Article

Webbing Semantified Scholarly Communication Datasets for Improved Resource Discovery

Journal of Digital Information Management (JDIM)

Suggested Citation: Latif, Atif; Tochtermann, Klaus (2012) : Webbing Semantified Scholarly Communication Datasets for Improved Resource Discovery, Journal of Digital Information Management (JDIM), ISSN 0972-7272, Digital Information Research Foundation, Chennai, Vol. 10, Iss. 4, pp. 245-253

This version is available at:
http://hdl.handle.net/11108/99

Webbing Semantified Scholarly Communication Datasets for Improved Resource Discovery

Atif Latif[1,2], Klaus Tochtermann[1,3]
[1]ZBW - German National Library of Economics
Leibniz Information Center for Economics,
Kiel, Germany.
[2]a.latif@zbw.eu
[3]k.tochtermann@zbw.eu

ABSTRACT: *The success of Linked Data project has played a vital role in the realization of the Semantic Web on a global stage. It has motivated people to publish datasets which are important for information linking, resource discovery and can further make contributions in shaping 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 and further on presentation of post-interlinking data are still researchable challenges. The search and identification of relevant datasets as well as devising a strategy for linking to these datasets with post-organization of interlinked data for human perceivable presentation is still a difficult task. In this paper, a novel approach is presented which 1) highlights and implements the steps involved in the interlinking process and 2) proposed a proof of concept application for automatic generation of an organized profile for post interlinked data presentation. These approaches are applied and presented as a case study focusing on interlinking scholarly communication datasets and highlighting the potential benefits of Linked Data. This study has interlinked two semantified scholarly datasets and was successfully able to discover new resources which were further organized into profile for showcasing potentials of Linked Data.*

Categories and Subject Descriptors:
I.2.8 (Problem Solving, Control Methods, and Search); *Heuristic methods*

General Terms: Linked Data, Web of Data, Semantic Web

Keywords: Linked Open Data (LOD), Semantic Web, information retrieval, information presentation, CAF-SIAL, URI, User Interfaces

Received:

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 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 to bootstrap the Semantic Web. 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 deference-able 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]:

- It removes data silos, turning the Web into a connected Giant Global Graph.
- All concepts are modeled with single Resource Description Framework (RDF), bringing in consistency to structured data representation leading to interoperability at various levels.
- Every concept has a unique identity (URI) in the document which is further discoverable, reusable and linkable.
- 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. Collectively, 295 data sets consisting of over 31 billion RDF triples which are interlinked by around 504 million RDF links, are recorded in September 2011.The heavy presence of Linked Data in various domains (e.g. government, geo-information, and scholarly communication)[1] 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: to structure the data (using RDF), put it on the Web, and include semantic links to other dataset. However the first two steps are rather easy to accomplish and are already covered by many case studies but the task of finding relevant external data to link to is still a challenge and lacks case studies. Another open challenge after interlinking step is to present information in organized way which can highlight the added values of Linked Data. The challenges usually involved in the interlinking and post-interlinked data presentation processes 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

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

- Adequate knowledge of SPARQL querying and semantic technologies as well as expertise in using SPARQL endpoints or dumped RDF datasets
- Disambiguation and identity resolution
- Presentation of post-linked information with simplified interface

In this paper we focus on said challenges and investigate the interlinking and post-interlinked data presentation possibilities in the Linked Open Data 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 use of available scientific scholarly datasets, i.e. DBLP [8], RKB Explorer [9], Semantic Web Dog Food [10] and DBpedia [21]. After the interlinking process, post-interlinked data is arranged and presented in an automatic generated profile by a prototype proof of concept application.

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 with pseudo algorithm are described in detail, followed by the prototype application generated profile to show results of the automatic interlinking. The paper closes with a discussion of the approach and an outlook on future research.

2. Related Work

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. In another study Latif et al [14] pointed towards importance of organized and perceivable presentations of Linked Data.

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

Based on the key-based as well as; similarity-based approaches many state of the art interlinking tools and case studies surfaced over Linked Data sphere [20]. These tools targets interlinking process usually in different context e.g. structured or un-structured datasets, similarity matrices, output and human effort. Some of the important interlinking tools, techniques and case studies for interlinking and post-interlinking presentation are discussed below.

