Bringing Scientific Blogs to Digital Libraries

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

[1]*ZBW - Leibniz Information Center for Economics, Kiel, Germany*

[2]*South East European University, Faculty of Contemporary Sciences and Technologies, Tetovo, Republic of Macedonia*

Keywords: Digital Libraries, Scientific Blogs, Semantic Web.

Abstract: Research publication via scientific blogging is gaining momentum, with an ever-increasing number of researchers accepting it as their main or complementary research dissemination channel. This development has prompted both scientific bloggers and Digital Libraries (DL) to explore the potential of streamlining these resources along DL collections for increased and complementary user selection. In this paper we explore a methodology for achieving the integration of DL and a blog post collections, together with some use case scenarios that demonstrate the values and capabilities of this integration.

1 INTRODUCTION

Web 2.0, or the ”read-write” Web, has enabled richer levels/channels of engagements/expression with the audiences. This, in turn, has stirred publication initiatives that find them practical in terms of networking, collaboration, ease of publication, (unofficial, peer) reviewing from the community, etc. In the case of scientific blogging, for example, authors can easily share their research work with the community, even as the research develops, whereas readers are able to provide continuous feedback to it. (Burgelman et al., 2010) report on ”Science 2.0” development as enabled by tools and changing research behavior practices, featuring increased number of authors, publications, and data available to consume, reuse, and comment by the community. In another study, (Mahrt and Puschmann, 2014) find ”a dual role of blogs as channels of internal scholarly communication as well as public debate” among the motivations for scientific blogging. Furthermore, the same study finds that science bloggers especially value the community feedback on their posts – an additional explanation for the development and acceptance of ”Science 2.0” from the research community.

Traditional publication repositories have already moved on to embrace the benefits of Semantic Web technologies. Projects that structure and represent Digital Library (DL) repositories as machine-readable and link them up and make them available to the Linked Open Data (LOD) cloud are pretty common. According to the ”State of the LOD Cloud 2014 report”[1] publications take the second largest data set on the LOD. The Library of Congress Linked Data Service[2] offering standards and vocabularies used by the library; the British Library’s LOD initiative[3]; the Swedish National Library Open Data project[4] including bibliography and authority data; German National Library of Economics – EconStor LOD project[5]; are just some of the LOD projects from the domain of DL repositories.

As scientific blogging is getting more contributions and prominence in the research community, we see major benefits from putting its publication contributions to use in different scenarios and environments. In this work we focus on (i) Integrating them with the more traditional DL publication archives, and (ii) ”Porting” them on the Web of Data for supporting current and future applications scenarios.

2 MOTIVATION AND USE CASES

2.1 Motivation

After many requests from the scientific blogging community, the German National Library of Economics

[1] http://linkeddatacatalog.dws.informatik.uni-mannheim.de/state/#toc1

[2] http://id.loc.gov

[3] http://bnb.data.bl.uk

[4] http://libris.kb.se

[5] http://linkeddata.econstor.eu/beta

Limani, F., Latif, A. and Tochtermann, K.
Bringing Scientific Blogs to Digital Libraries.
DOI: 10.5220/0006295702840290
In *Proceedings of the 13th International Conference on Web Information Systems and Technologies (WEBIST 2017)*, pages 284-290
ISBN: 978-989-758-246-2

(ZBW) is considering the opportunity of extending its repository by including scientific blog posts from the domain of economics and offer it to its users alongside the standard research publications. Science bloggers seek to make the most of DLs' dissemination channels and reach higher audience and visibility, whereas DLs seek to complement their collections and increase their value offer to its users. In the face of increasing scientific blog contributions and adoption in the scientific workflow, this is the single most important motivation for this study.

2.2 Use Cases

Following are the main use cases that motivated our research:

(i) Heterogeneous data integration: Blog collections do not adhere to a standardized metadata structure and often rely on different vocabularies from the ones adopted by DLs. In a situation like this, a user interested in resources in both DL and scientific blog resources would have to query these collections separately, using the different vocabulary terms. Thus, there is an opportunity to alleviate this situation and combine these collections in a uniform ”query space”. EconStor, our DL of choice, metadata are already structured and represented in RDF. As a framework, this representation is well equipped to handle combination of heterogeneous resources.

