Latif, Atif; Hoefler, Patrick; Tochtermann, Klaus

Conference Paper

Interlinking Scientific Authors with the LOD Cloud. A Case Study

Suggested Citation: Latif, Atif; Hoefler, Patrick; Tochtermann, Klaus (2012) : Interlinking Scientific Authors with the LOD Cloud. A Case Study, In: Networked Digital Technologies 4th International Conference, NDT 2012, Dubai, UAE, April 24-26, 2012, Proceedings, Part II, ISBN 978-3-642-30566-5, Springer, Heidelberg, pp. 99-108

Interlinking Scientific Authors with the LOD Cloud: A Case Study

Atif Latif[1], Patrick Hoefler[2] and Klaus Tochtermann[1]

[1] ZBW - German National Library of Economics Leibniz Information Center for Economics, Kiel, Germany
˜A.Latif@zbw.eu, K.Tochtermann@zbw.eu
[2] Know-Center Graz, Austria
˜phoefler@know-center.at

Abstract. Linked Data has played a vital role in the realization of the Semantic Web on a global level. It motivates people to publish datasets which can be important for information linking and discovery and can further make contributions in streamlining the Web as a single connected data space. This effort has successfully amassed a variety of Linked Data and has introduced many novel ways for the publishing of data. As a result, putting Linked Data online has become rather easy, but actually linking the data with already existing data in the cloud is still a challenge. The search and identification of relevant datasets as well as devising a strategy for linking to these datasets is still a difficult task. In this paper, a novel approach is presented which highlights and implements the steps involved in the interlinking process. This approach is further applied and presented as a case study focusing on interlinking scholarly communication datasets and highlighting the potential benefits.

1 INTRODUCTION

Linked Data is about ensuring best practices for producing and sharing structured data in a way that is understandable and processable by machines[1]. The main emphasis of this effort is to make data available openly as Linked Data in order to get added value and construct intelligent services, which previously was not possible due to walled gardens that gave no benefits and prevented access for machines and people. The ultimate goal of the Linked Data effort is to create a globally connected data space where related information is better connected, enabling both simple and sophisticated queries as well as intelligent web services. The W3C community project *Linking Open Data*[2] was founded in 2007. It is based on the Linked Data principles, four simple rules stated by Tim Berners-Lee[3]. They are:

1. Use URIs as names for things
2. Use HTTP URIs so that people (and machines) can look up those names (see also[4])
3. When someone looks up a URI, provide useful information

4. Include links to other URIs so that they can discover more things

Basically, these rules emphasize a set of practices to publish data in RDF[5] by giving each data chunk a unique URI, which is further deferenceable to present more meaningful information. By following this set of practices, any individual or organization can open up their datasets as structured data and can interlink with other datasets to bring more value to their datasets. Some of the potentials of exposing data as Linked Data are[6]:

→ Every concept has a unique identity (URI) in the document which is further discoverable, reusable and linkable.
→ It removes the data silos, turning the Web into a connected Giant Global Graph.
→ All concepts are modelled by a single Resource Description Framework (RDF), bringing in consistency to structured data representation and leading to interoperability at various levels.
→ Complex questions can be asked using a querying language called SPARQL[7]. Queries from different interconnected datasets can lead to the discovery of hidden patterns and relationships.
→ Increase in value and visibility of data by interlinking with external data resources.

In anticipation of these benefits, a growing number of linked datasets as well as supporting tools and techniques are emerging rapidly. The heavy presence of Linked Data in various domains (e.g. medicine, government, geo-information, and scholarly communication)[3] offer heaps of open and linked data for research and mash-ups.

