Limani, Fidan; Latif, Atif; Tochtermann, Klaus

Conference Paper — Accepted Manuscript (Postprint)

Linked Publications and Research Data: Use Cases for Digital Libraries

Suggested Citation: Limani, Fidan; Latif, Atif; Tochtermann, Klaus (2018) : Linked Publications and Research Data: Use Cases for Digital Libraries, In: Mendéz, Eva et al. (Ed.): Digital Libraries for Open Knowledge 22nd International Conference on Theory and Practice of Digital Libraries, TPDL 2018, Porto, Portugal, September 10–13, 2018, Proceedings, Springer, Cham, pp. 363-367,
https://doi.org/10.1007/978-3-030-00066-0_41

This Version is available at:
http://hdl.handle.net/11108/516

Linking publications to research data: bridging intra-institutional siloes

Fidan Limani[1], Atif Latif[1], and Klaus Tochtermann[1]

ZBW - Leibniz Information Center for Economics, Kiel, Germany
~f.limani@zbw.eu, a.latif@zbw.eu, k.tochtermann@zbw.eu

Abstract. Linking publications to ever increasing research data is becoming important for providing a more complete research picture. Siloes of publication and data, even within institutions surely hampers this picture. In this work we explore a linking strategy for scholarly resources – publications and research data.

Keywords: research publications, research data, digital libraries

1 Introduction

Research data (RD) is experiencing a more prominent presence, and rightfully so. Emerging as a 1st class research citizen, especially in the part of research dissemination, there is a need to consider different aspects of RD in order to maximize its research impact. Starting from metadata description provision, to (set of) relevant services development to surround it with, it is necessary to properly handle this new tenant of scholarly communication, similar to what has been done with respect to research publications. The outcome is expected to provide a more complete scholarly communication experience, where research publications, data, and other research-related resources are all easily addressable (to begin with), and provide a more comprehensive research picture.

In order to reap the benefits that RD bring forward, standardization efforts, funding agencies, and many (trans)national initiatives are already pushing for a set of (FAIR) criteria[1] that it needs to abide by. These principles provide a way of anchoring RD to scholarly communication ecosystem (alongside research publications, software, project and funding information, etc.) for a more contextualized use, enabling such elaborate feats such interdisciplinary research prospects, for example.

Digital libraries are already considering extending their services with the addition of RD in their catalogues and services (if they dont have a solution in place already) as there are rationales for sharing data that support this initiative; see Borgman [1] for more. Specifically, establishing links between publications and data is one of the building blocks of a holistic scholarly view something we are aiming at in this paper, focusing on research resources from the domain of economics and related fields.

[1] https://www.dtls.nl/fair-data/fair-principles-explained/

2 Motivation

There are few drivers for our work that enable several use case implementations. Following are the two motivating factors and corresponding use cases in the paper:

Publications and RD that stem from the same research work Since RD archiving is a more recent undertaking as compared to publications archiving, it is not uncommon these collection types to lack links. Usually, there exist parallel archiving initiatives for publications and (later for) RD, and researchers submit their scholarly resources accordingly (and separately!). As a result, often times we find that publication and RD collections are siloed even within the same institution. This prohibits readers a more complete view of research outcome.
The use case focuses on identifying and linking resources from these different collections for a more complete research "experience", supporting cases like measuring the usability of a dataset in a research domain, for example, etc.

Relevant publication-data links Data re-usability is key for research validation and implementation of new use cases, not envisaged in the original paper. In this case, we want to establish links that complement publications (with relevant data), or data (with relevant publications for more context). Same as before, these links will bring together (yes, link!) siloed initiatives within (and across) institutions so that users do not have to search over different collections, platforms, etc.
The use case focuses on identifying and linking resources from the RD collection based on criteria of interest (publication date, publisher, topic, etc.). Recommending RD across institutions for interdisciplinary research could be another useful case.

3 Related Work

Cross-linking scholarly resources is already in the focus of multiple initiatives. From encoding link semantics, to applying linking technologies, solutions rely on ontologies, metadata standards, persistent identifiers, user-provided classification terms, and more to further structure this linkage (see Mayernik et al. [2]; Aalber et al. [3]; and [4]).

Projects with different scopes also exist that establish links between scholarly resources. RMap relies on Semantic Web[2] and Linked Data[3] to model the relationships between such resources, extending beyond publications and data (see Hanson et al. [5] for more). Similarly, ResearchObject[4] includes multiple initiatives and relies on several mechanisms to represent scholarly resources as a bundle that can be accessed as a whole, all in a machine-readable way, with the final result of having it published according to FAIR principles.