2.1.1 RDF-AI:

RDF-AI [19] is an interlinking tool which provides modules for the pre-processing, matching and data fusions on RDF datasets. It delivers various outputs depending upon user provided output module specification. It uses string matching and taxonomical similarity measures for interlinking and connects two datasets on the owl:sameAs relationship. This tool is good for already RDFized dataset but for unstructured dataset, users have to take care of the conversion into RDF and parameter specification.

2.1.2 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 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.

2.1.3 Linking Music Data:

Yves Raimond et al. [12] looked at the problem of interlinking music-related data sets on the Web. They described a graph based matching interlinking algorithm that takes into account both the similarities of web resources and of their neighbors. It has produced very good results with interlinking of music data with other related datasets.

2.1.4 Linking Open Journal Data:

In another case study, Latif et al. [13] worked on the interlinking of open digital journal data with the LOD cloud by extending the CAF-SIAL application[2]. CAF-SIAL is a proof of concept system to discover and present informational aspect of resources describing people from Linked Data [14]. 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[3] 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.

2.2.1 SINDICE

Sindice [15] 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 [16] or Swoogle [17]. Sindice is used in this study, mainly due to its larger indexing pool and the ease provided in use of public API

2.2.2 SameAs

SameAs[4] 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

[2] http://cafsial.lod-mania.com/

[3] http://www.jucs.org/

[4] http://sameas.org/

the exact URI beforehand. The SameAs API returns result in multiple formats.

Table 1. Dataset Selection Results

Repository Name	Michael Granitzer	Stefanie Lindstaedt	Klaus Tochtermann	Total	Comments
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

3. KNOW-CENTER Test Dataset

For this experiment, our main test dataset is based on a Know-Center[5] publication dataset provided openly as RDF. The Know-Center is Austria's competence center for knowledge management and knowledge technologies, founded in 2001. 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 319 authors and 538 distinct 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[6] 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. Details about these dataset are provided in Table 1 above.

4.1 DBLP D2R L3S and FU BERLIN Server

The DBLP D2R L3S server[7] 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 [18]. 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[8]. 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[9], ACM[10], and DBLP[11]. All these datasets are published as Linked Data by the RKB [9]. RKB explorer is a service which, after applying 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 DBpedia

For our post-interlinking author profile generation, we selected DBpedia dataset for locating the author bio-information. DBpedia is a famous project, which extracts structured information (tables) from Wikipedia and publishes this created information as Linked Data on the World Wide Web. What makes DBpedia as one of the central interlinking and searching hub for person data information is its large datasets which contain about 416,000 persons (as of July 2011). These persons are further linked with persons, places and images…. within DBpedia and outside with linked datasets like FreeBase, DBLP and GeoNames etc. To locate and query Know-Center author information we accessed SPARQL Endpoint of DBpedia[12].

4.5 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

[5] http://know-center.tugraz.at/
[6] http://ckan.net/
[7] http://dblp.l3s.de/
[8] http://www4.wiwiss.fu-berlin.de/dblp/
[9] http://citeseer.rkbexplorer.com/
[10] http://acm.rkbexplorer.com/
[11] http://dblp.rkbexplorer.com/
[12] http://dbpedia.org/snorql/

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 are illustrated in Table 1 at above page.

5. Interlinking Framework

For the interlinking of the Know-Center 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 and rdf:seeAlso relationship. The framework for this strategy is illustrated in Figure 1. The strategy looked as follows:

