense.eu/sembib/onto/persons/Barr> ,
                                <http ://givingsense.eu/sembib/onto/persons/Angelini> ;
        ns4 :language           "en" ;
        ns4 :title              "Adaptive quantification... kov measure field mode" ;
        ns1 :publicationDate    "2014"|http ://www.w3.org/2001/XMLSchema#integer> ;
        ns2 :venue              <http ://givingsense.eu/sembib/onto/channels/WIMS_14> ;
        ns6 :Abstract           "The extent of pulmonary ... random fields (MRFs)." .
```

3 Choix de représentation

Nous allons ici présenter le modèle de représentation retenu par SemBib.

Au cœur du projet, nous avons un graphe -au sens de la représentation RDF- de représentation de nos publications. Chaque publication est représentée par une entité désignée par une URI à laquelle sont associées un ensemble de propriétés (prédicats selon RDF). Essentiellement, les propriétés utilisées dans ce graphe sont les propriétés intrinsèques d'une publication : titre, auteurs, année de publication, résumé éventuel.

Certaines propriétés ont une valeur littérale, comme le titre, d'autres ont pour valeur une URI vers une entité décrite dans un autre graphe. Ce graphe fait notamment référence au graphe des auteurs et affiliations ainsi qu'au graphe des canaux de publication. La table 1 donne un exemple de représentation.

Nous avons choisi d'exploiter des vocabulaires bien identifiés pour ce type de données. Le Dublin Core est un point de départ. Au niveau 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). Pour cette raison, à l'avenir, les évolutions successives de nos graphes vont intégrer de plus en plus de concepts définis par les ontologies SPAR.

4 Données initiales et données SemBib

Actuellement, nous travaillons sur 11311 publications référencées [1] dans notre base bibliographique depuis 1969. Au-delà des meta-données -titre, auteurs,lieu de publication, année-, la

1. en novembre 2017

base fournit seulement 1313 URL d'accès à la publication proprement dite. Nous avons mis en place des automatismes en vue de collecter l'ensemble de nos publications. A ce jour, environ 5000 textes intégraux ont été récupérés.

L'exploitation de ces données initiales a permis de construire

— un graphe de représentation des publications qui comporte actuellement 263147 triplets [2],
— un graphe de description des personnes et organisations impliquées dans nos publications qui comporte actuellement 50411 triplets[2],
— un graphe de description des canaux de publications que nous avons utilisé ; il comporte actuellement 6599 triplets[2].
— des graphes génériques de concepts à relier aux articles, aux auteurs et aux canaux de publication.

Ces graphes -enrichis au fil des travaux- sont accessibles sur un point d'accès SPARQL [3] et un dump sur github [4] en est assuré périodiquement.

Notre approche a ainsi comme conséquences d'améliorer la qualité des données de notre base bibliographique interne, de nous permettre de disposer d'informations qui n'ont pas leur place dans les bases externes (groupes de recherche ou sein des départements, projets, ...), de nous interconnecter avec d'autres bases sur les principes du LOD -Linked Open Data-, d'être une source de qualité pour alimenter des bases génériques comme HAL, de bénéficier d'une meilleure indexation de nos publications par les moteurs de recherche.

5 Conclusion

Nous avons présenté un jeu de données conçu comme base de travail pour des méthodes de représentation sémantique de données bibliographiques dans le domaine des technologies de l'information. Nous pensons que nos choix pour la représentation ouvrent un large champs pour l'exploration, l'exploitation et l'interconnexion de ces données.

Références

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.

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.

2. au 8/4/2018

3. `http://givingsense.eu/sembib/sparql/` ce point d'accès est utilise actuellement ARC2, qui n'implémente que partiellement SPARQL

4. `https://github.com/moissinac/sembib-graphs`