

Limani, Fidan; Latif, Atif; Tochtermann, Klaus

**Conference Paper — Accepted Manuscript (Postprint)**

## Bringing Scientific Blogs to Digital Libraries: An Integration Process Workflow

*Suggested Citation:* Limani, Fidan; Latif, Atif; Tochtermann, Klaus (2018) : Bringing Scientific Blogs to Digital Libraries: An Integration Process Workflow, In: Majchrzak T. Traverso P. Krempels KH. Monfort V. (Ed.): Web Information Systems and Technologies. WEBIST 2017, ISBN 978-3-319-93527-0, Springer, Cham, pp. 171-178, [https://doi.org/10.1007/978-3-319-93527-0\\_8](https://doi.org/10.1007/978-3-319-93527-0_8)

This Version is available at:  
<http://hdl.handle.net/11108/365>

### **Kontakt/Contact**

ZBW – Leibniz-Informationszentrum Wirtschaft/Leibniz Information Centre for Economics  
Düsternbrooker Weg 120  
24105 Kiel (Germany)  
E-Mail: [info@zbw.eu](mailto:info@zbw.eu)  
<https://www.zbw.eu/de/ueber-uns/profil-der-zbw/veroeffentlichungen-zbw>

### **Standard-Nutzungsbedingungen:**

Dieses Dokument darf zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden. Sie dürfen dieses Dokument nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen. Sofern für das Dokument eine Open-Content-Lizenz verwendet wurde, so gelten abweichend von diesen Nutzungsbedingungen die in der Lizenz gewährten Nutzungsrechte.

### **Terms of use:**

*This document may be saved and copied for your personal and scholarly purposes. You are not to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public. If the document is made available under a Creative Commons Licence you may exercise further usage rights as specified in the licence.*

# Bringing scientific blogs to Digital Libraries: An integration process workflow

Fidan Limani, Atif Latif, and Klaus Tochtermann

ZBW – Leibniz Information Center for Economics, Kiel, Germany,  
{f.limani, a.latif, k.tochtermann}@zbw.eu

**Abstract.** Scientific blogging is continuously gaining importance in research communities as a complementary support to scientists during different phases of research and publication lifecycle. By enabling early feedback from the community (peers commenting on one’s work as early as the first draft, for example), and providing faster time-to-publish cycles, or tracking audience reach to some extent (via Web 2.0 features, such as shares, likes, etc.), it is becoming an important medium for research(ers). While blogs certainly bring their differences when compared to the more traditional research papers (shorter than typical research articles; include tags – a way of community-driven controlled vocabulary – to aid during blog post retrieval, etc.), they are becoming attractive as complementary scientific resources for Digital Libraries (DL). In this work, we present a complete workflow that spans retrieval, processing, vocabulary handling, and finally integration of a scientific blog posts collection to a DL collection. Moreover, the Use case scenarios that demonstrate the value of the workflow outcome to Digital Libraries along with future development plans are also presented in this paper.

**Keywords:** Scientific blogs, Digital Libraries, Integration workflow, Linked Data

## 1 Introduction: Social scientific what?

Web 2.0, or the “read-write” Web, has enabled higher and easier levels of engagements with the audiences. This, in turn, has stirred individual and group publication initiatives that find these tools practical in terms of networking, collaboration, ease of publication and (unofficial) reviewing from the community, etc. In the case of scientific blogging, for example, authors can easily share their research work with the community, even continuously – as the research develops, be it a result or roadblock that is being shared; whereas the readers are able to provide feedback to this work as it progresses along those results or roadblocks. (Burgelman et al. [2]) report on “Science 2.0” development, as enabled by tools and changing research behavior practices (spurred by factors such as lowering entrance barriers, offering processing data and capabilities to wider audiences, etc.), characterized by increased number of authors, publications, and data available to consume, reuse, and comment by the community. In

