Latif, Atif; Borst, Timo; Tochtermann, Klaus

Compiling Scholarly Profile Pages by Integrating External Authority Data

Suggested Citation: Latif, Atif; Borst, Timo; Tochtermann, Klaus (2018) : Compiling Scholarly Profile Pages by Integrating External Authority Data, In: JCDL '18 - Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries Pages, Fort Worth, Texas, USA — June 03 - 07, 2018, ISBN 978-1-4503-5178-2, ACM, New York, pp. 411-412, https://doi.org/10.1145/3197026.3204473

Compiling Scholarly Profile Pages by Integrating External Authority Data

Atif Latif
ZBW - Leibniz Information Center for Economics
Kiel, Germany
A.Latif@zbw.eu

Timo Borst
ZBW - Leibniz Information Center for Economics
Kiel, Germany
T.Borst@zbw.eu

Klaus Tochtermann
ZBW - Leibniz Information Center for Economics
Kiel, Germany
K.Tochtermann@zbw.eu

ABSTRACT

Open Access (OA) Repositories provide users with barrier free access to scientific resources and play a significant role in the dissemination of scientific results and the increase of author visibility. Although these repositories are providing free resources, they are not well connected on the level of their metadata. One approach to tackle this issue is to linkup with the scattered pieces of bibliographic and biographic information residing in external sources. In this demonstration, we focus on the contributors of a repository by means of authority data that can be linked to several authority systems (WikiData,DBpedia, VIAF, ORCID, RePEc) to make repository data more diverse, interlinked and visible. With the proposed multistage approach and by developing a proof of concept application, we successfully link a selection of EconStor authors with external identifiers and showcase their aggregated information as a scholarly profile web page.

KEYWORDS

repository, linked data, authorities

1 INTRODUCTION

A well linked open access repository can even provide users with more innovative and unorthodox functionality, i.e. a) providing instant access to internal and external related resources in the way of recommendations[1], b) promoting persons in their role as contributors by keeping their scientific activities up to date [2] and c) fostering collaborations between institutions and contributors[3]. At the current state, these OA repositories are providing freely available scientific content, but not well connected with respect to their objects and metadata items. For example, an author may have published various articles in different repositories, but for users to locate them, they would have to visit and query any of these repositories for a collection of all the papers, potentially receiving duplicates. The same counts for biographical information, i.e. place of work, achievements etc. information which normally does not reside within a single repository, since it requires permanent maintenance and updates. From many decades libraries have used authority control files for the unique identification and better organization of bibliographic data. For instance, the Integrated Authority File (German: Gemeinsame Normdatei, also known as: Universal Authority File) or GND is an international authority file particularly used for the organization of person names, subject terms and corporate bodies from catalogues. On the other hand, Linked Open Data stresses the significance of open and machine readable structured data for unique identification, open data publishing and cross-linkage of (similar) digital resources. Both Linked Data developments and authority control converge with each other to solve the problem of unique identification and reliable linkage of resources across different domains.

2 MOTIVATION

We run our own Open Access repository named EconStor[1] which comprises of more than 150k publications from Economics, most of them being Working Papers as preprints and hosts contributions from more than 100k authors. Although the repository's items are mainly accessed through search engines (in particular Google Scholar), the application provides its own GUI with jump off pages, web statistics and a local search engine. To further promote its content, but also to normalize, to cluster and to enrich the metadata, we decided to introduce author pages into the application. These author pages will reflect both a researcher's local output in EconStor, but also his or her contextual scholarly record. Moreover, we also wanted to validate our rationale that a certain amount of EconStor authors have a valid identifier in terms of an ID, which can serve as a key for further linking and information aggregation. We are particularly focusing on authority name control identifiers due to following reasons:

- Names tend to be ambiguous. It is common that more than one person may have the same name (polysemes), or that a person may be known under different names (synonyms) which make them difficult to identify, authenticate and link to. The identifiers provide added advantage over this problem.
- Initiatives like Wikipedia, Wikidata, Freebase, VIAF and ORCID have already employed many authority control data files for the cross linkage. Linking to these these identifiers will help to enrich author information.

3 RELATED WORK

The enhancement of OA repositories resp. their underlying software with authority data has been a topic since years, though with

[1] https://www.econstor.eu/

JCDL '18, June 3–7, 2018, Fort Worth, TX, USA

ACM ISBN 978-1-4503-5178-2/18/06.
https://doi.org/10.1145/3197026.3204473

no satisfying 'out-of-the-box'-solution yet. Basically, there is work on authority control by means of authority values, and/or by means of authority linking. For instance, the ePrints system, by its major release 3, supports the integration of authority values in terms of internal predefined lists kept in files or database tables[4]. While this approach requires either indexing of only internally available and used values, or external data sources to be continuously replicated into a local environment, it is cumbersome in terms of local efforts to be repeated for each local repository. In that sense, a more sustainable approach was suggested in the context of establishing a CRIS-DSpace[5], where during the submission of a publication a temporary ORCID key is assigned to the [contributor] field inside a DSpace system, so that a local researcher profile can be generated and associated with the publication(s).