In a nutshell, the main idea of the Linked Data efforts so far is: Structure the data (using RDF), put it on the Web, and include semantic links to other data in order to create a *Giant Global Graph*[6]. Whereas the first two steps are rather easy to accomplish and are already covered by many case studies, the task of finding relevant external data to link to is still a challenge and lacks case studies. The challenges usually involved in the interlinking process are:

→ Searching and identifying the datasets that are candidates for the interlinking
→ Understanding the underlying ontologies and semantic structures
→ Looking for availability of live SPARQL endpoints
→ Adequate knowledge of SPARQL querying and semantic technologies as well as expertise in using SPARQL endpoints or dumped RDF datasets

In this paper we focus on these challenges and investigate the interlinking possibilities in the Linked Open Data (LOD) cloud. This study depicts a concrete use case where a relational database of scientific authors was "RDFized" and the authors were semi-automatically linked with an external database of scientific authors. For that purpose we have searched, identified, queried and made

[3] http://richard.cyganiak.de/2007/10/lod/

use of available scientific scholarly datasets, i.e. DBLP[8], RKB Explorer[9] and Semantic Web Dog Food[10].

This paper starts with a short overview of the state of the art and related literature. Then, the datasets which were investigated for this study are discussed in the test dataset section. In the next section, the actual use case and the technical implementation are described in detail, followed by the results of the automatic interlinking. The paper closes with a discussion of the approach and an outlook on future research.

2 RELATED LITERATURE

The concept of Linked Data is maturing, and though many problems are still unsolved, most of them are at least known. Tom Heath and Chris Bizer[6] summed up the current state of the art and looked at the problem of making links with external data sources. Regarding auto-generating RDF links, they gave an overview of key-based as well as similarity-based approaches.

The state of the art in this paper is divided into two parts, which are 1) Related Tools and Studies and 2) Services for Interlinking.

2.1 Related Tools and Studies

SILK: One of the popular similarity-based approaches is SILK. The SILK framework[11] provides a set of services, which are used to discover relationships of resources within different linked datasets. By using SILK (Link Specification Language), data publishers can specify the type of RDF links that need to be present in im the linked dataset. Additionally, the conditions and restrictions that should be validated during the process of interlinking can be specified. The SILK framework works on data sources that are interlinked with the SPARQL specification. The SPARQL endpoint is also made available for the community. To use SILK, however, one needs to be an expert. Moreover, if users want to convert from a relational database, they need to take care of the conversion into RDF themselves.

Google Refine Extension: Fadi Maali et al.[12] examined several approaches to implement reconciliation services in order to link data to so-called Linked Open Data hubs as part of the data publishing process. They also described their implementation in an extension to the data workbench application Google Refine[13]. Google Refine is a open source power tool which works with messy data, discovers, experiments, transforms, extends and links it to databases and knowledge bases like Freebase[14].

Linking Music Data: Yves Raimond et al.[15] looked at the problem of interlinking music-related data sets on the Web. They described an interlinking algorithm that takes into account both the similarities of web resources and of their neighbours.

Linking Open Journal Data: In another case study, Latif et al.[16] worked on the interlinking of open digital journal data with the LOD cloud by extending the CAF-SIAL application[4]. CAF-SIAL is a proof of concept system to discover and present informational aspect of resources describing people from Linked Data[17]. It is based on a methodology for harvesting a person's relevant information from the gigantic LOD cloud. The methodology is based on combination of information: identification, extraction, integration, and presentation. Relevant information is identified by using a set of heuristics. The identified information resource is extracted by employing an intelligent URI discovery technique. The extracted information is further integrated with the help of a Concept Aggregation Framework. Then the information is presented to end users in logical informational aspects. This system is currently used by the Journal of Universal Computer Science[5] and has successfully interlinked the journal authors with the LOD cloud.

2.2 Search Services for Interlinking

A few of the currently available services for finding relevant material from the LOD cloud are introduced next.

SINDICE Sindice[18] provides indexing and search services for RDF documents. Its public API allows forming a query with triple patterns that the requested RDF documents should contain. Sindice results often need to be analyzed and refined before they can be directly used for a particular use case. Similar kinds of services are provided by semantic search engines like Falcon[19] or Swoogle[20]. Sindice is used in this study, mainly due to its larger indexing pool and the ease provided in use of public API.