(ii) Semantic annotation of blog posts: More and more resources are being published as LOD, benefiting both publishers and consumers of those resources. Making their resources machine-understandable enables publishers increase the audience reach, including additional (re)use from software agents. On the other hand, LOD publishing enables consumers to (re)use these resources in new scenarios not covered originally by publishers. A final and important note at this point: in case scientific blog post publishers are interested in making their content available in this way, they should not face or be concerned with any technological barriers in the process.

(iii) Dataset profiling: Linking scientific blog resources with relevant entries in external collections and knowledge bases (KB) indexed using different controlled vocabularies (CV) – thesauri and classification schemes, in our case – is another value-adding step for the end user. In this way, depending on the external resource(s) it links to, we provide different ”profiles” for every blog post in our collection, enabling a more elaborate and rich (search) experience to the user. The user potentially benefits from related resources coming from different disciplines – economics, social sciences, or agriculture, in our case; retrieve additional information (and context) from a KB such as DBpedia; include relevant resources from a German-specific or international, multilingual collection; etc.

(iv) Dataset analysis: Summarizing datasets by offering useful statistics as exploration tips for users is quite important. This especially holds for large datasets that could prove challenging for users to comprehend (e.g., identifying resources of certain features, closer to their area of interest). Highly commented/discussed blog posts (expressed via user comments, shares, etc.), the most featured/covered subjects in the collection, ”trending” subjects for a given period of time (based on the number of blog posts for a given subject), top contributing authors (or ”expert groups”) per subject/topic (based on the number of blog posts that an author has for a given subject), or, in the context of aligned CVs of different (linked) datasets, relevant publications by authors in external KBs, are just some of the available analysis options.

3 RELATED WORK

A lot of diversified but concrete research work has been done on mapping relational data (RDB) to graph representation, as well as publishing and integrating heterogeneous collections, including social web data. (Auer et al., 2010) present common motivations for representing RDB in Resource Description Framework (RDF) data model; the use cases for integrating RDB with structured sources or existing RDF on the web (Linked Data) correspond to a great extent with our motivation for this work. Motivated by semantic annotation of dynamic web pages and mass generation of LOD data, (Spanos and Mitrou, 2012) survey the proposed approaches for mapping and integrating RDB content to/with that published as LOD, whereas (Powell et al., 2010) demonstrate fusing library and non-library data from disparate resources, relying on RDF as a common data model and using graph-based analysis and visualization to generate useful information on the resulting data.

Some contributions focus on reusing and enriching social (user generated) content from external resources, such as LOD Cloud, for example: (Holgersen et al., 2012) in their research consider bibliographic-related data (in the form of comments and ratings for books); (Hu et al., 2013) focus on social content from publication submission and review process of a journal, in the form of reviewers' comments, editors' decisions, author replies, etc.; whereas (Passant et al., 2010) reuse collaboratively-built knowledge in the enterprise, contained in different fragmented information resources and represented across heterogeneous data formats in an enterprise setting.

(Yoose and Perkins, 2013) survey the LOD adoption in libraries and report important and increasing number of LOD projects that ”free” library resources from specific library representation formats and enrich them with relevant resources from the LOD Cloud datasets. In a concrete implementation example, (Latif et al., 2014) go through both conceptual and practical aspects of publishing an Open Access (OA) repository as Linked Data.

In this work we integrate library and non-library resources – a DL and a blog post collection. A blog post is not pre-related or referring to any DL publication; bloggers write on any topic they deem important, regardless of the DL repository publications.

4 METHODOLOGY

This section details the applied methodology, including the dataset selection, (pre)processing, modeling and conversion, as well as enrichment from other resources. You can find more details pertaining to section 4.1 and 4.2 in our previous work (Limani et al., 2016).

4.1 Dataset Selection

In order to support our use cases, the dataset selection contains a DL repository and a blog post collection. A final component, a thesaurus, is also presented here for contextual information.

(i) DL repository: EconStor[6] is an open access publication platform for the domain of economics and related fields. It supports the publication dissemination for many institutions, as well as provides its collection metadata to other academic repositories. At the time of writing, it holds more than 123 K conference papers, proceedings, research reports and working papers.

