Limani, Fidan; Latif, Atif; Tochtermann, Klaus

Conference Paper

Scientific Social Publications for Digital Libraries

Suggested Citation: Limani, Fidan; Latif, Atif; Tochtermann, Klaus (2016) : Scientific Social Publications for Digital Libraries, In: Fuhr N. Kovács L. Risse T. Nejdl W. (Ed.): Research and Advanced Technology for Digital Libraries. TPDL 2016., ISBN 978-3-319-43997-6, Springer, Cham, pp. 373-378,
http://dx.doi.org/ 10.1007/978-3-319-43997-6_29

This version is available at:
http://hdl.handle.net/11108/297

Scientific Social Publications for Digital Libraries

Fidan Limani[1], Atif Latif[2], and Klaus Tochtermann[2]

[1] South East European University, Faculty of Contemporary Sciences and Technologies, Tetovo, Republic of Macedonia
~f.limani@seeu.edu.mk
[2] ZBW - German National Library of Economics Leibniz Information Center for Economics, Kiel, Germany
~A.Latif@zbw.eu, K.Tochtermann@zbw.eu

Abstract. Social web content is an important development in the scientific workflow. In this context, scientific blogs are an important medium: they play a significant role in the timely dissemination of scientific developments, and provide useful grounds for discussion and development via the readers feedback. Blogs from the domain of economics are no exception to this practice. A possible extension to Digital Libraries (DL) services, content- and service-wise, is to enable its users access to these blogs. This paper demonstrates an approach for seamlessly integrating scientific blogs in DLs and, with the developed proof of concept application, showcases the resulting benefits for the users and DLs.

Keywords: digital libraries, scientific social publications, Web 2.0

1 Introduction

Going beyond the daily news-, entertainment-, or commercially-related consumption, there is an emergence of a new breed of social web publications in the form of scientific blogs covering many domains and serving an ever-increasing audience, which is lately receiving more and more attention in the scholarly community (Latif et al. 2015). It is via this type of publication that scholars present their ongoing research, discuss latest scientific events, or follow and treat emerging developments (Powell et al. 2012).

Libraries are more and more emphasizing the social component, striving for a community-centered, "social DL" (Calhun, 2013) that engages its users more with their collections (Miller, 2006). The role and awareness of social web resources is already present in the research workflow supporting the whole research lifecycle as a complement to the "mainstream" scholarly publications (CIBER report, 2010). Furthermore, integrating non-library resources with DL collections would allow for a seamless navigation across the resulting collection (W3C LLD Incubator Group Report, 2011).

With over 150 million professional and amateur blogs in the blogosphere (Ragner and Bultitude, 2014) collecting and disseminating science to larger audiences in a wide variety of domains, a significant and growing portion of which

scientific blogs, we are prompted to further explore their contribution to the DL ecosystem. For this there remains a need for an efficient approach to seamlessly integrate and make these resources accessible within a DL environment.

2 Related Work

We see different levels of social web resource integration with different DLs, starting from cases that enable user-provided contributions annotate DL resources for improved search or user engagement, to cases that bring together separate, heterogeneous collections for increased resource selection. Chenu-Abente et al. (2012) explore the opportunity of integrating several components and services of the research process lifecycle by proposing a platform that brings together research publications and researchers with the aim of integrating social and scientific services. In another undertaking, Mutschke and Thamm (2012), via the ScholarLib model architecture, treat the opportunity of relating social web resources with DL repositories in the domain of social sciences. In enabling Web 2.0 features within the DL ecosystem, García-Crespo et al. (2011) with their semantic DL implementation – CallimachusDL, demonstrate the inclusion of user-provided semantic annotations of social web multimedia resources for improved resources browsing and search. In this line of contribution, Gazan (2008) stipulates that the inclusion of Web 2.0 user-annotations into DL can result with "encouraging increased exploration and engagement" of users in the DL environment.

At another level, Wielemaker et al. (2008) exploit the availability of controlled vocabulary (CV) annotation and requirements for integrating heterogeneous sources to support users with search capabilities for the domain of cultural heritage. Another related contribution, that of Holgersen et al. (2012), demonstrates the possibility of enriching bibliographic records of DLs by including user-generated contribution in the form of ratings, comments, tags, etc.; in this case, a set of DL resources is associated with corresponding user feedback, thus enabling the DL to enlist social contributions alongside its resources.

3 Test Dataset Components

For our experimentation purposes, we needed domain-specific social web and DL collections of high quality, and a CV. We choose The Wall Street Journal's (WSJ) economic blog section[3], EconStor[4] and STW, respectively.

EconStor – The DL dataset The publishing/archiving platform EconStor includes Open Access publications from the areas of economics, business administration, and social sciences. The publications in this portal contain high metadata quality and users can rely on a set of feature-rich services to find publications of their interest.