SameAs SameAs[6] from RKB explorer[9] provides a service to find equivalent URIs annotated with *owl:sameAs* links in Linked Data datasets. It facilitates finding related data about a given URI from different sources; however, it is necessary to know the exact URI beforehand. The SameAs API returns result in multiple formats.

3 KNOW-CENTER TEST DATASET

For this experiment, our main test dataset is based on a Know-Center[7] publication dataset provided openly as RDF. The Know-Center is Austria's competence center for knowledge management and knowledge technologies, founded in 2001.

[4] http://cafsial.lod-mania.com/
[5] http://www.jucs.org/
[6] http://sameas.org/
[7] http://know-center.tugraz.at/

We decided to link two important assets of the Know-Center – authors and papers – with the external datasets in the LOD cloud to 1) enrich Know-Center's publication dataset, 2) interlinking additional resources within Know-Center's corpus and 3) increase the visibility of authors and papers in the Linked Data sphere for discovery and interlinking. In this study we have worked with 296 authors and 524 publications from Know-Center's publication dataset.

4 EXTERNAL DATASET SELECTION

Today, a lot of scientific publishing services provide online access to journal and conference publications. Many digital libraries and repositories have developed archives with information about conferences, journals, authors, and papers, searchable by keyword, category, and publishing year. These resources are very helpful to scientists and researcher. However, in the context of Linked Open Data, there are only a few services which provide a semantic representation of these scientific resources. Before being able to use the scientific publishing data in the form of Linked Data, it is necessary to find all the available data sets in the LOD cloud. The CKAN initiative[8] is currently building a comprehensive directory of (linked) data repositories, which should prove helpful in locating relevant repositories in the future.

We have identified and selected six Linked Data bibliographic datasets which are described next.

4.1 DBLP D2R L3S and FU BERLIN Server

The DBLP D2R L3S server[9] is based on the XML dump of the DBLP database. The DBLP database provides bibliographic information on major computer science journals and conference proceedings. The database contains more than 800.000 articles and 400.000 authors[21]. To query the DBLP L3S data set, the D2R Server, a semantified version of DBLP bibliography, was accessed via its SPARQL endpoint. Another DBLP bibliographic dataset which we considered for our study is DBLP D2R Server FU Berlin[10]. The FU Berlin dataset is also available as Linked Data and accessible via its SPARQL endpoint.

4.2 DBLP, CITESEER and ACM RKB Explorer

Other important and popular bibliographic datasets which we considered for this study were Citeseer[11], ACM[12], and DBLP[13]. All these datasets are published as Linked Data by the RKB[9]. RKB explorer is a service which, after applying

[8] http://ckan.net/
[9] http://dblp.l3s.de/
[10] http://www4.wiwiss.fu-berlin.de/dblp/
[11] http://citeseer.rkbexplorer.com/
[12] http://acm.rkbexplorer.com/
[13] http://dblp.rkbexplorer.com/

their co-reference mechanisms with enriched ontology description, has provided these mentioned dataset as SPARQL endpoints.

4.3 Semantic Web Dog Food Dataset

We considered this dataset due to its repository providing newly conducted conferences data related to the research field of the Semantic Web.

4.4 Dataset Selection Process

After manually identifying the repositories relevant for this use case – DBLP (L3S), DBLP (RKB Explorer), DBLP (FU Berlin), ACM (RKB Explorer), Citeseer (RKB Explorer), and Semantic Web Dog Food – a selection process was required to decide which repository would yield the most potential for the enrichment of the author data set at hand. We manually queried these repositories to get the publication count for the three authors of the dataset with the most publications – *Klaus Tochtermann*, *Michael Granitzer* and *Stefanie Lindstaedt*. After comparing the results, we selected the DBLP L3S data set for interlinking due to its large index and availability of maximum results. The details of manual investigation about these datasets is illustrated in Table 1.