(ii) Scientific blog collection: The Wall Street Journal[7] dataset holds over 40 K blog posts from the domain of economics. In our previous work we annotated this collection with STW thesaurus terms (see section 4.2 for details) to make it terminologically up to par with the EconStor collection.

(iii) The standard thesaurus for economics[8] *(STW)* with about 6.000 descriptors (both in German and English) is a rich vocabulary, primarily covering the domain of economics. STW is used for annotation/indexing and retrieval operations for EconStor publications. Moreover, it is aligned with other CVs (thesauri, classification systems) and has relations to a KB (DBpedia).

4.2 Dataset (Pre)Processing

In order to bridge the ”terminology” gap between the heterogeneous datasets and for identifying the important metadata elements of blog post to our experiment, we conduct:

(i) Automatic indexation: Having the DL dataset collection annotated with STW thesaurus terms determined a similar take on the blog post collection annotation. As described in our previous work, (Limani et al., 2016), we used MAUI[9] for the automatic annotation of every blog post from the WSJ collection with the STW as a vocabulary of choice. This is an important decision as it enables several benefits: bridges the terminology gap between DL and blog post collections; generates semantic annotations for the blog post collection without involving the author/blogger at all; does not require complex integration rules or application of ontology mapping techniques to terminologically ”cross” from one collection to the other.

(ii) Blog post metadata of interest: Blogs tend to differ in the vocabularies they select to annotate their collection, although they usually have common (to not say standard) metadata elements in their blog posts. In our case, besides the usual blog post metadata, such as author, title, content, and publication date, we also retained comments for each blog post. Some social web features such as ”shares” and ”tweets”, although present in the data set, were left out of the current research analysis for future studies.

4.3 Dataset Modeling and Conversion

This part covers three activities that complete the dataset modeling and conversion process, including: (i) selected vocabularies for the blog post collection, (ii) technical details of mapping the blog post collection from a relational database to RDF; and (iii) EconStor and blog post collections merge for a unified query space, available for querying and exploration via an HTTP server for RDF data.

(i) Blog post vocabulary selection: Despite the Web 2.0 nature, blog posts reflect the common metadata that a DL publication has, such as post title, publication date, content, terms describing the post; whereas also having some additional aspects that

[6] http://econstor.eu

[7] blogs.wsj.com

[8] http://zbw.eu/stw

[9] https://github.com/zelandiya/maui

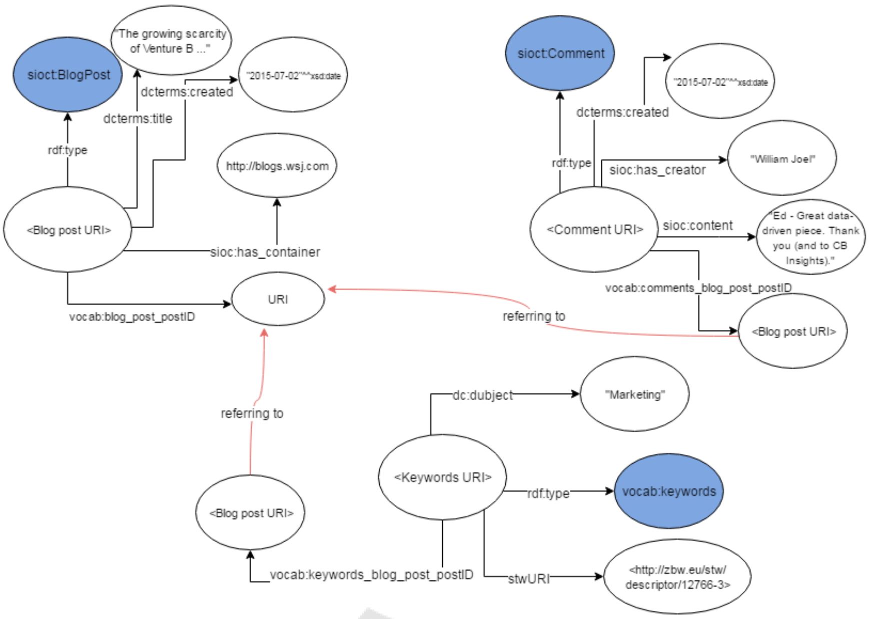


Figure 1: Classes (colored in blue) and properties modeling a blog post instance.