another study, (Mahrt and Puschmann [9]) 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 argument 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 2017"<sup>1</sup>, publications take the second largest data set on the LOD). The Library of Congress Linked Data Service<sup>2</sup> offering standards and vocabularies used by the library; the British Library's LOD initiative<sup>3</sup>; the Swedish National Library Open Data project including bibliography and authority data<sup>4</sup>; German National Library of Economics's EconStor LOD project<sup>5</sup>; 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 all the author, publications, and research data contributions to use in different scenarios and environments. In this work we focus on (1) Integrating them with the more traditional DL publication archives, and (2) "Porting" scientific blogs on the Web of Data for supporting applications benefiting from these resources in the future.

## 2 Motivation and Use cases

### 2.1 Motivation

After many requests from the scientific blogging community, the German National Library of Economics (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. This is the single most important motivation for this study. This motivation is just part of a broader picture of increasing scientific blog contributions and their adoption in the scientific work flows, which furthermore emphasizes the importance of publications from the scientific blogging community.

### 2.2 Use cases

Following are the main use cases that motivated our research

---

<sup>1</sup> <http://lod-cloud.net/>

<sup>2</sup> <http://id.loc.gov>

<sup>3</sup> <http://bnb.data.bl.uk>

<sup>4</sup> <http://libris.kb.se>

<sup>5</sup> <http://linkeddata.econstor.eu/beta>

1. *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[6]. As a framework, this representation is well equipped to handle combination of heterogeneous resources.
2. *Semantic annotation of blog post collections:* More and more resources are being added to the Web of Data, benefiting in this way both publishers (making their resource machine-understandable and available to an additional pool of users) and consumers of those resources (for example when integrating these resources in environments of interest, DL repositories in our case). In case scientific blog post publishers are interested in making their content available on this platform, they should not be concerned with any technological barriers to achieve this goal.
3. *Dataset profiling:* If "isolated", the scientific blog post collection has limited value for the user. Linking these resources with relevant entries in external collections and knowledge bases (KB), indexed using different controlled vocabularies (CV) – thesauri and classification schemes, in our case – from the one used by scientific blogs, but that have been aligned with, is another value-adding step for the end user. In this way we provide different "profiles" for every blog post in our collection, all depending on the external resource(s) it links to, enabling the user a more elaborate and rich (search) experience. The user search experience could see improvements from related resources to their current reading from the fields of economics, social sciences, or agriculture; retrieve additional information (and context) from a KB, such as DBpedia; include relevant resources from a German-specific or international, multilingual collection; etc.
4. *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 grasp in order to use them to their full capacity (e.g., identifying resources of certain features, closer to their area of interest). In this regards, scientific blog posts semantification lends us various analysis options, such as highly commented/discussed (expressed via user comments to) blog posts; most featured/covered subject in the collection (based on their topic coverage); "trending" subjects for a given period of time (based on the number of blog posts for a given subject); the top contributing authors per subject/topic, which would, in a way, identify expert groups on certain subjects/topics; or, in the context of aligned CVs of different (linked) datasets, suggest publications by an author in external KBs; etc.

### 3 Related work

Blog post collections require established publication and dissemination infrastructure, as well as increased visibility of their content. DL repositories have this, and would consider an added value by offering them alongside their original bibliographic content. There is solid research done on mapping relational data (RDB) to graph representation, as well as publishing and integrating heterogeneous collections, part of which also includes social web data. Furthermore, DL repositories are more and more moving to the Web of Data making this possible inclusion even easier with already established approaches and practices.

Auer et al. [1] 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<sup>6</sup>) correspond to a great extent with our motivation for this work. Moreover, Spanos et al. [13] survey the proposed approaches for mapping and integrating RDB content to/with the Web of Data, with citing semantic annotation of dynamic web pages and mass generation of Semantic Web data as some of the key motivations of the work.

