Hajra, Arben; Latif, Atif; Tochtermann, Klaus

Conference Paper — Accepted Manuscript (Postprint)

Retrieving and Ranking Scientific Publications from Linked Open Data Repositories

Suggested Citation: Hajra, Arben; Latif, Atif; Tochtermann, Klaus (2014) : Retrieving and Ranking Scientific Publications from Linked Open Data Repositories, In: Proceedings of 14th International Conference on Knowledge Technologies and Data-driven Business - i-KNOW '14. Graz, Austria. 16-19 September 2014, ISBN 978-1-4503-2769-5, Association for Computing Machinery (ACM), New York, pp. 29:1-29:4,
https://doi.org/10.1145/2637748.2638436

This Version is available at:
http://hdl.handle.net/11108/153

Retrieving and Ranking Scientific Publications from Linked Open Data Repositories

Arben Hajra
South East European University (SEEU)
Bul. Ilindenska 335
1200 Tetovo, Macedonia
a.hajra@seeu.edu.mk

Atif Latif
Leibniz Information Centre for Economics (ZBW)
Düsternbrooker Weg 120
24105 Kiel, Germany
a.latif@zbw.eu

Klaus Tochtermann
Leibniz Information Centre for Economics (ZBW)
Düsternbrooker Weg 120
24105 Kiel, Germany
k.tochtermann@zbw.eu

ABSTRACT
Content enrichment of publications stored in different cross domain Digital Libraries can facilitate the scholarly communication in big way. However, current DL still entails limitation of interoperability between cross domain repositories. This paper emphasizes on this limitation and proposes an innovative approach for finding and recommending scientific publications which are stored in disparate repositories. At first Linked Open Data is considered by exploring existing alignments between Econstor and other datasets within the current LOD cloud through the STW Thesaurus. Moreover, incorporation of other relevant metadata is proposed by implementing a data mining approach which improves the semantic relativeness of the publications from the recommended list.

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]: Information filtering
H.3.7 [**Digital Libraries**]: Miscellaneous

General Terms
Knowledge Management, Recommender Systems, Algorithms

Keywords
Linked open data, semantic web, digital libraries, data mining

1. INTRODUCTION
The Digital Libraries (DL) represents an important place for publishing, discovering and sharing scientific findings. Their usages bring a huge avail for the community and for scholars especially. However, not always what is required can be found in a single location. Publications stored in a repository in most cases belong to a particular domain, described or catalogued according to predefined metadata. Furthermore, according to Paepcke et al. [15], the interoperability among different repositories continues to be a challenge for digital libraries.

As part of this, our goal is to enrich scientific publications stored in a specific repository with other related data, such as information about authors, correlations with other authors' information about conferences, events, projects etc. However, one of the main aims at this stage has to do with finding and extracting other publications, from different repositories that may belong to entirely different areas. Such interoperability would facilitate the scholarly communication by bringing information from many repositories. In order to achieve this, the main direction will be to leverage the already available contents on the semantic web, such as Linked Open Data (LOD) repositories, as one of the most promising data sources [10]. The majority of the evaluations take place at Econstor[1] repository for targeting publications at OpenAgris[2].

This paper begins by highlighting the motivation and problem statement about the current digital libraries. We continue by exploring the research track and proposals, as the main metadata, Linked Open Data and Data mining approaches. The remaining parts will focus on the designed prototype, results and the evaluation.

2. MOTIVATION AND PROBLEM STATEMENT
It is an undisputed fact that libraries represent the most important place for scholarly communication. However, in most cases the user needs to be very lucky to find the "proper" resource with few clicks in a relatively short timeframe. Each search provides a large list of resources offered from the digital library. Therefore, very useful information can be offering a list of recommended papers related to the publication initially sought. Current DL systems mainly offer such services, however, the recommendations are retrieved from the same repository where all publications are. Thus, it would be beneficial for readers to be offered publications related to the initial one, from different repositories and domains, such as social science, agriculture or medicine, which will enrich the given result.

