Enriching scientific publications with semantically related data

Arben Hajra
Bul. Ilindenska 335
1200 Tetovo
FYR of Macedonia
+38944356187
a.hajra@seeu.edu.mk

Klaus Tochtermann
Düsternbrooker Weg 120
24105 Kiel
Germany
+49-431-8814-333
k.tochtermann@zbw.eu

Vladimir Radevski
Bul. Ilindenska 335
1200 Tetovo
FYR of Macedonia
+38944356000
v.radevski@seeu.edu.mk

ABSTRACT
Linked Open Data (LOD) and semantic technologies present an enormous potential for a variety of applications. LOD cloud is changing shape in a permanent way with the continuous addition of new datasets. This raises new possibilities for usage and affects the creation of applications on top of these data for various purposes.

The purpose of this paper is to describe the use of LOD and semantic technologies to allow content and search results enrichment for digital libraries, with an emphasis to EconBiz case study. The main challenge for achieving this, present the process of contextualization of scientific publications with semantically related data. More specifically we describe the process of creating a profile of publication from the content that is available in a semantic web representation. This would allow detection and alignment of similar publications from different areas, author details, co-authors relativeness, events, organizations, etc.

This is claimed to affect significantly the perspective of what should be of digital libraries in the future.

Categories and Subject Descriptors
H.3.4 [**Semantic Web**], H.3.7 [**Digital Libraries**]

Keywords
Linked open data, LOD, Digital libraries, Semantic web, Publications, Scientific Papers

1. INTRODUCTION
The focus of this paper is to highlight the usage of semantic data, i.e. Linked Open Data (LOD) in order to enrich scientific publications with other relevant information. The publication of a large number of data, such as linked data, provides an excellent opportunity to use them in different scenarios. We describe the possibility of using these data for creation of a specific profile with other information for a given scientific publication. The idea is that scientific publications, mainly from the field of economics, can be enriched with additional data for making a wider, still relevant overall "picture" of them. One of the main desirable outcomes in this context is to find and obtain scientific publications from different fields, stored in other repositories (different from the repository of the initial search) and retrieved as recommended literature for that profile. The ultimate, "enriched" profile would include information about the author, the aspects of his scientific work (wider area, scientific community etc.) correlations with other authors (colleagues and co-authors), information about conferences, events, projects etc.

This research was realized at ZBW, German National Library of Economics - Leibniz Information Centre for Economics[1]. This institution represents the world's largest specialist library for economics. The library hosts more than four million publications in printed or electronic format and subscriptions to 31,970 periodicals and journals.[2] The ZBW offers EconBiz[3] as a single point of access to the world's economics literature and information and the database ECONIS[4] with more than five million datasets. Through EconStor[5], the ZBW offers a platform for Open Access publishing to German researchers in economics. It is the leading Information Centre for developing and applying the latest semantic technologies and Web 2.0 technologies for highly innovative information services. In 2012, ZBW in collaboration with other institutes of the Leibniz Association and several university institutes from all over Germany has established a strategic research network known as Science 2.0.[6]

Nowadays, the German National Library of Economics (ZBW) maintain the Standard Thesaurus Wirtschaft (STW)[7], the Thesaurus for Economics and uses it for indexing purposes [1]. Besides this thesaurus, for the purposes of our research at this phase, we will consider some other thesauruses and repositories,

[1] http://www.zbw.eu/

[2] http://zbw.eu/e_about_us/e_library_profile.htm

[3] http://www.econbiz.de/en/

[4] http://www.zbw.eu/e_catalogues/e_econis.htm

[5] http://www.econstor.eu/

[6] http://science20.zbw.eu

[7] http://zbw.eu/stw/versions/latest/about

such as, Social Sciences Thesaurus (TheSoz)[8] and Food and Agriculture Organizations Thesaurus (Agrovoc)[9].[2]

2. PROBLEM STATEMENT AND PROPOSAL

Digital Libraries (DL) are the crucial place for scholarly communication (publishing, discovering and sharing scientific findings). However we do not always get the most relevant, up to date and cross domain literature as a result of our search, because the main reason of such limitation is the static metadata structure of DLs. [3] Thus, we get limitations in the literature search to a specific field, correspondingly catalogued and indexed. [4], [5]. Christine Borgman [6] since 2003, noted that DLs are like monolithic systems, where metadata describe the data rather than usages.