4 STUDY DESIGN AND EXPERIMENT

We followed a multistage approach to link the EconStor authors with external identifiers. We used a EconStor dataset dump from April 2016 that is made available as Linked Open data[2]. The data set consists of 111,107 publications with which 64490 authors are associated. To gauge the distribution of GND person identifier in EconStor we acquired the ECONIS database in JSON format. ECONIS is the catalogue of the German National Library of Economics and is completely integrated in the EconBiz search portal[3].It contains more than five million title records.

4.1 Locating the GND of the authors:

In a first step, we used a Perl language script to search for all EconStor publications within an external library catalog, where authors are assigned with a GND identifier. The script returned the handle URL of those publications including author names, associated GND identifiers and other meta data elements. Afterwards, the handle URLs from the results were matched with the publications from EconStor database to retrieve the EconStor publication id and its corresponding author. In a last step, names of authors from the EconStor publication records were matched with the listed GND authors, so that the GND identifier could be added to the EconStor database. With this preprocessing, we were able to answer the following questions:

How many EconStor authors are already associated with a GND identifier (thus identifying an author on the level of a national library system)? We found that 25461 out of 64490 EconStor authors have valid GND identifiers. That is approximately 40% of total authors, who contributed to 69588 publications.

Are the most publishing EconStor author already provided with a GND Identifier? We hypothetically set the publication count number to 25 as a baseline to determine the most publishing authors in Econstor and discovered that 1121 authors have at least 25 publications. By further analysis on recent articulated GND list, we found that approximately 70% i.e., 750 out of 1121 author are associated with GND identifiers showing a good score for our repository power researchers.

[2]http://linkeddata.econstor.eu/beta
[3]https://www.econbiz.de/

4.2 Locating additional identifiers:

The next step of our multistage approach is to query and connect with other external identifier. For this purpose, we repetitively SPARQL query Wikidata endpoint with EconStor author names and GND identifiers. Predominately, we choose WIKIDATA for querying, as it is considered as one of the major connecting hubs to e.g., DBpedia, VIAF, ORCID and REPEC etc. Moreover, it already connects more than 1500 sources of (national)authority data files which may lead to better external linkages. As a result of querying, we found 1780 matches to Wikidata items, 1775 matches to VIAF files, 465 matches to RePEc items, 279 DBpedia resources and 44 matches to ORCID identifiers. Further analysis on the results revealed that most of the prominent EconStor authors have external identifiers. This reconfirms our assumption that prominent authors of EconStor have scattered bibliographical and biographical information within these external identifier system, which can be actively aggregated to build a proof of concept application to showcase corresponding author profiles.

4.3 Proof of concept application

A proof of concept application is developed to showcase the initial idea of an author's profile page. On the start page, users are provided with an index page where all of the prominent authors are listed with their name as hyper-linked. When a user clicks on the link, an author's profile page is generated on the fly with biographical information collected from external identifier. In addition, bibliographical information is compiled from publication lists both from the local repository (EconStor) and a subject portal as external biblographic data source (EconBiz).

5 CONCLUSION AND FUTURE WORK

In this preliminary work, we demonstrated how to introduce author pages by means of automatic linking to other resources. The approach relies heavily on the identification of persons by linking their names to identifier systems, which originally has been an intellectual or manual cataloging workflow. We emphasized on the use of sources for authority data like GND, and showcased how Wikidata can be used as a connecting hub to find author related information. As future work, first we would like to expand the publication list of author by including other external sources, and secondly we would like to investigate the opportunities for enriching Wikidata with missing GND EconStor authors both by editorial means and by bots, the latter to bypass rather static and cumbersome library work flows.

REFERENCES

[1] Borst, T., Lehmann, M., Pianos, T., and Witt, N. (2016). Taking Libraries Cultural Content to the User - Approaches and Experiences from the EEXCESS Project. LIBER Quarterly, 26(3), 125-140.
[2] Latif, A., Afzal, M .T., Helic, D., Tochtermann, K., Maurer, Hermann A. (2010). Discovery and Construction of Authors' Profile from Linked Data (A case study for Open Digital Journal). In Proceedings of the Linked Data on the Web (LDOW2010) (Raleigh, NC, 2010).
[3] Stankovic, M., Wagner, C., Jovanovic, J. and Laublet, P. (2010). Looking for Experts? What can Linked Data do for You? In Proceedings of the Linked Data on the Web (LDOW2010) (Raleigh, NC, 2010).
[4] ePrints (2015), Authority Lists. https://wiki.eprints.org/w/Authority_Lists
[5] Bollini, A. (2016), ORCID Integration. https://wiki.duraspace.org/display/DSPACECRIS/ORCID+Integration