Considering these facts, it can be said that using current DLs, we not always get the most relevant, up to date and cross domain literature [3] [14]. The static metadata structure of DLs present them as monolithic systems, where metadata describe the data rather than usages [2] [4]. Based on this, we raise the need to add other information about a publication to provide a richer and more detailed description. With this goal in mind, we are considering to include "everything" that exists around about the publication; other publications from other disciplines, authors' details, co-authors relations, information about the institute or organization, events, etc.

i-KNOW '14, Sep 16-19 2014, Graz, Austria
ACM 978-1-4503-2769-5/14/09.
http://dx.doi.org/10.1145/2637748.2638436

[1] http://www.econstor.eu/
[2] http://aims.fao.org/openagris

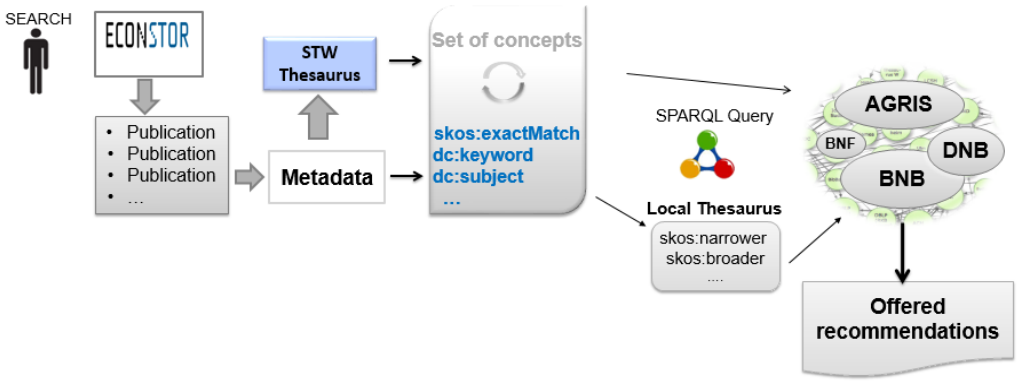

Figure 1. Generating and using the set of concepts from the publication metadata for querying other repositories

The main focus for enriching scientific publications with other information, will be constrained to the content available in a Semantic Web representation, i.e., Linked Open Data (LOD). The publication of a large number of data, such as linked data, provides an excellent opportunity to use them in different scenarios. Hence, the primary source for this purpose will be the LOD cloud where the repositories and thesauri will be highlighted, like STW[3], Agrovoc[4], OpenAgris, TheSoz, DBLP[5], etc. [1] [6]. The idea is to design a generic framework that would include different repositories (taxonomies, thesauri and ontologies) for achieving an interoperability.

3. RESEARCH TRACK AND PROPOSAL

Analysis of the existing metadata which is used to describe a publication in the Econstor repository will be the first step to achieve interoperability goal. Each paper stored in Econstor is described by a wide range of metadata. Besides the commonly included data for title, abstract, and authors, the application of the STW thesaurus provides enrichment with a huge set of descriptors and concepts with the respective mappings to other datasets [13]. Table 1 show the potential metadata set behind a publication in Econstor.

Table 1. The set of all possible metadata behind a publication in Econstor

Title	$T=\{t_1, t_2, t_3, \ldots, t_n\}$
Abstract	$A=\{a_1, a_2, a_3, \ldots, a_n\}$
Descriptors	$D=\{d_1, d_2, d_3, \ldots, d_n\}$
Narrower	$N_1=\{(n_1, d_1), (n_2, d_1), (n3, d_1), \ldots (n_n, d_1)\}$
Broader	$B_1=\{(b_1, d_1), (b_2, d_1), (b3, d_1), \ldots (b_b, d_1)\}$
Related	$R_1=\{(r_1, d_1), (r_2, d_1), (r3, d_1), \ldots (r_r, d_1)\}$
Mappings (Alignments)	$M=\{m_1, m_2, m_3, .., m_n\}$
Keywords	$K=\{k_1, k_2, k_3, \ldots, k_n\}$
Synonyms	$S_1=\{(s_1, k_1), (s_2, k_1), (s3, k_1), \ldots (s_s, k_1)\}$

To get more cross domain alignment for Econstor publication we are experimenting with OpenAgris and Agrovoc thesauri which are a comprehensive multilingual agriculture thesaurus used for indexing the data in OpenAgris [1]. Between Agrovoc and STW there already exists 1136 linked concepts with *skos:exaxtMatch* links. By considering the existing alignments between STW and other Thesauri/Repositories, a key question here would be to find out "How inclusion of additional metadata element i.e. keywords, title, etc., along with existing alignment can improve the quality of the retrieved results". According to this, the implementation of algorithms for vector space model and text mining (calculating the frequency of used term in different documents, similarity between them or relevance ranking) will facilitate the process.