The relevance and the usage of a scientific publication may depend on many elements which make its determination a complex task, while the process of collection, classification and evaluation is typically time consuming. According to this, having a wider "profile" or a "picture" of a publication with several details can help the process in general. The "picture" should include "everything" that exists around about the publication; other publications from other disciplines, authors' details, co-authors relations, information about institute or organization, events, etc. For example this institute has published these papers in that field, they are produced by this working group and similar.

In the Fig.1 we represent an overview of the contextualization process of a scientific paper. Thus, for each publication from EconBiz, a detailed picture would be offered. This picture will represent the profile of the publication.

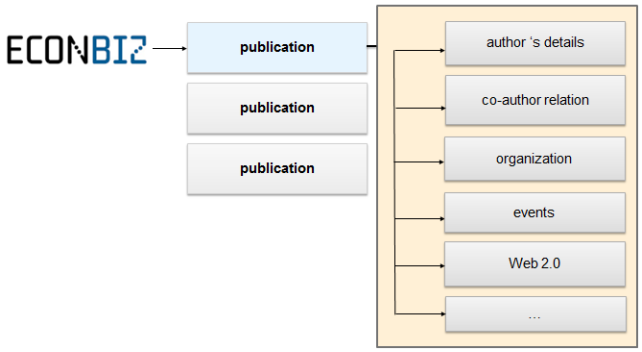

Figure 1. Possible elements of publication profile

The primary source for enriching the profile of a item subject to search will be the LOD cloud. The main focus of our research are repositories and thesauruses like STW, Agrvoc, OpenAgris[10], TheSoz, DBLP[11], etc. They are considered as most promising sources for the elements for enriching the profiles. To further improve the retrieved results, we can link up to other "nonscientific" repositories, from web 2.0.

A challenge at this point will be to answer the following questions: What can be an ideal profile for making the full "picture" of a publication? How many information would be available in order to create this profile considering the fact that not much content is available in a Semantic Web representation?

3. RESEARCH TRACK

In order to achieve the required results we will follow different paths which converge into the main goal. Staring with LOD cloud, by discovering and using the aligned concepts between datasets and then continuing to find other existing ontologies. Further, the Web 2.0 resources can be considered.

3.1 Using Linked Data Cloud (LOD)

We will consider Linked Open Data by exploring existing alignments between EconStor and other datasets within the current Linked Open Data (LOD) cloud. The STW Thesaurus for Economics is the semantic representation of EconStor and is part of LOD cloud[12]. It provides a vocabulary on any economic subject: more than 6,000 standardized subject headings and about 19,000 entry terms to support individual keywords. The alignments between STW thesaurus and the concepts in other repositories is one of the key factors for retrieving resources from these repositories. Currently STW has a large number of aligned concepts with different datasets; it is mapped to DBpedia[13], SWD[14], TheSoz and Agrovoc.

Between Agrovoc and STW there are 1136 linked concepts with *skos:exaxtMatch* type of links.[15] While between STW and TheSoz 4927 links (2844 exact matches, 631 related matches, 1418 broad matches, 34 narrow matches).[7]

Let's consider the following example: The concept *RISK* in Econstor (http://zbw.eu/stw/descriptor/ 10057-0) is linked to *Risiko* in TheSoz (http://lod.gesis.org/ thesoz/concept/10045555) and *Risk* in Agrovoc (http://aims.fao. org/aos/ agrovoc/c_6612).