[3] http://blogs.wsj.com
[4] http://econstor.eu/

The Wall Street Journal – The Scientific Blog dataset The WSJ as a domain-specific collection has its focused categories, such as: Bankruptcy beat, Digits, Private equity, Venture capital, etc. Publications from this collection contain high-quality content, making it an attractive proposal to the EconStor portal and a good candidate for our use cases. It is important to mention that blogs usually rely on their own set of terms (tags) when describing posts, different that any complex CV that a DL might use. "Data processing" has the details of how we deal with the "terminology gap".
Thesaurus for Economics (STW) – The CV With about 6,000 descriptors, 20,000 non-descriptors, as well as rich set of "hierarchical" and other associations, STW represents a valuable asset used to index publications on EconStor portal. Moreover, its (manual) alignments to other thesauri allow users to retrieve publications indexed with vocabularies other than STW terms. Any collection can use STW for indexing its resources as long as they are from the domain of economics and related sciences.

4 Approach

The proposed approach entails several activities that streamline blog posts to the DL environment. Following are the details of each activity.
Blog Post Collection Retrieval Based on the URI pattern of the WSJ blogs categories and the Cascading Style Sheet (CSS) rules identifying blog post elements, we retrieved over 41 K blog posts and their corresponding elements during the time period of July – August 2015.
Blog Post (Meta)Data Augmentation Regardless of the blog post category, each WSJ post has the same set of elements: a URL address, title, tags (that describe the post), author, content, publication date, reader comments, share counts, etc. In addition to the original tags, we assign our own set of descriptors to each post using the same vocabulary used by EconStor (see "Data processing" next for more).
Data Processing In addition to the available blog post elements, we conducted automatic indexing of posts based on the STW thesaurus[5]. For this task we used MAUI[6], a mature and open source solution for term assignment with CVs. We decided to assign no more than 5 STW terms to every blog post as the average number of terms (or, in blog's parlance, tags) per blog post dictated this limit. This effectively brings blog posts at the same vocabulary level as publications from EconStor.

5 Proof of Concept Application

In this section we introduce the proof of concept prototype to support use case scenarios motivating our research. We first explain the typical use case scenario steps, and then conclude with an example demonstration.

[5] http://zbw.eu/stw/version/latest/about
[6] https://code.google.com/p/maui-indexer/

5.1 System Prototype

The system prototype is given in Fig. 1; in completing users' requests, it interfaces with both the DL portal (EconStor) and the blog post collection.

The user submits a query; EconStor look-up service processes it and displays the results (Fig. 1, "EconStor Results"). The user selects an article from the result set and its thesaurus-related metadata are retrieved to further support her refine the results (Fig. 1, "Filtering options"). The prototype then uses the same metadata to query the blog post collection for related results (Fig. 1, "Blog post Results").

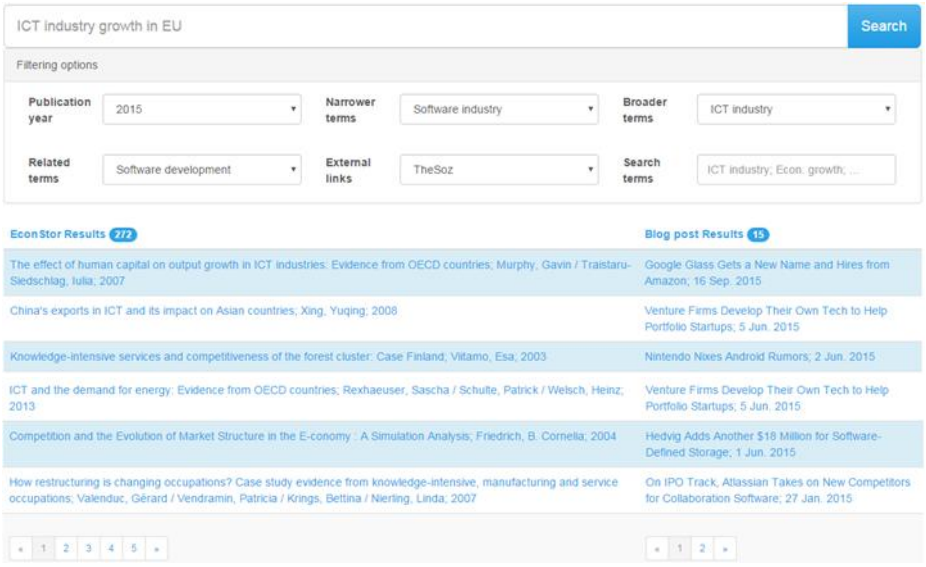

Fig. 1. System prototype

5.2 A Use Case Scenario: Description and Example

The proposed system architecture lends us different use case scenarios to explore. Let's follow with the search scenario illustrated in Fig. 1. A search with "ICT industry growth in EU" presents 272 results from EconStor; the STW terms used in this search are "ICT industry" and "economic growth". The user narrows down the search to "software industry"[7] which reduces the results to 246. In a similar way she can use STW structure and associations to further (re)define her search.

She selects the top-ranked article from the result set; its STW descriptors are "human capital", "ICT industries", and "Economic growth". The prototype

[7] http://zbw.eu/stw/versions/latest/descriptor/24708-1/about.en.html

queries the blog collection with the terms "ICT industry" and its narrower term "software industry" with no results. She chooses another related term, "telecommunications industry", which results in 36 matches; a related term, "telecommunications", results in 13 blog posts. She can further filter out blog posts by date, leaving only the most recent ones in the result set.