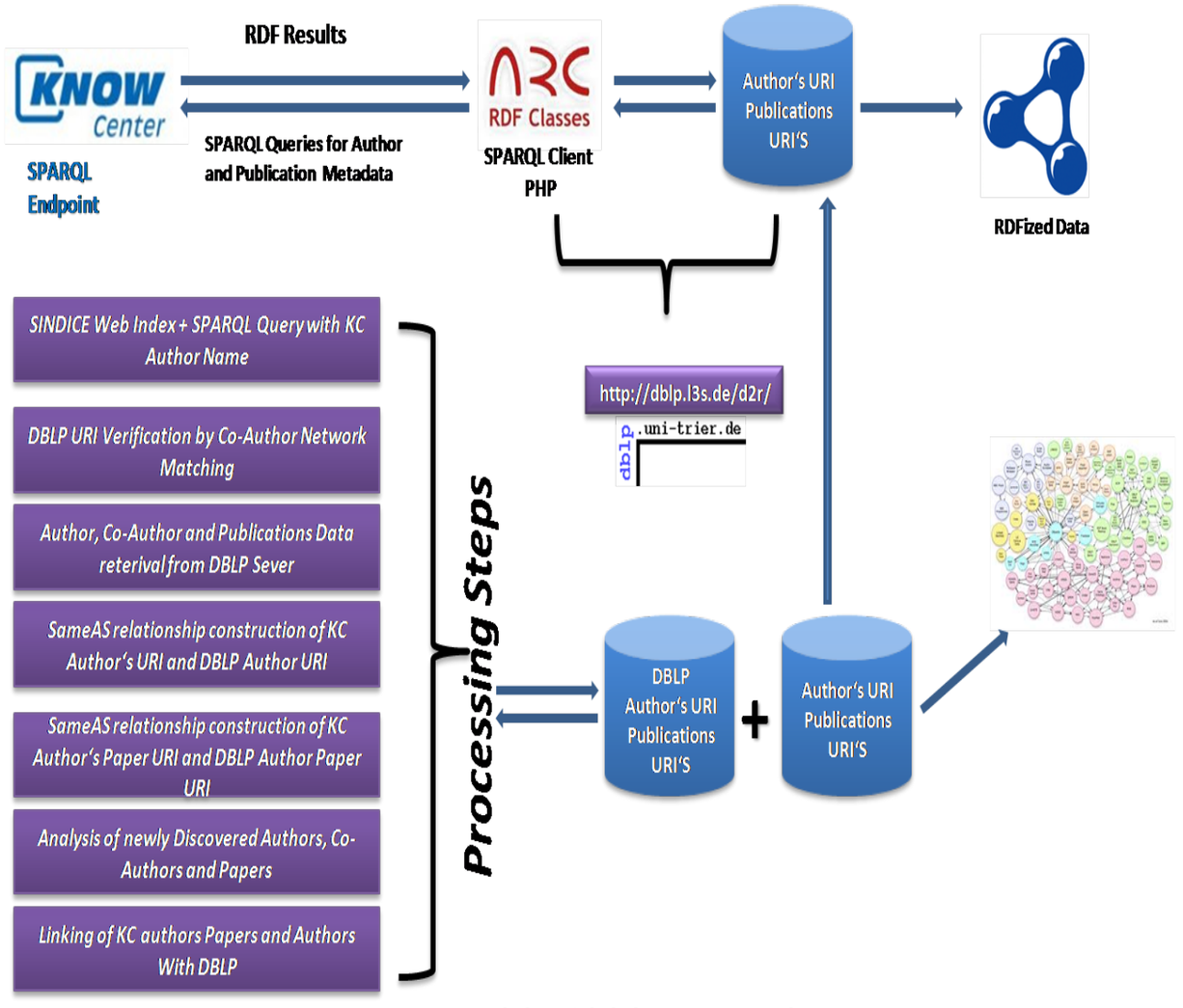

Figure. 1 Interlinking Scholarly Data Framework

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[13] 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 for Auhtor DBLP records

The searching for authors URI from DBLP was divided into two parts 1) Sindice Search API call and 2) DBLP SparqlEnpoint query. We decided to perform these two steps for improving our result set. The details of these processing steps are given below.

First, the Sindice Search API[14] 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 was then filtered automatically on the basis of heuristics to make sure that they belonged to the DBLP. In this process, URIs for 112 out of the 319 authors in question was found. Keeping in mind the limitations of Sindice -- basically the same recall and precision problems that every search engine displays to a certain degree – additionally, have to query DBLP Sparql endpoint as a second step.