[2] https://www.w3.org/standards/semanticweb/
[3] https://www.w3.org/standards/semanticweb/data
[4] http://www.researchobject.org/

Standardization efforts that propose solutions at a general level also exist. Such is the case with the Scholix Framework[5]. Driven by the RDA/WDS Publishing Data Services Working Group[6] and its partners, it represents a set of guidelines that foster interoperability between scholarly resources. The Data-Literature Interlinking (DLI) Service represents an implementation instance adhering to these guidelines, and currently offers more than 8 M links, as well as services to access the collection in different ways (see Burton et al. [6, 7]).

4 Methodology

In this section we present the methodology: dataset selection, metadata set supporting our use cases, and workflow that supports our approach.

4.1 Data collection

As mentioned earlier in the paper, often times there are silos of collections (pulications, research data, regardless) even within institutions. For our study we select two collections, operated by a single organization - the ZBW[7]:

- Journal Data Archive (JDA): a ZBW project that targets research data from journals in the domain of economics. Researchers that have data (raw data, scripts or implementation code, etc.) to share, upload it to JDA, and these entries with assigned DOI's become findable and available. Interested journals in economics, in a way, "delegate" data storage and sharing responsibilities to the JDA. The current collection of JDA includes 66 datasets from different economics journals.
- EconBiz is a publications portal that focuses on the domain of economics. It supports many types of publications that researchers can store there, such as conference or journal papers, book chapters, "work in progress" papers, etc. With well over 10 M publications across participating databases and its set of services, it offers a great support to researchers in finding relevant publications.

4.2 Metadata observation of target collections

Although driven by different requirements, there is a set of metadata elements describing both JDA and EconBiz collections that we take into consideration during harvesting/accessing and matching operation. Following are brief information these elements.

- EconBiz metadata elements:
  Title; *ResourceProvider*; *Creator/Person* The entity that authored the resource; *Date*; *Identifier*; *Publisher*; *Subject* A terms describing the resource

[5] http://www.scholix.org/home
[6] https://rd-alliance.org/groups/rdawds-publishing-data-services-wg.html
[7] https://www.zbw.eu/de/

topic; *Type* and *Type_genre* The nature or genre of the resource, and narrower categories, correspondingly; *Identifier_url* Electronic address for accessing online resource.
- JDA metadata elements: *Title*; *ResourceProvider*; *Dara_authors*; *Dara_PublicationDate*; *Identifier*; *Organization*; *Tags* and *Num_tags* The terms (and number) used to describe the resource; *Dara_jels* Terms from JEL[8] describing the resource. These terms are already mapped to the ones used for publications in EconBiz; *Type* A controlled list for denoting the type of the resource. `dataset` is an example value; *Notes* A brief description of the dataset (research aims, methodology used, etc.). Its application is optional at the moment as we do not see a need of this element in any of our use cases at this phase. *Num_resources* The number of files constituting the dataset. This is another attribute that could support a use case in the future, especially during filtering results, where the users will have the opportunity to focus on datasets with more/less/certain number of files for a dataset entry.

The "subject" in EconBiz and "tags" in JDA are aligned: the STW[9] thesaurus that is used to describe EconBiz resources is aligned with JEL classification, used to describe JDA resources. This represents a nice opportunity to identify relevant resources across both collections, either by direct matching, or by narrowing/broadening search criteria, depending on the results. It is important to note, thought, that few of the metadata elements are mandatory when registering RD in repositories.With JDA dataset, there were cases that authors, JEL classification terms, or other elements were missing. As a result, we had to rely on as few metadata elements as possible in implementing our approach – the title and, at times, classification terms.

4.3 Approach workflow

Figure 1 depicts our publications-to-data linking workflow:

Harvest collections: Both collections provide REST interfaces; JDA entries are stored as JSON files, whereas the EconBiz search is conducted on-the-fly for every JDA entry, with only the highest-ranking result considered for matching.

Establish links: Having in mind the size of both collections, we start the process from the JDA collection. For every entry of the collection, we search for an entry in EconBiz that contains the same publication title (Use case 1). In case such a link cannot be established (the original publication that used the data is not hosted on EconBiz), we use other attributes to conduct search, such as subject terms used to classify resources in both collections (Use case 2). Other search scenarios are also possible: filtering candidate links based on publication year, publisher, access policy (open or closed), etc., are all viable refinements of this part.

Link re-use: Once established, there is great potential for sharing the publication-data links. Adopting Scholix Framework principles, or Linked Open Data as a

[8] https://www.aeaweb.org/econlit/jelCodes.php
[9] http://zbw.eu/stw/version/latest/about

publication medium, or any other method mentioned in the related work, are just few of the options that could further increase the impact of the links. Our approach recognizes and plans for this aspect, too.