Holgersen et al. [3], with their semantic-web based framework, part of the Swedish "The Open Library" project implementation, transform social content repository data (user-generated content in the form of tags, reviews, comments, etc.) in a Resource Description Framework (RDF) machine-readable format, to further enrich it by related data from external collections, and provide (as a Web Service) the resulting data to libraries that hold related bibliographic records to benefit from (this enrichment). As a result, libraries can associate and show all their bibliographic records and related social content to their users. Hu et al. [4] publish a journal collection based on Linked Data principles, which, besides bibliographic data, includes information on the publication submission and review process of the journal, such as reviewers' comments, editors' decisions, author replies, to name few of the (meta)data. The data is further enhanced by relevant resources from the LOD Cloud, such as DBpedia and Semantic Web Dog Food<sup>7</sup>, and made available as part of the LOD Cloud<sup>8</sup>.

Atif et al. [6] go through both conceptual and practical aspects of publishing an Open Access (OA) repository as Linked Data. To demonstrate the potential from this undertaking, they link up the repository collection with the LOD Cloud datasets (via mappings to an economics thesaurus, lexicon-related service, and an economics classification system – all part of the LOD Cloud) for contextualizing the OA repository collection and enabling more discerning queries to be conducted over it.

Powell et al. [10] demonstrate fusing library and non-library data from disparate resources by applying Semantic Web technologies to the task. They rely on RDF as a common data model, and use graph-based analysis and visualiza-

<sup>6</sup> <http://linkeddata.org/>

<sup>7</sup> <http://data.semanticweb.org/>

<sup>8</sup> <http://lod-cloud.net/>

tion to generate useful information on the resulting data for the user. In another work, Yoose and Perkins [14], in a survey on LOD adoption in libraries, including archives and museums, report important and increasing number of L(O)D projects that free library resources from specific library representation formats, benefiting the consumption of this data from interested parties, and further improve the library experience for its users by enriching this (meta)data with relevant resources from the LOD Cloud datasets.

Passant et al. [11] research the application of Semantic Web technologies in an industry context for enhancing the Enterprise 2.0 setting. In their study they present a paradigm for reusing collaboratively-built knowledge in the enterprise, contained in different fragmented information resources and represented across heterogeneous data formats, which, to also provide more insightful query capabilities to the end users.

Due to the said prominence of blog posts, there are other research efforts for collecting and developing value-adding services based on these resources. In a related undertaking, Papadokostaki et al. [12] develop a platform for storing, indexing and searching blog posts. In their effort they apply semantic web technologies throughout the project, starting from developing a custom ontology for modeling the blog posts, storing the resulting posts (as RDF) in a triplestore, as well as relying on a Linked Data format (JSON-LD) for their REST services; when modeling the domain news articles, the authors develop a new ontology. All blog posts added to the collection are also made available to the Linked Open Data cloud. Users of the portal can benefit from searching the blog post collection created by automatic ingestion of blog posts, or have the possibility of manually adding blog posts to the collection. The BlogForever platform (see Kalb et al. [5]) focuses on archiving operations for the blogosphere; in addressing the potential information silos from this undertaking (blogs from different domains, using different vocabularies, etc.), the authors adopt Linked Data principles in their approach. A common domain model for blog archiving and a set of vocabularies enable exposing resulting blog archives as LOD. The benefits from this work are manifold, such as: possibility to search across different blog archives; link archives to external resources or collections; etc.

Harnessing the value from connecting heterogeneous resources – in this case blogs and DLs – is specifically the motive for our work in this paper. In this work, we integrate blog post collections external to and independent of the DL collection for increased visibility to the former, and enriched offer of the latter. Blog posts typically are not pre-related or do not refer to DL publications (such as via a publication DOI, or other identification mechanisms); and 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 for this work, its (pre)processing and augmentation, modeling and conversion, as well as enrichment (via linking) from other resources. For additional information

pertaining to section 4.1 and 4.2, we refer you to our previous work in Limani et al. [7]).

#### 4.1 Data selection

In order to support our use cases, our dataset selection contains a DL repository and a blog post collection. A final component, a thesaurus, is also presented here for contextual information; its role will be detailed as we go further in the section.