[13] http://know-center.tugraz.at/sparql/

[14] http://sindice.com/developers/searchapi

In this phase, we first iteratively construct a sparql query with the names of an author in question and then on returned results employed a string-matching algorithm that compared the names of the authors in question with author names from the DBLP data set. In this process, URIs for 120 out of the 319 authors in question was found. This way we found a few additional URIs which had not been retrieved by Sindice. After combining results of these two steps, in total 133 authors URIs from DBLP was located.
We combined the results from steps 1 and 2 and constructed a relational database which stored the names and URIs of each author for future processing

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. Furthermore, 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, rdf:seeAlso relationships between authors as well as publications were established and published.
One of the limitation which we spotted in our validation service was it's inability to validate the authors who have papers with associated co-authors which are not indexed in DBLP. For example "Soren Auer" has only one publication associated with three co-authors in Know-Center dataset. Interestingly, the DBLP Uri of author was found in DBLP dataset in step 2 which lead us to 58 distinct papers with associated 53 co-authors. Apparently this record gives us a healthy resource discovery in form of publications and co-authors but in validation service this record cannot be validated because the resources present in Know-Center and DBLP produced no matches. We found this limitation is purely dependent on the nature of the data at our test Know-Center dataset and vice versa.
In summary our validation service will ensure the maximum validity by dropping the records who shared same names as of authors but not a publisher as well as; those authors who have URIs in Know-Center and DBLP but not have resources which produce match between two datasets. The matching of at least one co-author or publication title is essential for author validation.

6. Results

In this section the results achieved by the proposed multi-step strategy are discussed. At the start, a local data set (Know-Center) with 319 authors and 538 publications was semi-automatically matched with a remote data set based on data from the DBLP, provided by the L3S. Using the Sindice Search Engine, 112 DBLP URIs were found where author names had a Levenshtein distance of less than 4.
By querying the DBLP SPARQL endpoint using the author's names, 120 URIs were found. After combining both the results of Sindice and the direct SPARQL query, the total number of matched URIs was 133. Out of the 133 identified matches, 79 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 1 of them had to match. If 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 538 local publications, 172 distinct paper titles 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. It is important to note here that distinct paper title is taken as one record contrary to the number of author attached to it. 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 172 publications that were identified as potential matches, 123 could be validated.
In the context of discovery of new resources from DBLP, we have found 854 distinct publications and distinct 758 authors. These results are those records which have not matched at sameAS criteria and belongs to Know-Center authors as rdf:seeAlso additional publications and co-authors. Similarly in case of the Know-Center dataset we have found 366 distinct papers and 240 authors which were not matched as sameAS criteria and can be taken as additional resources for inclusion in DBLP bibliographic dataset.
Hence the resource discovery statistics has shown success of this study and has proved that lot of scholarly linked data with addition to sameAs linking is available for building various useful case studies. Furthermore, this additional data can be used for enhancing records and to bring in added value with discovered new publications. Importantly, all these additional discovered recourses have deference-able unique URIs which will lead toward more resource discovery about co-authors and publications

7. Presentation of Post-Interlinked Data

After the interlinking step, it was also of great importance to showcase the results in an organized way to project the added benefits of the discovered results. We decided to build a prototype application. The goal of this application was to automatically generate author profile which will display interlinking results along with biography information of an author from DBpedia (if presents) in an organized way. We are of view that this proof of concept application will help users to understand the resource discovery and interlinking added value in a better way.

7.1 Pseudo Algorithm

Algorithm Author_Interlinked_Profile (Author)

Start

1. Create '*empty author profile*'
2. **for each** *author* **do**
3. **get** *KCAuthorUri&Publications (KC SparqlEndpoint query)*
4. **remove** '*anomalies in author name*'
5. **get** *AuthordblpUri&Publications(kcauthorName)*
6. **validate** *Author(kcAuthor,kcCo-author, dblpAuthor,dblpCo-author)*
7. **if** (*validated*)
8. **generate** *sameAslinking(kcAuthordetail, dblpAuthordetail, kcPublication,dblpPublication)*
9. **link** *AdditionalResources (kcdata,dblpdata)*
10. **get** *authorprofile (sameAsPublication, sameAsAuthor, seeAlsoPublication, seeAlsoCo-author)*
11. **else**
12. **return** '*emptyProfile*'
13. **end**
14. **end**
15. **return** '*Author_Interlinked_Profile*'