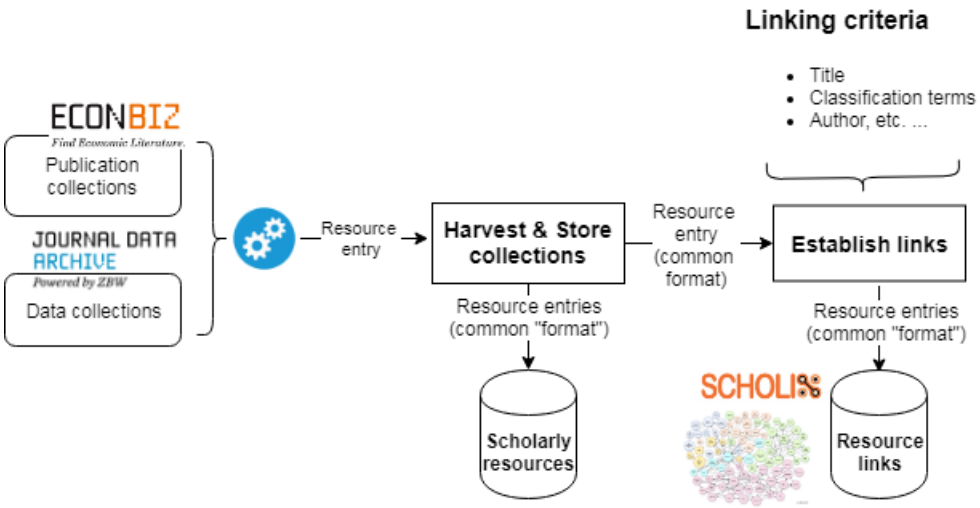

Fig. 1. Publication-to-data workflow

5 Results

Resources linking: For the first use case (publications-data linking) our approach matched 28 (out of 66) JDA entries (around 42%) to publications in EconBiz. For the second use case, we successfully searched for RD based on certain criteria. For example, finding RD that correspond to Germany, retrieves 8 such resources from the JDA collection. The search criteria can surely vary. We could as easily track RD publication by country, institution, publisher; specify the research domain and publication date, or uniquely address certain dataset; to name but a few scenarios.

Link re-use: Once the publication-data links are established, there are opportunities for their re-usability. Besides institutional repositories (e.g., recommending relevant resources that link together), the links themselves represent a valuable asset to external parties. For example, sharing the links according to the Scholix Framework principles is a way to increase the exposure of linked resources beyond the scope of local repository.

6 Conclusion and Future Work

In this paper we showcased the approach to link scholarly publications and research data which resides in intra-institutional siloes. We initially identify publications-data links based on the title of entries in the collections. In the future, as other metadata elements get more present – and mandatory (such as JEL terms), semantically aligned (such as identifiers in both collections), our approach will adapt accordingly. Moreover, extracting important terms from the title, using controlled vocabularies to further expand/narrow the term, etc., enables more capable search for related publications.

On the short term, quantify these links presents an interesting goal: There are many more metadata elements that could be used to realize new use cases that further explore the value-adding effect of scholarly resources linking. Moreover, based on these links, we can auto-complete missing metadata values from either resource after matching (if the title matches, chances are that authors, publication date, etc., are the same).

When it comes to data selection, there is a potential for broadening the scope: including external research collections, targeting domain-specific repositories, is one of the immediate dataset extensions planned next.

References

1. Borgman, C.L., 2012. The conundrum of sharing research data. *Journal of the Association for Information Science and Technology, 63*(6), pp.1059-1078.
2. Mayernik, M.S., Phillips, J. and Nienhouse, E., 2016. Linking publications and data: Challenges, trends, and opportunities. *D-Lib Magazine, 22*(5/6).
3. Aalbersberg, I.J., Dunham, J. and Koers, H., 2013. Connecting scientific articles with research data: New directions in online scholarly publishing. *Data Science Journal, 12*, pp.WDS235-WDS242.
4. Nüst, D., Konkol, M., Schutzeichel, M., Pebesma, E., Kray, C., Przibytzin, H. and Lorenz, J., 2017. Opening the Publication Process with Executable Research Compendia. *D-Lib Magazine, 23*(1/2).
5. Hanson, K.L., Di Lauro, T. and Donoghue, M., 2015, June. The RMap Project: Capturing and Preserving Associations Amongst Multi-Part Distributed Publications. In *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 281-282). ACM.
6. Burton, A., Koers, H., Manghi, P., Stocker, M., Fenner, M., Aryani, A., La Bruzzo, S., Diepenbroek, M., Schindler, U. and Authr, C., 2017b. The Scholix Framework for Interoperability in Data-Literature Information Exchange. *D-Lib Magazine, 23*(1/2).
7. Burton, A., Koers, H., Manghi, P., La Bruzzo, S., Aryani, A., Diepenbroek, M., Schindler, U., 2017a. The data-literature interlinking service: Towards a common infrastructure for sharing data-article links, Program, Vol. 51 Issue: 1, pp.75-100, https://doi.org/10.1108/PROG-06-2016-0048.