4. PROTOTYPE AND RESULTS

The approaches mentioned in previous sections are implemented and evaluated through a prototype. The prototype has been deployed by using EasyRDF[6] PHP library and OWLIM-SE[7] Semantic repository through Sesame[8]. Let us consider a concrete example. For the publication with title: "**The long run impact of biofuels on food prices**", the prototype creates a set of metadata (see Table 1). Among them, all main descriptors *dc:subject*, are listed followed by the ongoing links, skos:*exactMatch* alignments, to other repositories, as Agrovoc, DBPEDIA, GNDB[9], and GESIS[10]. In addition, the prototype also creates all the narrowed related and broadened concepts. Figure 2 shows detailed view of this for a single descriptor: "Investment".

From these voluminous set of metadata, a special focus will be given to the alignments between repositories. Thus, for the publication mentioned above, specific subsets of mapped links are generated for each dataset in particular. Hence, in total 13 links (concepts) are found, that are mapped between STW and Agrovoc.

By performing a SPARQL query in the OpenAgris repository, with each of these URIs (concepts) in particular, a broad list of 232.154 papers is recommended. In order to deliver more details, only the concept "**Investment**" (see Figure 2) aligned to Agrovoc with the URI http://aims.fao.org/aos/agrovoc/ c_3930, is used as a descriptor in **19.292** papers, where the majority of them are not close with our Econstor paper regarding the similarity. Trying to include all of them with an "and" between, bring a completely

[3] http://zbw.eu/stw/versions/latest/about
[4] http://aims.fao.org/standards/agrovoc/about
[5] http://www.informatik.uni-trier.de/~ley/db/

[6] http://www.easyrdf.org/
[7] http://www.ontotext.com/owlim
[8] http://www.openrdf.org/
[9] http://www.dnb.de/
[10] http://www.gesis.org/

opposite result. The differences in the set of concepts that are used to describe a publication in different repositories result in a null returned publication. Almost impossible to be found a paper described with the same set of descriptors and concepts, in different repositories.

Figure 2. The list of alignments to other dataset together with narrowed, broadened and related concepts, about a single descriptor for an Econstor paper

Another tentative would be to use the thesaurus of the target repository, such as Agrovoc, in a way to narrow or broaden a specific concept there. Afterwards a query could be performed on OpenAgris. Using the local thesaurus of the target repository can be helpful for extending the set of descriptors and concepts, as the hierarchical navigation to them, however the raised challenge as how many and which of them to consider, still remain.

In such situation, we will refer to the set of metadata for a better solution. Hence, the title, abstract, the list of keywords, and synonyms, are promising for gaining a greater precision.

By considering the above example, we are using specific keywords from the title, T={long, run, impact, biofuels, food, prices}, such as "**biofuel**", "**food**" and "**price**". This approach will result in a shorter list of recommended publications from OpenAgris, only **80** titles, with an improved precision.

Thus, the choice, combination and lexical form (plural, singular, synonyms) of the metadata can play crucial role for retrieving publications, in particular by taking into account the semantic relativeness among them. In order to automate this selection, the implementation of vector space model algorithms and text mining techniques will be applied. The main intention would be to calculate the frequency of used terms in different document, similarity between publications, or a way to achieve a ranking among papers in a recommended list.

In line with this, as most promising for this phase, we use Cosine Similarity (CS) and Term frequency–inverse document frequency (TF–IDF) [12]. Hence, a multidimensional array **A** is created and consisted from various terms from the corpus of the publication's metadata (refer to Table 1).

$$A = \{T \cup A \cup \ldots \cup K\} =$$

$$= \{ (t_1, t_2, t_3, \ldots, t_n) \cup (a_1, a_2, a_3, \ldots, a_n) \cup \ldots \cup (k_1, k_2, k_3, \ldots, k_n)\}$$

While on the other side, **B** is an array consisted by the possible metadata from each publication at the target repository. In fact, for each paper in the target repository, a particular array is built, B1, B2, B3, …, Bn. The content of the array **B** depends from the set of metadata offered in the target repository.