The Thesaurus for the Social Sciences (TheSoz) is a SKOS-based[16] German thesaurus for the domain of the social sciences. The TheSoz is available in three languages (German, English and French) and contains overall about 12,000 keywords, from which 8,000 are descriptor and 4,000 non-descriptors (terms). [7]

Agrovoc is a comprehensive multilingual agriculture thesaurus, which is used for indexing the data in AGRIS[17]. Agrovoc contains close to 40,000 concepts in over 22 languages covering agriculture, forestry and fisheries, food security and other domains [2],[8].

In the described real-world context, we can enlarge the retrieving space by consideration of retrieve resources from other

[8] http://www.gesis.org/en/services/research/thesauri-und-klassifikationen/social-science-thesaurus/

[9] http://aims.fao.org/standards/agrovoc/linked-open-data

[10] http://aims.fao.org/openagris

[11] http://www.informatik.uni-trier.de/~ley/db/

[12] http://lod-cloud.net/

[13] http://dbpedia.org/About

[14] https://wiki.d-nb.de/display/LDS

[15] http://aims.fao.org/standards/agrovoc/linked-open-data

[16] http://www.w3.org/2004/02/skos/

[17] http://viaf.org

repositories and on the basis of an initial search starting point. If we begin to search with a keyword in local repository, e.g. EconStor, there will be a set of concepts, terms and metadata which match that keyword. By using *skos:Narrower* or *skos:Broader*, we can narrow or broaden the concepts. Then, we continue to perform a SPARQL query with "selected" concepts to STW thesaurus and by considering the alignments of these concepts to other datasets. The output of this process would be a list of publications from other domains that meet our initial search concepts.

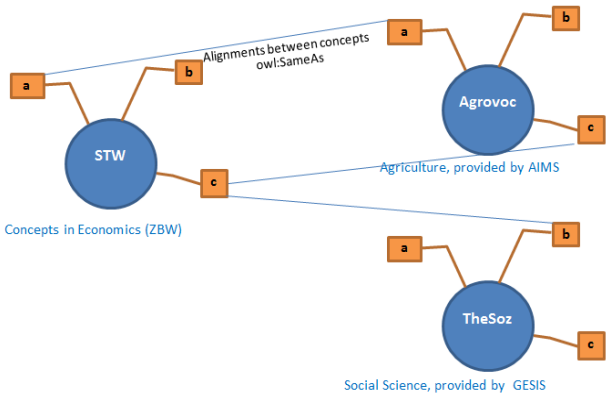

Figure 2. Aligned concepts between datasets.

The quality of returned resources may depend on many factors; however most of them are related to the alignments. The key elements here are:

- Which of the returned elements, terms or concepts should be considered?
- How many arrows (alignments) to be taken between repositories for retrieving a publication (semantic relativeness).
- Do we need an inference, and up to which extent can we use it, to find new concepts?

3.2 Using existing ontology to link up with linked data

By using the alignments of the concepts between the repositories we are limited in a specific domain. To expand the domain, it will be necessary, besides the aligned concepts (*owl:sameAs*), to also find an existing ontology (FOAF[18], VIAF[19], etc.). In this way we can achieve information from various domains and repositories. For instance retrieving person's profile, contribution, organization and related events. Further on we can discover and analyze the correlations between *foaf* persons (authors), and similar.

Let us consider a situation when we know the author (and co-author(s) if any).

In this case we can use FOF profile (from files of social network) to identify with whom is this author related.

By executing this listing into the BDLP SPARQL Endpoint[20], a list of co-authors will be showed from DBLP.

```
SELECT DISTINCT ?name ?other   WHERE {
?a   foaf:name "Arben Hajra".
?pub dc:creator ?a;
     dc:title ?title;
     dc:creator ?other.
?other foaf:name ?name.
}
```

Now we can search for literature of people who are related to that particular author. In this manner we explore new semantic relativeness.

