Collecting and Controlling Distributed Research Information by Linking to External Authority Data - A Case Study

Atif Latif(✉), Timo Borst, and Klaus Tochtermann

ZBW - Leibniz Information Center for Economics, Kiel, Germany
{A.Latif,T.Borst,K.Tochtermann}@zbw.eu

Abstract. With respect to the world wide web, scientific information has become distributed and often redundantly held on different server locations. The vision of a current research information system (CRIS) as an environment for constant monitoring and tracking of a researcher's output has become vivid, but still fighting with issues like legacy information and institutional repository structures to be established yet. We therefore suggest to gather those scattered research information through identifying its authors by means of authority data already associated with them. We introduce author pages as a proof-of-concept application collecting research information not only from a local source such as an institutional repository, but also from other external bibliographic sources.

Keywords: Digital library · Research information · Authority data · Linked open data

1 Introduction

Platforms for digital libraries have become the gateway for the dissemination and provision of scientific publications. However, recent advancements in Open Access and Open Science require digital libraries to transform their publication centered model. At current state of affairs, digital libraries are striving for the acquisition and management of supplementary digital science artifacts such as datasets, software and learning resources. Many of these artifacts are managed by open access repositories that foster the integration with a digital library setup [1]. Open access repositories provide users with barrier free access to scientific resources and already play a significant role in the dissemination of scientific results and the increase of author visibility. However, apart from providing freely available resources, these repositories are not well connected with respect to their metadata.

Basically, we see two challenges for traditional information systems relying on biographical information. The first challenge is that we are facing a landscape of distributed research information residing on different server locations, preventing

P. Manghi et al. (Eds.): IRCDL 2019, CCIS 988, pp. 331–339, 2019.
https://doi.org/10.1007/978-3-030-11226-4_26

us from scanning the current information space at a glance. Search engines had been introduced to handle this issue, but they are still quite weak in *identifying* information in terms of authorship or provenance. This leads us to the second challenge: with information often being redundantly published and distributed, it becomes uncontrolled in the sense that it is often unclear which person in fact is accountable for a piece of scientific information. We therefore suggest to tackle these two challenges by linking distributed research information to external authority data that is preferably exposed as linked open data (LOD). Through this study, we want to emphasize that a well linked and connected open access repository can provide users with author related information that promote persons in their role as contributors.

From long, libraries have used authority control files for the unique identification and better organization of bibliographic data. Authority control uses a unique heading or a numeric identifier to identify entities such as persons, subjects or affiliations. It has helped libraries to make bibliographic information less ambiguous and better findable. For instance, the Integrated Authority File (German: Gemeinsame Normdatei, also known as: Universal Authority File) or GND is an authority file particularly used for the organization and cataloging of names, subject terms and corporate bodies.

The program of LOD, on the other hand, stresses the significance of open and machine readable structured data for the purpose of unique identification, open data publishing and cross-linkage of (related) digital resources. Throughout the last decade, it has yielded a bunch of open, structured and interlinked heterogeneous datasets [2]. Moreover, it has inspired various national and international libraries to provide their catalog data [3] and authority control files as LOD. These authority data files were subsequently reused and interlinked by popular LOD hubs like DBPedia, Wikidata or FreeBase. In summary, both developments with respect to authority data and LOD converge to solve the problem of identification and reliable linking of resources across different domains.

2 Related Work

The work of this paper touches several aspects that are subject to ongoing research in information practice and computer science. [4,5] provide an overview of the whole topic, while [6] addresses the topic of authority data for persons from the classic viewpoint of cataloging as a more formal and context independent endeavour. [7] focuses more on potential usage application scenarios where authority data can unfold its full potential. Apart from the decision which descriptive information to be included in an authority record, there are several models for maintaining authority data: from a library-centered approach to automatic clustering to a more community-based effort, giving researchers the opportunity to claim their (suggested) publications [8–10]. With respect to identifiers for researchers and related systems for current research information (CRIS), [11] address the need to integrate internal names with external identifiers, while [12] focus more on additional and proprietary data to be managed by CRIS, neglecting the fact that those systems presuppose a certain data infrastructure that must be established yet. [13] identifies the opportunities for authority resp. library data serving as a backbone for the Semantic Web. Pages with scholar profiles already had been introduced approximately ten years ago by libraries [14], search engines [8] and publishers [10], but with a focus on global visibility and access, rather than local and contextual linking [15]. In that sense, there are efforts to conduct author identification already during the early stage of publishing by assigning a temporary ORCID key to a metadata field inside a DSpace system, so that a local researcher profile can be generated and associated with the publication(s) [16].