Table 1. External Dataset Selection

Repository Name	Granitzer	Lindstaedt	Tochtermann	Total	Comment
DBLP L3S	49	60	55	164	Large corpus and up to date dataset
DBLP RKB	27	46	30	103	Multiple URIs for each author
DBLP FU-Berlin	10	20	43	73	Limited dataset, November 2006 version of the DBLP dataset
ACM RKB	4	7	8	19	Multiple URIs for each author
CITESEER RKB	0	1	8	9	Multiple URIs for each author
SWDF	0	2	1	3	Too few results

5 STEPS IN INTERLINKING SCHOLARLY DATA

For the interlinking of the Know-Ceter publication dataset, a multi-step strategy was devised to find similar and related resources in form of authors and papers. These discovered resources were further processed for interlinking of resources with *owl:sameAs* relationship. The framework for this strategy is illustrated in Figure 1. The strategy looked as follows:

5.1 Know-Center Author and Publication Acquisition Service

The names and URIs of the authors from the Know-Ceter that should be interlinked with the LOD cloud were retrieved via a SPARQL query from Know-Center SPARQL endpoint[14] with the help of a web service. Then all the publications and co-authors of the authors in question were queried via SPARQL and stored in a relational database for further processing.

5.2 Search Service

The Sindice Search API[15] was used to search for URIs of the authors in question. We wrote a web service which took the authors name iteratively as an input and automatically called the API with formulated search queries. The resulting URIs were then filtered automatically on the basis of heuristics to make sure that they belonged to the DBLP. In this process, URIs for 91 out of the 296 authors in question were found. Keeping in mind the limitations of Sindice – basically the same recall and precision problems that every search engine displays to a certain degree – we additionally employed a string-matching algorithm that compared the names of the authors in question with author names from the DBLP data set. This way we found a few additional URIs which had not been retrieved by Sindice in the previous step.

We combined the results from steps 1 and 2 and constructed a relational database which stored the names and URIs of each author.

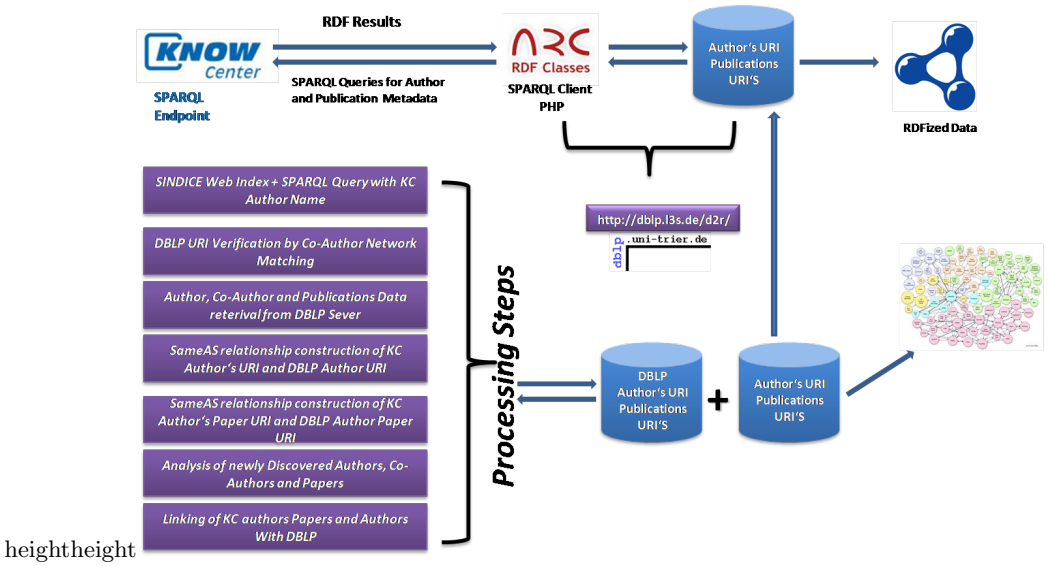
5.3 Know-Center Author's DBLP Co-Author and Publication Acquisition Service

For the acquisition of the publications and co-author of the previously tentatively matched authors we queried the DBLP SPARQL endpoint with the help of our web service specifically designed for DBLP dataset. This step provided us with the discovery of additional publications and co-authors of the Know-Center authors. Futher on, we stored these results in our local relational database with the name of DBLP data tables for the validation process.