The selected EconStor article and its related blog posts show a meaningful relationship. Some of the top-ranked posts discuss the relationship of human capital and ICT-related developments. One post treats the initiative to revamp a Google project (Google Glass) for current and new market segments as well as information on new hires of engineering background to support the initiative, whereas another post discusses applications that could help venture capitals identify investment opportunities. The underlying key terms of these results are: "software development", "software", and "enterprise". There are other terms associated with these top-ranked posts that affect the results such as "team" and "engineers" for the first post, "self help" (used to mean empowerment in the context of social relations) for the second, and "short selling" (in the context of securities trading) for the third post.

6 Resulted Benefits

In this section, in perspective of our proof-of-concept application, we describe the resulted benefits for DL and users as a whole. Lastly we highlight the challenges and limitations faced in the experimental setup.
DL benefits (1) Indexed over 41 K blog posts with DL repositorys CV of choice. This, in turn, enables a seamless search across the resulting collection; (2) Augmented the original blog post meta-data with STW term description, effectively bridging the terminology gap between the DL and blog post collections; (3) Applied a process flow that automates streamlining blog posts to the DL environment; (4) Enabled reuse of existing EconStor services on the newly-added blog posts. **User benefits** (1) Increased resource selection; (2) Social blog authors can rely on the presented process flow in describing their posts with CVs used by the target DL. **Challenges and Limitations** (1) Short blog posts render MAUI unable to propose description terms. (2) Small blog post collection, which, although primarily designed to support our use cases, needs a continuous update in order to provide good basis for resources match-up. (3) Our process flow relies on CSS rules to identify blog post elements, which limits scaling.

7 Conclusion and Future Work

Scientific blogs are getting closer to the traditional research workflow with ever-increasing quality and audience. In this paper we have proposed a strategy to streamline them to DL collections. Moreover, we have developed a proof-of-concept application that highlights the potential benefits of this approach. We are of a view that the integration of DL repository with scientific blogs introduces the issue of information overload for users and needs to be dealt with. In the

future, we plan to experiment with various similarity measures to tackle this problem.

References

Latif, A., Scholz, W. and Tochtermann, K. (2015). Science 2.0 – mapping European perspectives. *Report on the General Stance of Organizations on European Commissions Public Consultation on Science 2.0.*

CIBER report. (2010) Social media and research workflow. University College London, Emerald Group Publishing Ltd.

Calhun, K. (2013) Digital libraries and the social web: scholarship. In Exploring Digital Libraries: Foundations, Practice, Prospects. London: Facet Publishing; Chicago: ALA Neal-Schuman, 2014. ISBN: 9781856048200.

Chenu-Abente, R., Menéndez, M., Giunchiglia, F., De Angeli. A. (2012) An Entity-Based Platform for the Integration of Social and Scientific Services. In *8th International Conference Conference on Collaborative Computing: Networking, Applications and Worksharing*, Collaboratecom 2012 Pittsburgh, PA, United States, October 14-17, 2012.

García-Crespo, Á., Gómez-Berbís, J., Colombo-Palacios, R., García-Sánchez, F. (2011) Digital libraries and Web 3.0. The CallimachusDL approach. In Computers in Human Behavior. 27 (4). p. 14241430.

Gazan, R. (2008). Social annotations in digital library collections. In *D-Lib Magazine*. [Online] 14(11/12). Available from - http://www.dlib.org/dlib/november08/gazan/11gazan.html. [Accessed: 17 January 2016]

Holgersen, R., Preminger, M., Massey, D. (2012) Using Semantic Web Technologies to Collaboratively Collect and Share User-Generated Content in Order to Enrich the Presentation of Bibliographic RecordsDevelopment of a Prototype Based on RDF, D2RQ, Jena, SPARQL and WorldCats FRBRization Web Service. In 4Lib Journal, Issue 17. ISSN: 1940-5758. URL: http://journal.code4lib.org/articles/6695 .

Library Linked Data Incubator Group Final Report. W3C Incubator Group Report. (2011) [Online] Available from: http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025. [Accessed: 15 of January 2016].

Miller, P. (2006) Library 2.0: The challenge of innovation. *Talis* whitepaper.

Mutschke, P., Thamm, M. (2012) Linking Social Networking Sites to Scholarly Information Portals by ScholarLib. In *WebSci 2012*. June 2224, 2012, Evanston, Illinois, USA. ACM 978-1-4503-1228-8.

Powell, D., Jacob, C., and Chapman, B. (2012) Using Blogs and New Media in Academic Practice: Potential Roles in Research, Teaching, Learning, and Extension Douglas. In *Innovative Higher Education*. Volume 37, Issue 4, pp 271-282. Netherlands: Springer.

Ragner, M., Bultitude, K. (2014) The kind of mildly curious sort of science interested person like me: Science bloggers' practices relating to audience recruitment. Public Understanding of Science. DOI: 10.1177/0963662514555054.

Wielemaker., J., et a. (2008) Thesaurus-based search in large heterogeneous collections. (pp. 695-708). Springer Berlin Heidelberg.