are inherent to them, such as user-generated feedback (blog post comments and other Web 2.0 ”features” like shares, tweets, etc.). The key vocabularies selected for modeling blog post collections are SIOC[10](and SIOC Types module) and the Dublin Core Metadata Initiative[11]; the former covering especially well the user-generated content, whereas the latter the typical publications metadata. Figure 1 shows an example representation of a blog post instance (see below) with a single comment and subject keyword (”Marketing”, in this case). One of the key parts of the modeling refers to blog posts and related comments, with SIOC Types `BlogPost` and `Comment` classes used for the blog post and their comments, respectively. The default D2RQ mapping for the entity that covers the keywords used to annotate every blog post completes the instance modeling.

(ii) Mapping relational data to RDF: The DL collection is represented as RDF triples, whereas the scientific blog post collection is stored in a relational database. To bring both collections into the same model, we used the D2RQ platform[12] to generate RDF data representation of the blog post collection. This required mapping the relational database tables and columns to RDF classes and properties based on the vocabularies identified beforehand. With it, both datasets are combined via their common data model (RDF).

(iii) Unified query space over datasets: There is an important thing to mention at this point: since both collections use the STW thesaurus to annotate/index their resources, this eliminates need for any ontology alignment between the two datasets for the use case scenarios. Furthermore, this implies a closer connection between the datasets and renders them more integrated (at a vocabulary level) than just two datasets sharing the same representation (RDF, in this case). This condition enables us to address the combined datasets as being part of a single ”information space”; and we can solely rely on STW terms to search resources in both datasets. Both collections are loaded as a single dataset in a SPARQL endpoint, Apache Jena Fuseki[13], each being part of a named graph within the single dataset (`econstor` and `wsj`, respectively).

5 USE CASE SCENARIO DEMONSTRATION

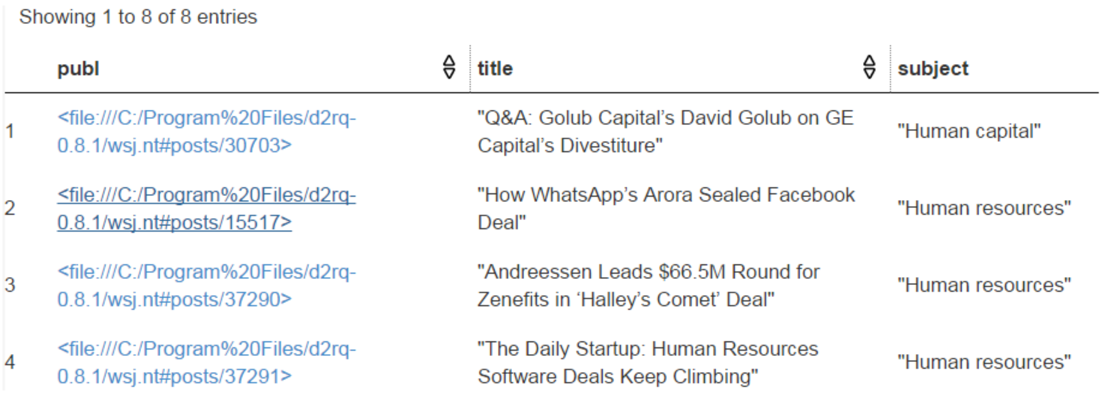
In this section we represent several use case scenarios implemented from integrating the DL repository and the scientific blog post collection that directly address the motivation for this paper. Although DL users are not expected to know SPARQL in order to search the collection, by exploring some query scenarios, we want to demonstrate that this collection can serve as a data store on top of which we can build

[10] http://rdfs.org/sioc/spec/

[11] http://dublincore.org/specifications/

[12] http://d2rq.org

[13] https://jena.apache.org/documentation/fuseki2/index.html



Showing 1 to 8 of 8 entries

	publ	title	subject
1	<file:///C:/Program%20Files/d2rq-0.8.1/wsj.nt#posts/30703>	"Q&A: Golub Capital's David Golub on GE Capital's Divestiture"	"Human capital"
2	<file:///C:/Program%20Files/d2rq-0.8.1/wsj.nt#posts/15517>	"How WhatsApp's Arora Sealed Facebook Deal"	"Human resources"
3	<file:///C:/Program%20Files/d2rq-0.8.1/wsj.nt#posts/37290>	"Andreessen Leads $66.5M Round for Zenefits in 'Halley's Comet' Deal"	"Human resources"
4	<file:///C:/Program%20Files/d2rq-0.8.1/wsj.nt#posts/37291>	"The Daily Startup: Human Resources Software Deals Keep Climbing"	"Human resources"