5.4 Validation Service

In order to ensure the validity of the matched URIs, a validation service was written. This service automatically took the co-author network of individual authors from the locally stored Know-Center and DBLP databases as an input and compared them for each author. This step helped us to drop incorrectly matched authors who shared the same name but were actually different people from different disciplines or research areas. In the end, the names of authors and publications from the local data set and the data set from the DBLP were compared once again, and the *owl:sameAs* relationships between authors as well as publications were established and published.

[14] http://know-center.tugraz.at/sparql/
[15] http://sindice.com/developers/searchapi



heightheight

Fig. 1. Discovery Framework for Interlinking of Scholarly Data

6 RESULTS AND DISCUSSIONS

In this section the results achieved our by multi-step strategy are discussed. A local data set with 296 authors and 524 publications was semi-automatically matched with a remote data set based on data from the DBLP and provided by the L3S. Using the Sindice Search Engine, 97 DBLP URIs were found where the author names had a Levenshtein distance of less than 4.

By querying the DBLP SPARQL endpoint using the authors names, 85 URIs were found. After combining both the results of Sindice and the direct SPARQL query, the total number of matched URIs was 111. Of those, 72 were found by both Sindice and the SPARQL query, 25 only by Sindice and 14 only by the SPARQL query. Out of the 111 identified matches, 67 were validated using co-author network matching. This was done by comparing the co-authors of every author retrieved from the local data set and the DBLP data set. Due to differences in spelling (e.g. German umlauts), a Levenshtein distance of less than 3 was considered a match. If an author had less than 5 co-authors, then at least 2 of them had to match. In there were more than 5 co-authors, then at least one third of all co-authors of this author had to match to successfully validate the author.

Besides the authors, publications were also matched. Out of 524 local publications, 161 were found in the DBLP. Publications with a title length of more than 20 characters and a Levenshtein distance of less than 13 were considered a tentative match. In the next step, the authors of each publication were considered a match if the author name was longer than 8 characters and Levenshtein

distance less than 4. Out of the 161 publications that were identified as potential matches, 106 could be validated.

In the context of discovery of new resources from DBLP, we have found 667 distinct publications and distinct 845 distinct authors. These results are those which have not matched at sameAS criteria and may belong to Know-Center authors as publications and co-authors. Similarly in case of the Know-Center dataset we have found 363 distinct papers and 190 authors which were not matched as sameAS criteria and can be taken as additional resources for inclusion in DBLP bibliographic dataset. Hence this study indicates that lot of semantic data in addition to sameAs linking is discovered which can be used for dataset enrichment.

7 CONCLUSIONS AND FUTURE WORK

The Linked Open Data movement provides a heap of Linked Data and motivates people to open up their dataset for interlinking. This effort's main goal is to turn the Web into a single global data space and importantly give power to machines for understanding the data in order to make intelligent decisions. At present, the process of publishing data as Linked Data is rather easy, but interlinking data with external datasets already present in the Linked Data cloud is still quite a challenge. In this paper, we investigated the interlinking challenge. We proposed and implemented a multi-step strategy to interlink the Scientific publication dataset of Know-Center with external Linked Data resources and showcased an important case study for scholarly datasets interlinking. This interlinking will bring more enrichment to the repository of the Know-Center publication dataset and will provide a medium for the papers and authors for more visibility in the outside world of Linked Data. In the future, we will extend our semi-automatic approach to fully automatic for interlinking purpose and will also collect other informational aspect related to authors from the LOD cloud.

8 ACKNOWLEDGEMENT

This contribution is partly funded by the Know-Center, which is funded within the Austrian COMET program – Competence Centers for Excellent Technologies – under the auspices of the Austrian Federal Ministry of Transport, Innovation and Technology, the Austrian Federal Ministry of Economy, Family and Youth, and the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG.