By considering an example, for a randomly chosen title from Econstor, like "**The impact of tax reform on new car purchases in Ireland**", a preliminarily search will be performed based on the existing aligned concepts used to describe this paper. So, from more than 4 million bibliographic records, by applying the alignments we get a list of **85.331** titles. Now each of these publications will be measured according to the similarity with the terms at the array **A**.

Thus, for each term in the corpus of **A** we are looking in the corpus of each **B**, from $\mathbf{B_1}$ to $\mathbf{B_n}$. The term with the highest frequency becomes more relevant and showed in top.

$$A = \{T_a \cup A_a \cup K_a\} \text{ looking at}$$

$$B_1 = \{T_{1b} \cup A_{1b}\} \text{ to } B_n = \{T_{nb} \cup A_{nb}\}$$

For example, if **A** is consisted of the Title (T), Abstract (A) and Keyword (K), while **B** only from Title (T_b) and Abstract (A_b), the retrieved results are listed as in Figure 3. The column "Similarity" gives the overall similarity between **A** and $\mathbf{B_x} = \{T_{xb} \cup A_{xb}\}$ while "Title" gives the similarity among the tiles of this paper and titles found in the target repository, **A** and $\mathbf{B}\{T_{xb}\}$. All publications are ranked based on the "Similarity". Following this line of reasoning, we have performed several experiments, by applying different combination into the array **A** and **B**, by giving different weight to several of them, as the title, abstract, keywords, and synonyms.

5. EVALUATION

Based on the current stage of these analyses, alignments between repositories/thesauri are very important for having a preliminary search, especially for reformulating a search query from one vocabulary to another. The presence of a local thesaurus at the target repository can be useful to extend the corpus of concepts for narrowing or broadening the list of retrieved resources. However, incorporating other metadata with data mining approaches improves the list of retrieved publications and offers ranking possibilities according to similarity.

To evaluate these methods, 112 aligned papers from Econstor to OpenAgris are manually experimented with help of prototype. In most of the cases the top ranked publications, according to the similarity are close to the initial paper from Econstor. To obtain a more detailed conclusion, we should consider involving either the experts of the field or a long term end-users usage.

On the other side, the different combinations of arrays **A** and **B**, result in different list of retrieved publications and different similarity percentage. The usage of Keywords in most of the cases significantly improves the results. This can be according to the fact that at Econstor keywords are assigned to each paper manually or semi-automatically by the domain experts. So, their use gives a greater reliability to the array **A**. Otherwise, including the set of Synonyms related to the Keywords, in several cases impairs the result, due to the fact that some irrelevant terms receive a certain weight. Additionally, any external service such as WordNet[11] would be considered for extending the list of terms. All these analyses rise the question which combination works best?

To give an in-depth assessment of our approach, we are going to implement the same methodology in just one repository, i.e.

[11] http://wordnet.princeton.edu/

Nr.	Title	Similarity	Title	AVG
1	EATING OUT: AN IMPORTANT SOURCE OF FOOD FOR THE POOR AND THE FOOD INSECURE	39.87 %	39.28 %	**39.58 %**
2	Consumer choice process when purchasing the staple food	39.66 %	23.57 %	**31.62 %**
3	World food crisis and threats in the food sphere	37.05 %	30.15 %	**33.6 %**
4	Efficiency, energy and stoichiometry in pelagic food webs; reciprocal roles of food quality and food quantity	36.75 %	27.22 %	**31.99 %**
5	Rising food prices intensify food insecurity in developing countries	36.33 %	30.15 %	**33.24 %**
6	THE DEMAND FOR FOOD IN EGYPT: PROBLEMS AND PROSPECTS	34.75 %	21.08 %	**27.92 %**
7	Insegurança alimentar intrafamiliar e perfil de consumo de alimentos	33.49 %	0 %	**16.75 %**
8	The main problems of food allergic consumers concerning food labeling: an ethnographic study	33.19 %	34.43 %	**33.81 %**
9	ANALYSING RUSSIAN FOOD EXPENDITURE USING MICRO-DATA	32.85 %	13.61 %	**23.23 %**

Figure 3. The retrieved list of recommendations from OpenAgris based on a particular Econstor paper

Econstor. To validate our approach, we make a comparison between the results offered by Econstor services and our prototype. Every time a search is performed, the top ten publications on both sides are approximately the same. Only some minor ranking positions might differ by combining different subset of metadata for performing the query. In the same manner, as [5] [11] indicate, any crowdsourcing and pooling techniques can be applied with a smallest subset of papers from different repositories.