3 Motivation

At our institution, we run an open access repository named EconStor [17]. Currently it comprises more than 150k publications from Economics, most of them being working papers. EconStor has contributions from more than 100k authors, with more than 1000 persons contributing more than 20, and 27 persons contributing more than 100 publications (cf. Fig. 1).

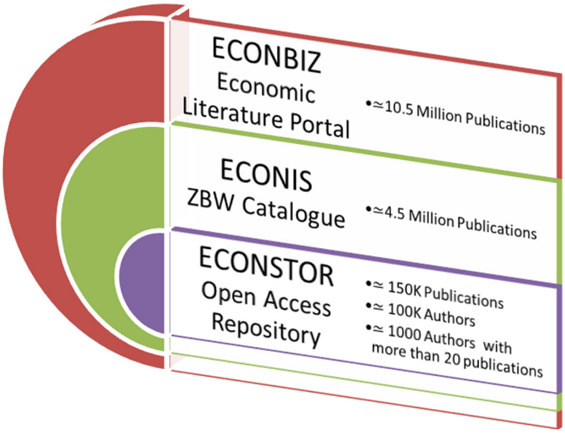
ECONBIZ
Economic Literature Portal
•≃10.5 Million Publications
ECONIS
ZBW Catalogue
•≃4.5 Million Publications
ECONSTOR
Open Access Repository
•≃ 150K Publications
•≃ 100K Authors
•≃ 1000 Authors with more than 20 publications

Fig. 1. Distribution of bibliographic databases in EconBiz subject portal

Although EconStor items are mainly crawled by search engines (in particular Google Scholar), the repository provides its own interface with jump off pages, web statistics and a local search engine. To promote its content even better, but also to normalize, to cluster and to enrich it, we decided to introduce author pages into the application, which reflect both a researcher's local output in EconStor and his or her contextual scholarly record that is compiled from our subject portal EconBiz.

Before we step into the details of our data processing, we want to point out that this work is primarily not concerned with the classic topic of disambiguating persons, e.g. by their publications, citations or co-author networks. Rather, we take a pragmatic attitude towards this issue by identifying those authors which are already associated with an identifier as part of a larger identifier system and maintenance workflow. In the following, we particularly make use of the GND identifier for persons provided by the German National Library, and the handle URLs of EconStor publications.

4 Datasets

Given below are the details of the datasets and authority data which we use in our work.

4.1 EconStor

For the purpose of our analysis and showcase application, we used a EconStor dataset dump from April 2016 that is made available as LOD since 2014 [1]. As of April 2016, the data set consists of **111107** publications with which **218185** author names are associated.

4.2 ECONIS

ECONIS is the catalogue of the German National Library of Economics being completely integrated into the EconBiz search portal [18]. It includes title records for indexed literature and subject specific information procured by the German National Library of Economics. ECONIS contains more than five million title records, most of them manually linked to a GND identifier.

4.3 Integrated Authority File (GND)

According to the German National Library (DNB), the GND is a dataset for describing and identifying persons, corporate bodies, conferences and events, geographic information, topics and works [19]. Centrally provided by the DNB, the data is constantly maintained by several other national and university libraries, requiring a mandate to introduce or to update central information on e.g. a new researcher. By cataloging EconStor authors, they will be associated with a GND identifier according to the general cataloging rules. As a national contribution, the GND dataset is integrated into the VIAF authority file.

4.4 Wikidata

Since its launch in 2012, Wikidata has become one of the major connecting data hubs and has been created to support roughly 300 Wikimedia projects. Despite of interlinking all Wikipedia pages to the relevant items, it also connects more than

1500 sources of (national) authority data files. The biographical information aggregated and linked by Wikidata is taken from different sources, in the first place maintained and updated by editors, if not by the communities or the authors themselves (in contrast to GND, which is maintained only by dedicated supporting library staff). Hence, it may prove to be both the most up-to-date and accurate source despite a certain likelihood to provide wrong or even manipulated data. Linking to Wikidata resp. to other connected identifier systems is a means for associating names with persons. Currently, in total Wikidata has more than 4.29 million persons listed as items, from which more than 509K persons have GND identifier and more than 25K persons are listed as economists (according to Wikidata property p106:occupation) with approximately 11743 represented with GND identifier. This distribution is shown in the scattered Venn diagram (Fig. 2).