References

1. Bizer, C., Heath, T., Ayers, D., Raimond, Y.: Interlinking Open Data on the Web. Demonstrations Track at the 4th European Semantic Web Conference, Innsbruck, Austria. (May 2007) `http://www.eswc2007.org/pdf/demo-pdf/LinkingOpenData.pdf`

2. W3C community project Linking Open Data. (2007) http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData
3. Berners-Lee, T.: Linked Data – Design Issues. (July 2006) http://www.w3.org/DesignIssues/LinkedData.html
4. Sauermann, L., Cyganiak, R., Ayers, D., Vlkel, M.: Cool URIs for the Semantic Web. W3C Interest Group Note. (2008) http://www.w3.org/TR/2008/NOTE-cooluris-20081203/
5. Resource Description Framework. (2004) http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/
6. Heath, T. and Bizer, C.: Linked Data: Evolving the Web into a Global Data Space (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, Morgan and Claypool 1:1. (2011) 1–136 http://linkeddatabook.com/editions/1.0/#htoc56
7. SPARQL Query Language for RDF, (2008). http://www.w3.org/TR/rdf-sparql-query/
8. Digital Bibliography and Library Project. (2009) http://www.informatik.uni-trier.de/~ley/db/
9. Glaser, H. Millard and I. C.: RKB explorer: Application and infrastructure. In: Proceedings of Semantic Web Challenge. (2007)
10. Semantic Web Dog Food. (2009) http://data.semanticweb.org/
11. Volz, J., Bizer, C., Gaedke, M., Kobilarov, G.: SilkA Link Discovery Framework for the Web of Data. In: Proceedings of CEUR-WS Vol-538 of 2nd Linked Data on the Web Workshop (LDOW2009), Madrid, Spain. (2009) http://events.linkeddata.org/ldow2009/papers/ldow2009_paper13.pdf
12. Maali, F., Cyganiak, R. and Peristeras, V.: Re-using Cool URIs: Entity Reconciliation Against LOD Hubs. In: Linked Data on the Web (LDOW2011). (2011) http://events.linkeddata.org/ldow2011/papers/ldow2011-paper11-maali.pdf
13. Google Refine Project. (2012) http://code.google.com/p/google-refine/
14. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: a collaboratively created graph database for structuring human knowledge, In: Proceedings of ACM SIGMOD international conference on Management of data. (2008) 1247-1250
15. Raimond, Y., Sutton, C. and Sandler, M.: Automatic Interlinking of Music Datasets on the Semantic Web. In: Linked Data on the Web (LDOW2008). (2008) http://events.linkeddata.org/ldow2008/papers/18-raimondsutton-automatic-interlinking.pdf
16. Latif, A., Afzal, M.T., Helic, D., Tochtermann, K., Maurer, H.: Discovery and Construction of Authors' Profile from Linked Data (A case study for Open Digital Journal). Linked Data on the Web (LDOW 2010). (2010)
17. Latif, A., Afzal, M.T., Ussaeed, A., Hoefler, P., Tochtermann, K.: Harvesting Pertinent Resources from Linked Data. In Journal of Digital Information Management (JDIM) 8 (3). (2010) 205–212
18. Oren, E., Delbru, R., Catasta, M., Cyganiak, R., Stenzhorn, H., Tummarello, G.: Sindice.com: A Document-oriented Lookup Index for Open Linked Data. International Journal of Metadata, Semantics and Ontologies. 3(1) (2008) 37-52
19. Cheng, G., Ge, W., Qu, Y.: Falcons: Searching and Browsing Entities on the Semantic Web. In: Proceedings of 17th International World Wide Web Conference, Beijing, China. (2008) 21–25
20. Ding, L., Finin, T., Joshi, A., Pan, R., S. Cost, R., Peng, Y., Reddivari, P., C. Doshi, V., Sachs, J.: Swoogle: A Search and Metadata Engine for the Semantic Web. In: Proc. Thirteenth ACM Conference on Information and Knowledge Management, Washington, D.C., USA. (2004) 8–13

21. Michael, L.: DBLP - Some Lessons Learned. PVLDB 2(2). (2009) 1493–1500