Figure 2: Results illustration in a SPARQL server.

a standard user interface for search that users understand (keyword-based search, in the same way they use a search engine or search a document in their computer).

(i) Search across the "unified query space:" The user searches for publications related to the subject of "technology transfer" in EconStor and WSJ datasets. As mentioned earlier, there are several types of publications archived in EconStor, but, in this case, the user is interested in research papers (i.e., `swc:Paper`), published since 2014. The search returns four results in total, with three results coming from the EconStor dataset and one coming from the blog post collection. This just demonstrates the possibility for the user to search across two different datasets described with the same thesaurus term(s), and receive publication results from corresponding datasets (treating the datasets as if they are one source of information).

(ii) Retrieve relevant blog posts from the collection: This scenario is related to the previous one: the user initially searches the DL collection (i.e., EconStor) and selects a publication that she wants to further examine. We search the blog post collection for additional publications that could be of interest to her based on the STW term(s) that describe the publication she is currently reading. The user searches for (`swc:Paper`) publications from EconStor that cover the subject of "Human capital", published from 2014 and onwards. The user selects the publication titled "Labour market integration, human capital formation, and mobility" from the result list. Using the same STW term ("Human capital") that describes the selected publication, gives us 1 blog post from the WSJ collection, titled "Q&A: Golub Capital's David Golub on GE Capital's Divestiture", as well as 7 other posts described with the "related" STW term "Human resources" that could further complement user' reading experience (Figure 2 shows part of the retrieved results in Apache Jena Fuseki). This further emphasizes the role that the (STW) thesaurus can play in providing alternative results for the user by using its structure, such as via "narrow", "broad", or "related" terms.

(iii) Search the scientific blog post collection alone: In another scenario, the user searches for the newest blog posts covering a certain subject. During this scenario, the user can decide to factor in the number of comments that a blog post has, i.e. the post that stirred the most feedback/discussion on a given subject, or explore the most used STW terms from the collection, in order to have a better understanding on the variety of blog posts that constitute the collection. Let's see how these two search strategies work for our blog post collection:

– Highest number of comments: This search, filtered by posts published from 2015 and on, lists the following top 3 blog posts with the highest number of comments: "Facebook Plans a 'Dislike' Button, but Only for Empathy, Zuckerberg Says" with 18 comments; "Microsoft Expected to Unveil Next-Gen Windows Phone and Surface Tablet" with 12 comments; and "Alabama Judges 'Reprehensible' Conduct Merits Impeachment, Judiciary Says", the last post stressing a judicial misconduct by a judge, with 8 user comments from the blog readers.

– The most featured blog posts by STW term: This is an attempt to mimic "topic trending" in the blog post collection – showing the extent to which certain subjects are covered (via posts) in the blogging community. In our case, searching for the most used STW terms in the blog post collection results with the top 3 most used terms "Enterprise", "Personalization", and "Share". This provides some hints to DL users about the most represented/covered subjects from the blog post collection, in case they want to use that information to guide their exploration of this collection.

– A combination of the two: Having identified the most used STW terms, the user can further explore the most commented on blog post from a popular subject, which with regards to our blog post collection results to combining posts on the subject of ”Enterprise”, ”Personalization”, or ”Share” (as discussed above), and blog posts that attracted the most attention in the blog community (the top 3 blog posts listed above).

6 RESEARCH BENEFITS

The key benefits relate to the automatic indexing and semantic annotation of the scientific blog post collection, its integration with the DL collection in a unified (in terms of querying and resource description via the STW thesaurus) dataset, as well potential data profiling and analysis operations. Following are the emphasis on these aspects:

(i) Semantic annotation and representation of blog post collections: Having the DL collection published as LOD dictates the methodology of blog post collection integration with the DL. Without any effort from the bloggers’ side, we have modeled and represented this collection in the same way as the DL collection – thus making them part of the same ”model” (RDF, in this case), and automatically indexed it (based on the STW thesaurus) – thus bridging the terminology gap between these different resources and integrating them at a terminology level.