Another approach to obtain much reliable results in terms of evaluation, we think to follow some of the practices presented by Benno Stein in [7] [8] and [9], about the evaluation of several successfully applied plagiarism detection algorithms and different strategies for keywords searching.

6. SUMMARY AND FUTURE WORK

A possible content enrichment for publications stored in a DL along with the list of recommendations from different areas and disciplines would facilitate the scholarly communication. In such way, as most promising source we are considering the available semantic web content, especially bibliographic data represented as Linked Open Data. The existing alignments between a specific repository/thesaurus to other repositories are highlighted as crucial for data retrieval and query formulation.

However, the quality of the retrieved publications from the other datasets can be improved by applying different data mining techniques. Hence, in addition to the aligned concept we are also considering all possible metadata behind a paper such as title, abstract, keywords, descriptors, synonyms, etc. Accordingly, the algorithms for vector space model and text mining as TF-IDF in combination with Cosine Similarity (CS) are applied in several scenarios. A ranking among the retrieved publications is achieved as a result of this approach.

The results described in this paper are satisfactory, aside from the fact that we are facing a complex task for a detailed and precise evaluation, if the retrieved results are really the best recommendation. This remains an issue, depending on the right combination of the different metadata for query formulation.

7. ACKNOWLEDGMENTS

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

8. REFERENCES

[1] Anibaldi, S., Jaques, Y., Celli, F., Stellato, A., and Keizer, J. 2013. Migrating bibliographic datasets to the Semantic Web: The AGRIS case. SWJ, IOS Press.

[2] Borgman, C. L. 1999. What are digital libraries? Competing visions. *Inf. Process. Manage. Vol. 35, 3*, p.227-243.

[3] Borgman, L. C. 2013. Digital Scholarship and Digital Libraries: Past, Present, and Future. *In 17th Inter. Conference on Theory and Practice of Digital Libraries*. Valletta, Malta.

[4] Buchanan, G. and Hinze, A. 2009. Semantic alerting for digital libraries. *JCDL '09*, p. 363.

[5] Büttcher, S., Clarke, Ch., Yeung, P. and Soboroff, I. 2007. Reliable information retrieval evaluation with incomplete and biased judgements. *SIGIR '07*, ACM, p.63-70.

[6] De Vocht, L., Softic, S., Mannens, E., Ebner, M. and Walle, R. V. 2014. Aligning web collaboration tools with research data for scholars. *WWW Companion '14,* Geneva.

[7] Gollub, T., Hagen, M., Michel, M. and Stein, B. 2013. From Keywords to Keyqueries: Content Descriptors for the Web. *SIGIR '13*, p.981-984.

[8] Gollub, T., Potthast, M., Beyer, A., Busse, M., Rangel, F., Rosso, P., Stamatatos, E. and Stein, B. 2013. Recent Trends in Digital Text Forensics and its Evaluation. *CLEF '13*, Springer

[9] Hagen, M. and Stein, B. 2010. Search Strategies for Keyword-based Queries. *TIR '10*, IEEE, p. 37-41.

[10] Heath, T. and Bizer, C. 2011. Linked Data: Evolving the Web into a Global Data Space. *Synthesis Lectures on the Semantic Web*

[11] Kazai, G., Kamps, J., Koolen, M. and Milic-Frayling, N. 2011. Crowdsourcing for book search evaluation: impact of hit design on comparative system ranking. *SIGIR '11*. ACM, p.205-214.

[12] Moerland, M., Hogenboom, F., Capelle, M. and Frasincar, F. 2013. Semantics-based news recommendation with SF-IDF+. *WIMS '13*. ACM, Article 22.

[13] Neubert, J. 2012 Linked Data Based Library Web Services for Economics, *Dublin Core and Metadata Applications 2012*, p.12-22.

[14] Ojha, R. C. and Aryal, S. 2009. Digital libraries : Challenges and Opportunities. *In Infolib, Vol.3. Nr.3*, p. 3–10.

[15] Paepcke, A., Cousins, B. S., Garcia-Molina, H., Hassan, W. S., Ketchpel, P., S., Röscheisen, M., and Winograd, T. 2014. Towards Interoperability in Digital Libraries.Stanford University