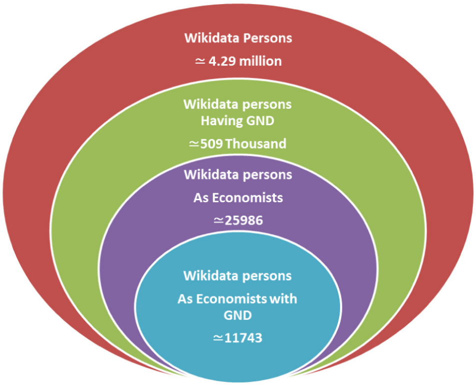

Fig. 2. Wikidata person items (May 2018)

5 Study and Showcase Application

Given below are the details of a multi-stage approach which we employed for linking the EconStor authors with external identifiers. It is important to note that the approach is specific with respect to the local data sources (ECONIS, EconStor) and identifiers it processes. On the other hand, it might serve as an example for globalizing and contextualizing those data sources by means of commonly used identifier systems such as handles or data hubs like Wikidata. An illustration of this multi-stage approach is given in Fig. 3.

5.1 Locating the GND-ID of the Authors from ECONIS Database

To conduct this step, we first transformed the recent ECONIS database from native PICA into JSON format for better processing. We then used a Perl script to search for all EconStor publications whose authors are assigned with a GND identifier, listing them in a data table. The script returns the Handle URL of

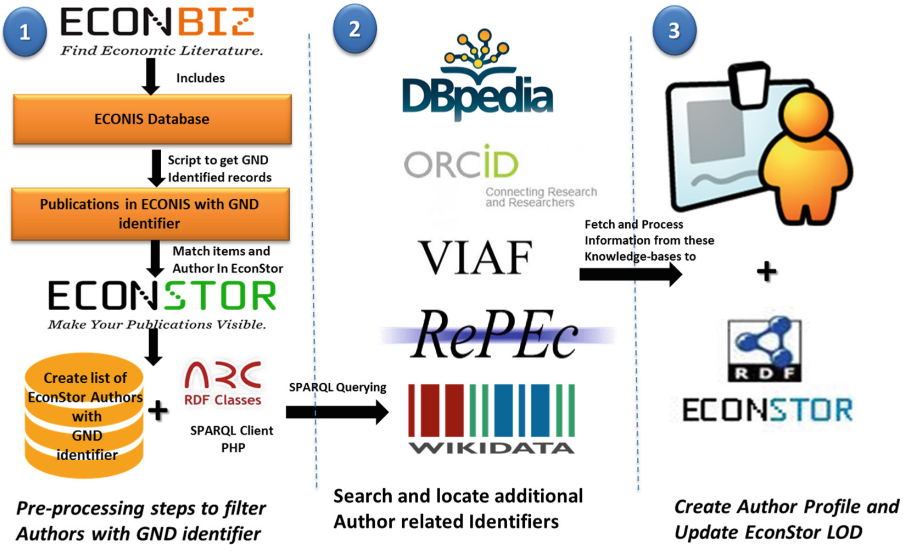
1
ECONBIZ
Find Economic Literature.
Includes
ECONIS Database
Script to get GND Identified records
Publications in ECONIS with GND identifier
Match items and Author In EconStor
ECONSTOR
Make Your Publications Visible.
Create list of EconStor Authors with GND identifier
+
ARC
RDF Classes
SPARQL Client PHP
SPARQL Querying
Pre-processing steps to filter Authors with GND identifier
2
DBpedia
ORCID
Connecting Research and Researchers
VIAF
RePEc
WIKIDATA
Search and locate additional Author related Identifiers
Fetch and Process Information from these Knowledge-bases to
3
+
RDF
ECONSTOR
Create Author Profile and Update EconStor LOD

Fig. 3. Multistage approach for linking EconStor authors with external identifiers

the publications (with the unique Handle prefix 10419 indicating an EconStor publication) together with the author names including their GND identifiers, plus some other metadata elements. The Handle URL is critical here, as it is the only means for mapping publications between EconStor and ECONIS databases.