(ii) Integration of (originally) heterogeneous collections for a seamless and unified ”query space”: Meeting DL’s interest to include heterogeneous resources – blog posts from the same domain, we have integrated the latter and made it available as a resource collection to the former. The users of the DL library, as shown with our queries over the resulting dataset collection, are able to retrieve relevant resources via different scenarios.

(iii) Data profiling and analysis (both indirect benefits from relying on STW for indexing the blog post collection): STW’s alignment with other thesauri, classification systems, and external KBs enables us to enrich the user search experience by linking up scientific blog posts of interest to the user with external, related resource collections. Moreover, we are able to provide useful information about the dataset to the user, such as ”trending” topics/subject for a given time period, the most popular topic/subject, or the blog posts that sparred the most debate with the users. For more details, see the implemented use case scenario implementations from section 5 of the research paper.

7 CONCLUSION AND FUTURE WORK

In this paper we have addressed a Digital Library requirement by integrating non-library resources – a scientific blog post collection – and making it available to its users as a complementary content in their search operations. In doing so, we have pre-processed the non-library resources in order to bring them up to par (vocabulary-wise) with the DL practices (assigning STW terms, in this case); modeled them according to the DL collections representation (RDF, in this case) by selecting a set of suitable vocabularies (and corresponding classes and properties); and finally converting them from a relational database to an RDF representation using the D2RQ platform. Furthermore, in order to support the use case scenarios, we loaded both the library and non-library datasets on separate named graphs of a single dataset on a SPARQL server.

One of the future work task is to develop a prototype, hence enabling evaluation scenarios with the users. Currently, in order to query the unified dataset implies knowledge of SPARQL, which is not a skill that common DL users should have in exploring a DL collection.

Another research follow up direction is that of analysis that would bring more value to the user (search) experience in view of the newly-added blog post collection, such as publications similarity based on the STW thesaurus structure and graph representation properties, to name a few.

REFERENCES

Auer, S., Feigenbaum, L., Miranker, D., Fogarolli, A., and Sequeda, J. (2010). Use cases and requirements for mapping relational databases to rdf.

Burgelman, J.-C., Osimo, D., and Bogdanowicz, M. (2010). Science 2.0 (change will happen.). *First Monday*, 15(7). Accessed: 28 June 2016.

Holgersen, R., Preminger, M., and Massey, D. (2012). Using semantic web technologies to collaboratively collect and share user-generated content in order to enrich the presentation of bibliographic records. *Code4Lib Journal*, 17. Accessed: 20 June 2016.

Hu, Y., Janowicz, K., McKenzie, G., S. K., and Hitzler, P. (2013). A linked-data-driven and semantically-enabled journal portal for scientometrics. In *12th International Semantic Web Conference*. Springer Berlin Heidelberg.

Latif, A., Borst, T., and Tochtermann, K. (2014). Exposing data from an open access repository for economics as linked data. *D-Lib Magazine*, 20(9/10). Accessed: 7 June 2016.

Limani, F., Latif, A., and Tochtermann, K. (2016). Scientific social publications for digital libraries. In *20th International Conference on Theory and Practice of Digital Libraries*.

Mahrt, M. and Puschmann, C. (2014). Science blogging: an exploratory study of motives, styles, and audience reactions. *Journal of Science Communication*, 13(3).

Passant, A., Laublet, P., B. G., and Decker, S. (2010). Semslates: Weaving enterprise 2.0 into the semantic web.

Powell, J., Collins, L., and Martinez, M. (2010). Semantically enhancing collections of library and non-library content. *D-Lib Magazine*, 16(7/8). Accessed: 10 June 2016.

Spanos, D.-E., S. P. and Mitrou, N. (2012). Bringing relational databases into the semantic web: A survey. *Semantic Web*, 3(2):169–209.

Yoose, B. and Perkins, J. (2013). The lod landscape in libraries and beyond. *Journal of Library Metadata*, 13(2-3):197–211.