Function KCAuthorUri&Publications(sparql query)

Start

1. **If** '*authorDetail*' exists in '*Know-Center Sparql endpoint*'
2. **query** and **process** '*Author, Co-Author, Publications*' detail
3. **return** *kcAuthorandPublicationsDetail*
4. **else**
5. **return** *null*
6. **end**

End

Function AuthordblpUri&Publications(kcauthorName)

Start

1. **If** '*authorDetail*' exists in '*DBLP Sparql endpoint*'
2. **return** *kcAuthorandPublicationsDetail*
3. **else if** '*authorDetail*' exists in '*SINDICE*'
4. **query** and **process** '*DBLP endpoint for kcAuthordblpUri*'
5. **return** *kcAuthorandPublicationsDetail*
6. **else**
7. **return** *null*
8. **end**

End

Function validateAuthor (kcAuthor,kcCoauthor, dblpAuthor, dblpCoauthor)

Start

For each *author* **do**

1. **If** (*kcAuhtorDetails AND dblpAuthorDetails*)
2. **match** '*at-least two co-author and one publication title in KC and DBLP*'
3. **return** *validatedAuthor*
4. **else**
5. **return** *noValidation*
6. **end**
7. **end**

End

Function Sameaslinking(kcauthor, dblpauthor, kcpublication, dblpPublication)

Start

1. **Levenshtein distance matching** of '*Author and Co-Author network (KC and DBLP)*'
2. **Levenshtein distance matching** of '*Publications Titles (KC and DBLP)*'
3. **create** rdf:sameAs linking of the '*matched entities URI (KC and DBLP)*'
4. **return** *sameAsAuthorandPublications*

End

Function linkAdditionalResources (kcdata,dblpdata)

1. **If** ('*sameAsAuthorandPublication*')
2. **locate** *additional Co-authors and publications*
3. **create** *rdf:seeAlso Co-authors links*
4. **create** *rdf:seeAlso publications links*
5. **return** *additional linked resources*
6. **else**
7. **return** *null*
8. **end**

End

Function authorprofile (sameAsPublication,sameAsAuthor, seeAlsoPublication,seeAlsoCoauthor)

Start

1. **If** '*author validated*'
2. **If** '*AuthorUri*' exists in '*DBpedia Sparql Endpoint*'
3. **read** *abstract from DBpedia*
4. **construct** '*brief Introduction Tab*'
5. **else**
6. *do-nothing*
7. **end**
8. **get** *rdf:sameAs publications and Co-Author*
9. **construct** '*rdf:sameAs publications and Co-Authors Tab in KC and DBLP*'
10. **get** *rdf:seeAlso publications and Co-Author*
11. **construct** '*rdf:seeAlso publications and Co-Authors Tab in KC and DBLP*'
12. **generate** '*interlinked RDF file for download*'
13. **return** *author_interlinked_Profile*
14. **end**

End

Listing. 1 Pseudo Algorithm for Automatic Author Profile Generation

For the presentation of interlinked data it was important to devise an intelligent algorithm which can automatically query, process, and display the results. Building on that, we have written a modular algorithm comprises of six strictly interdependent procedures. This algorithm starts with issuing a SPARQL query to *Know-Center Sparql Endpoint* for author information and ends with construction of author profile by utilizing each corresponding module output. The pseudo code of an algorithm is given in above Listing1 for better understanding.

7.2. Author Profile Illustration

In this section, the description of an author profile generated by our proof of concept application is discussed with the help of Figure 2. Initially, list of entire Know-Center authors who has been validated with owl:sameAs DBLP links are presented to the users. When user clicks on one of the listed authors, backend algorithm functions to generate a profile. For description, we selected “Klaus Tochtermann” author and major contributor in Know-Center datasets as an example. The algorithm first locate (if present) the equivalent DBpedia URI of the person from DBpedia datasets to acquire the brief introduction of the person which is then embedded to the “Brief Introduction” section of the profile as shown in Figure 2.