3.3 Using Web 2.0 Recourses

The existing interoperability within scientific resources offered as Linked Open Data can be a good track for reaching the goal of enriching the search results. In a case where the results which are obtained from interoperability with LOD do not satisfy the needs, we can refer to Web 2.0 resources as for example the scholarly digitally maintained libraries. Nowadays there are several socially maintained scholarly libraries, based on the folksonomy approach, similar to a social resource sharing systems. It means that a resource at Web 2.0 can be tagged, commented, ranked thus in an "informal" way it gets a kind of semantic [9], [10], [11]. A specific publication from these repositories then can be judged by popularity (e.g. most visited, most downloaded), comments, ranking, etc. The challenge here is to discover scholarly resources from socially maintained library services, such as, Bibsonomy[21], Mendeley[22] , CiteULike[23] , etc. Thus, the shared views on relevant literature and other socially influenced approaches can enrich the scholars' discovery experience.

3.4 Result Evaluation

Taking into account the fact that the data will be read and "understood" by the "machine", the feedback of humans at this point is very crucial. Initial evaluations of the prototype will be made by a small group of people who are aware with the expectations of the results. We will use a platform for getting the feedback electronically in the phase of evaluation. A system that will include the feedback automatically and improve the inferences is anticipated.

4. PROTOTYPE

For each publication stored in EconStore we extract the metadata, a set of concepts - as authors, co-authors, titles, keywords, terms, etc. From the offered set, we offer the possibility of choice of concepts for performing SPARQL queries in other repositories. In section 3.1 some challenges about this issue are described, as the nature and the number of the chosen concepts, the alignments that these concepts has to other datasets and outgoing links.

For example, the publication "*Employment protection, technology choice, and worker allocation*" from Eric Bartelsman is retrieved

[18] http://www.foaf-project.org/

[19] http://viaf.org/

[20] http://dblp.l3s.de/d2r/snorql/

[21] http://www.bibsonomy.org/

[22] http://www.mendeley.com/

[23] http://www.citeulike.org/

from EconStor. There are several metadata extracted, as the "keywords", which are shown in Table 1.

Table 1. Metadata of a publication

Keyword	Link	Description
dc:subject	<http://zbw.eu/stw/descriptor/10040-3>	Allocation
dc:subject	<http://zbw.eu/stw/descriptor/10057-0>	**Risk**
dc:subject	<http://zbw.eu/stw/descriptor/10460-2>	Productivity
dc:subject	<http://zbw.eu/stw/descriptor/10464-1>	High technology
dc:subject	<http://zbw.eu/stw/descriptor/10473-0>	Choice of technology
dc:subject	<http://zbw.eu/stw/descriptor/16124-2>	Employment protection
dc:subject	<http://zbw.eu/stw/descriptor/16757-5>	Information technology
dc:subject	<http://zbw.eu/stw/descriptor/17829-1>	United States
dc:subject	<http://zbw.eu/stw/descriptor/17983-5>	EU countries
dc:subject	<http://zbw.eu/stw/descriptor/19044-6>	Comparison

Each of these concepts is presented at STW Thesaurus with additional descriptions; the term in German language, the narrowed or broader terms, related terms and the alignments that this concept has in other datasets. If we take the keyword "Risiko" from the Table 1, all of these details can be showed in Fig.3. At the end are shown the alignments to other Thesauri and Vocabularies. As we can note, this concept is aligned to TheSoz at this link http://lod.gesis.org/thesoz/concept/10045555 or at Agrovoc with http://aims.fao.org/aos/agrovoc/**c_6612**.