In a following step, we matched the Handle URLs from the results with EconStor publications to retrieve the EconStor publication id and its corresponding author(s). Afterwards, we reconciled the names of authors from EconStor with the listed GND authors, so that the EconStor authors become assigned a GND-ID (at this stage, outside of the repository application). Thus we were able to answer the following basic questions:

How many EconStor names are already associated with a GND-ID (according to our catalog)?

We found that 114185 out of 218185 EconStor names have valid GND identifiers according to ECONIS records. In distinct numbers, 25461 GND identified authors contributed to 69588 EconStor publications, which already is a good percentage (37%), but still quite incomplete with respect to a total of 111107 publications. Hence we considered introducing author pages at least for the most active researchers in terms of publishing.

Are the most publishing EconStor authors already provided with a GND-ID in ECONIS?

For determining the most publishing authors, we set the number of publication threshold to 25. We found that there are 1121 authors contributing at least 25 publications to our EconStor repository. By applying the same condition to our analysis, we found that 750 out of those most publishing 1121 authors are already associated with a GND-ID (67%), so this looked like a good basis to us.

5.2 Search and Locate Additional Author Related Identifiers via Wikidata

The next step of our multistage approach is to query, search and connect with additional external identifier systems apart from GND. The primary reason for this is to find out the coverage of different identifier systems, and to prove our preference for GND. In our context, the most common additional identifier systems are:

- VIAF as the international authority dataset provided by OCLC,
- RePEc author service as the largest identifier inventory for researchers in economics,
- ORCID as being promoted by the publishing industry, and
- Wikidata as a community endeavour and central data hub aggregating all the aforementioned identifier systems.

We looked up EconStor distinct author names by the following listed SPARQL query to the endpoint provided by Wikidata (to be more precise, we setup a local endpoint by ourselves). As a result, with reference to the baseline of 25461 GND identified EconStor authors, we got 1780 matches to Wikidata items, including 1775 matches to VIAF items, including 465 matches to RePEc items, and including another 44 matches to ORCID identifiers. Overall, approximately 7% of total 25461 distinct EconStor authors are associated with a different identifier system than GND, cf. Fig. 4. By analysing the results again, we found a critical mass of the prominent EconStor authors (more than 25 publications) who have additional external identifiers along with bibliographical and biographical information (most of them from Wikidata and VIAF). This is adequate enough to build a proof of concept application for author profiles which is accessible at: http://beta.econbiz.de/atif/.

Listing 1.1. SPARQL query at Wikidata endpoint to lookup external identifiers for a GND identified author

```
SELECT ?WikidataID ?viafID ?RePEcID ?ORCID WHERE {
 # WIKIDATA–ID of EconStor author ”Dennis Snower” over GND
  ?WikidataID wdt:P227 ”124825109”.
   # request for VIAF ID
  OPTIONAL { ?WikidataID wdt:P214 ?viafID. }
    # request for RePEc ID
  OPTIONAL { ?WikidataID wdt:P2428 ?RePEcID. }
    # request for ORCID ID
  OPTIONAL { ?WikidataID wdt:P496 ?ORCID. }
  }
```

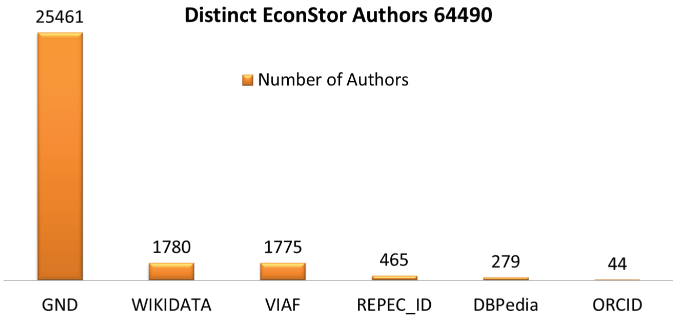

Fig. 4. Number of external identifiers linked with EconStor authors

5.3 Proof of Concept Application

To showcase the initial idea of a scholar's profile page in this proof of concept application, users are provided with an index page where all of the prominent authors are listed with their name and the hyperlink to the corresponding author page. When a user clicks on the link, the profile page is generated on the fly with biographical information queried and collected from external identifier systems. In addition, bibliographical information is compiled from publication lists both from the local repository (EconStor) and our subject portal as external bibliographic data source (EconBiz). An automatically generated author profile page of "Dennis J. Snower" is accessible at: https://tinyurl.com/yc9xd4d8.