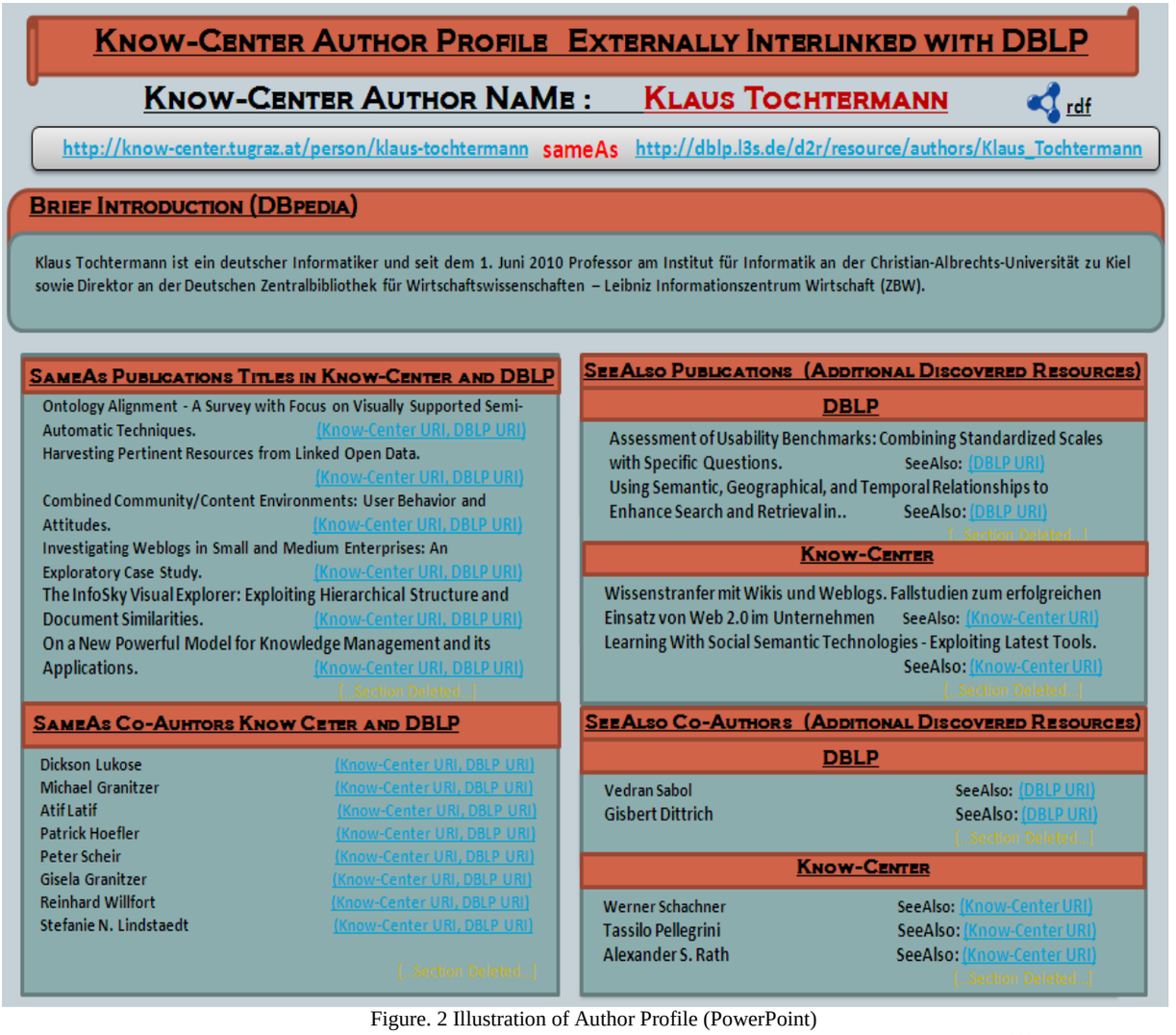

Figure. 2 Illustration of Author Profile (PowerPoint)