For delivering more details, if we perform a query in the OpenAgris SPARQL endpoint[24] with the above concept, a list of publications will be showed.

```
SELECT DISTINCT ?title
 {
a dcterms:type "Article";
  dcterms:title ?title;
  dcterms:subject ?s.
 FILTER (regex(str(?s), "c_6612" , "i"))
} ORDER BY ?s
```

Table 2. List of articles retrieved from OpenArgis

Title
"Stewardship and Risk: An Empirically Grounded Theory of Organic Fish Farming in Scotland"
"Private Decisions and Public Goods: Trade-Offs in the Conservation Programs in the New Farm Bill: Discussion"
"Subsurface Drip Irrigation Versus Center-Pivot Sprinkler for Applying Swine Effluent to Corn"
"Rainfall reliability, drought and flood vulnerability in Botswana"
"The Economic Rationale of Recycling Hybrid Seeds in Northern Tanzania"

[24] http://202.45.142.113:10035/repositories/agris

Using different concepts we get different results. Thus, the combination and the number of the concepts determine the quality of the retrieved results.

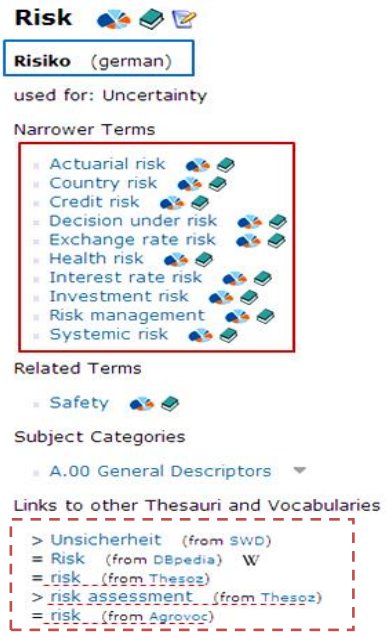

Figure 3. Description of a concept in STW Thesaurus.

5. SUMMARY

In this paper we described the use of LOD and semantic technologies to make content enrichment of digital libraries - by considering the EconBiz repository as a case study. Adding a list of recommendations is provided by enabling linking data from different areas and disciplines. In such way we obtain contextualization of scientific publications and their enrichment with semantically related data. This allows us to create a profile of a publication from the content that is available in a semantic web representation. The work is considered to be a contribution in the perspective of what digital libraries should look like in the future.

6. ACKNOWLEDGMENTS

Our thanks go to German National Library of Economics - Leibniz Information Centre for Economics (ZBW) for using the services, repositories and thesauri.

7. REFERENCES

[1] J. Neubert, "Bringing the ' Thesaurus for Economics ' on to the Web of Linked Data," *LDOW2009*, vol. 25964, no. April 20, 2009, Madrid, Spain, 2009.

[2] K. W. Onn, M. Lim, S. Niu, G. Johannsen, and J. Keizer, "Framework for Matching and Linking Large Ontologies."

[3] R. C. Ojha and S. Aryal, "Digital libraries : Challenges and Opportunities," pp. 3–10.

[4] A. Paepcke, C. K. Chang, H. Garcia-molina, and T. Winograd, “Interoperability for Digital Libraries : Problems and Directions,” pp. 1–23.

[5] G. Buchanan and A. Hinze, “Semantic alerting for digital libraries,” *Proceedings of the 2009 joint international conference on Digital libraries - JCDL ’09*, p. 363, 2009.

[6] C. L. Borgman, “What are digital libraries ? Competing visions,” vol. 35, pp. 227–243, 1999.

[7] P. Hitzler and K. Janowicz, “TheSoz : A SKOS Representation of the Thesaurus for the Social Sciences,” 2009.

[8] L. Y. Sean, A. A. Sadanandan, and D. Lukose, “Scientific Publication Retrieval in Linked Data,” 2012.

[9] P. Anderson, M. Hepworth, B. Kelly, and R. Metcalfe, “What is Web 2 . 0 ? Ideas , technologies and implications for education by.”

[10] A. Hotho, J. Robert, C. Schmitz, and G. Stumme, “BibSonomy : A Social Bookmark and Publication Sharing System.”

[11] D. Benz, A. Hotho, R. Jäschke, B. Krause, F. Mitzlaff, C. Schmitz, and G. Stumme, “The social bookmark and publication management system bibsonomy,” *The VLDB Journal*, vol. 19, no. 6, pp. 849–875, Dec. 2010.