6 Conclusion and Future Work

We demonstrated how to collect and reconcile distributed research information by means of automatic interlinking between different biographical sources. The approach relies fundamentally on the identification of persons by linking their names to identifier systems, which originally has been an intellectual or manual cataloging workflow. Hence, the approach does support neither the identification of an author nor the de-duplication of his or her works, but the controlled aggregation across different sources by means of intermediary datahubs. In this respect, we emphasized on the use of sources for authority data such as GND file, and showcased how Wikidata can be used as a connecting hub to find author related information. We are of the view that our study of author data linking service can be a reference enabler for the Digital Libraries to contribute for the digitization of science cause. As future work, 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 both by editorial means and by bots, the latter to bypass rather static and cumbersome library workflows.

References

1. Latif, A., Borst, T., Tochtermann, K.: Exposing data from an open access repository for economics as linked data. D-Lib Mag. (2014). http://www.dlib.org/dlib/september14/latif/09latif.html
2. The Linked Open Data Cloud. https://lod-cloud.net/
3. Yoose, B., Perkins, J.: The LOD landscape in libraries and beyond. J. Libr. Metadata **13**(2–3), 197–211 (2013)
4. Rotenberg, E., Kushmerick, A.: The author challenge: identification of self in the scholarly literature. Cat. Classif. Q. **49**(6), 503–520 (2011). https://doi.org/10.1080/01639374.2011.606405
5. Gasparyan, A.Y., Nurmashev, B., Yessirkepov, M., Endovitskiy, D.A., Voronov, A.A., Kitas, G.D.: Researcher and author profiles: opportunities, advantages, and limitations. J. Korean Med. Sci. **32**(11), 1749–1756 (2017). https://doi.org/10.3346/jkms.2017.32.11.1749
6. Tillett, B.: Authority control: state of the art and new perspectives. In: Tillett, B., Taylor, A.G. (eds.) Authority Control in Organizing and Accessing Information. Definition and International Experience, pp. 1–48. Taylor & Francis, Oxford (2004)
7. Gorman, M.: Authority control in the context of bibliographic control in the electronic environment. Cat. Classif. Q. **38**(3–4), 11–22 (2004). https://doi.org/10.1300/J104v38n03_03
8. Google Scholar. https://scholar.google.de/
9. RePEc Author Service. https://authors.repec.org/
10. ORCID. https://orcid.org/
11. Jörg, B., Höllrigl, T., Sicilia, M.A.: Entities and identities in research information systems. In: EuroCRIS (2012). http://dspacecris.eurocris.org/handle/11366/102
12. Nabavi, M., Jeffery, K., Jamali, H.R.: Added value in the context of research information systems. Data Technol. Appl. **50**(3), 325–339 (2016). https://doi.org/10.1108/PROG-10-2015-0067
13. Neubert, J., Tochtermann, K.: Linked library data: offering a backbone for the semantic web. In: Lukose, D., Ahmad, A.R., Suliman, A. (eds.) KTW 2011. CCIS, vol. 295, pp. 37–45. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-32826-8_4
14. WorldCat Identities. https://www.worldcat.org/identities/
15. Latif, A., Afzal, M.T., Helic, D., Tochtermann, K., Maurer, H.A.: 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 Workshop (LDOW2010), Raleigh, North Carolina, USA, 27 April 2010. CEUR Workshop Proceedings (2010). ISSN 1613-0073
16. Bollini, A.: ORCID Integration (2016). https://wiki.duraspace.org/display/DSPACECRIS/ORCID+Integration
17. https://www.econstor.eu/
18. http://www.econis.eu/
19. http://www.dnb.de/EN/Standardisierung/GND/gnd.html