Next, a section with "SameAs Publications Titles in Know-Center and DBLP" is generated where all the matched publications and Co-Authors in Know-Center DBLP are displayed. In addition, for further de-referencing and lookup, the URI of each matched resource in Know-Center and DBLP is provided. Next, in the last section “SeeAlso Publications (Additional Discovered Resources)” discovered publications and co-authors of current authors from Know-Center and DBLP are listed by the algorithm automatically. Importantly, links to publications and Co-Authors in Know-Center and DBLP are also provided for more resource discovery. For example, user wants to see publication of current author co-author. Clicking on provided DBLP link will take users to desired resources.

3.2 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 along with presentation of interlinking results are still quite a challenge. In this paper, we investigated the interlinking and post-interlinking presentation challenges. We proposed and implemented a multi-step strategy to 1) interlink the scientific publication dataset of Know-Center with external Linked Data resources and 2) present the post-interlinked results. Our strategy showcased an important case study for scholarly communication datasets interlinking and highlighted the potentials in efficient resource discovery by help of our study results. This interlinking has brought more enrichment to the repository of the Know-Center publication dataset and has provided a medium for the papers and authors to get more visibility in the outside world of Linked Data. In addition, we have introduced a new post-interlinking data presentation strategy with our automatic generated "Author profile". In the future, we will collect other information related to authors from the LOD cloud. We also have a plan to improve and generalize our prototype proof of concept application to make it available for other scholarly communication datasets interlinking case studies.

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., Völkel, 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.: Silk–A 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] 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

[13] 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)

[14] 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

[15] 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

[16] 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

[17] 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

[18] Michael, L.: DBLP - Some Lessons Learned. PVLDB 2(2). (2009) 1493—1500

Authors' Biographies

[19] Scharffe, F., Liu, Y., Zhou, C.: RDF-AI: an Architecture for RDF Datasets Matching, Fusion and Interlink. In Proceedings of the IJCAI 2009 workshop on Identity, reference, and knowledge representation (IR-KR), Pasadena, CA US, 2009.

[20] Wölger, S., Siorpaes, K., Bürger, T., Simperl, E., Thaler, S., Hofer, C.: A Survery on Data Interlinking Methods. STI Technichal Report, March 2011.

[21] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z (2007). DBpedia: A Nucleus for a Web of Open Data. In: Proceedings of the 6th International Semantic Web Conference (ISWC). Springer, Busan, Korea .

Dr. Atif Latif has earned his PhD degree in Computer Science with distinction from Technology University of Graz, Austria in 2011. He was affiliated with Institute of Knowledge Management and Know-Center, Austria's COMET Competence Center for Knowledge Management. He received his master's degree in Computer Science from Quaid-i-Azam University, Islamabad, Pakistan in 2007 and was awarded with excellent grade in his final thesis. His main research areas are Linked Open Data, Semantic Web, Social Semantic Web and Digital Libraries. Dr. Atif Latif is the author of book and has published several articles appeared in reputable journal and conferences. He is involved in scientific services (Editor/Committee-member/Session-chair/) for number of international conferences and Journals.

Prof. Dr. Klaus Tochtermann is a professor for Computer Media at the Christian-Albrechts University of Kiel (Germany). In addition to his professorial duties, he is also the Director of the ZBW – Leibniz Information Centre for Economics. With about 270 employees maintaining more than 4,2 Million documents related to economics, ZBW is the world's largest library for economics. From October 2000 to Jun 2010, Prof. Tochtermann was the director of Austria's first industry-based research institute on knowledge management Know-Center. The work of the Know-Center includes projects on knowledge management, knowledge relationship discovery, semantic technologies, workplace-integrated learning, Web2.0. Klaus Tochtermann had also been head of the institute for knowledge management at Graz University of Technology from 2004 to 2010. Klaus Tochtermann studied Computer Science and earned his Dr. degree in 1995. In 1996 he spent his post-doc at the A&M University in Texas.