1. *DL repository*: Our requirements for this part are open access policy of the repository holding the DL collection, and potentially the existence of a controlled vocabulary in order to emulate the DL environment as closely as possible. The former eases the aspects of retrieving and using collection resources, whereas the latter brings up the challenge of streamlining resources with different vocabularies. EconStor<sup>9</sup>, the DL representative choice for this work, is an open access publication platform for the domain of economics and related fields, supporting publication (at the time of writing with more than 140 thousands working papers, journal articles and conference proceedings, etc.) dissemination for many institutions, as well as providing its collection metadata to other academic repositories.
2. *Scientific blog collection*: For this part we needed qualitative and rich blog collection to pair up with the standard that EconStor collection employs. With over 40 K blog posts from the domain of economics that we harvested for this work, we chose the The Wall Street Journal<sup>10</sup> blog for our work. A few notes are in order to provide the rationale of our choice of the WSJ as our scientific blog representative. There are two main reasons for our decision:
  - (a) Although not scientific blog per se, being a popular authority on economics publishing, it provides a useful resource for a DL that covers the economics (and related fields') domain. We feel this could be a real use case that users of the DL will appreciate (and there is an evaluation plan to assess this as our future work); in the same way, other similar publications – relevant and of high quality as the DL collection – could serve in the same way for our research process workflow.
  - (b) For our workflow we needed a rich selection of blog posts in order to demonstrate its usefulness in a DL setting. For that we needed a single, rich blog to save us from having to deal with different approaches especially during the blog posts retrieval, and one with regular updates and high-quality posts and relevant authors. The WSJ is just one of the many such collections that offer blogging on relevant and up-to-date topics from the economics domain.

---

<sup>9</sup> <http://econstor.eu>

<sup>10</sup> [blogs.wsj.com](https://blogs.wsj.com)

3. *The standard thesaurus for economics (STW<sup>11</sup>)*: With about 6.000 descriptors (both in German and English), it is a rich vocabulary primarily covering the domain of economics, as well as related fields (law, sociology, politics, etc.), used for indexing and retrieval operations for EconStor publications. Furthermore, its alignment with other CVs and collections increase its "reach" and value: thesauri for social sciences and agriculture (TheSoz and Agrovoc, respectively); classification system for publications from the fields of economics (JEL), as well as with publications used in German libraries (German National Library Integrated Authority File); or even relations to a KB (DBpedia). These alignments are important to mention as they enable the realization of some of the benefits listed in section 2 of this paper, in "Dataset profiling" and "Dataset analysis" subsections.

## 4.2 Dataset (pre)processing and augmentation

The "Data selection" subsection emphasized on the differences between selected data sets – DL and scientific blog collections. To account for this aspect, our methodology includes activities that (i) bridge the "terminology" gap between the heterogeneous data sets, and (ii) identify blog post metadata that are important for our use cases. Specifically, we conduct:

1. *Automatic term assignment based on a CV*: Having to deal with heterogeneous resources brings its own challenges to the table. Namely, after having selected our data set as described in the previous subsection, we need to work with their inherent differences seamlessly. With vocabularies used in either collection being the most prominent difference, we need to be able to conduct all our analysis on these collections as if it were a single, homogeneous data set, regardless of any of the differences mentioned so far.  
Having the DL data set (EconStor) described by the STW thesaurus determined a similar requirement for the blog post collection – i.e., describe it using the same set of vocabulary terms. Due to the fact that blog posts are created at a frequent pace (every blog could have many posts added every day), we needed to automate this step. In this work we used a mature and popular automatic indexer (MAUI<sup>12</sup>) to automatically assign terms to every WSJ blog post. One advantage that MAUI gives us is the possibility to specify a CV as a source from where to choose terms from during the term assignment process. This is exactly what we needed for our case: MAUI suggested terms to WSJ blog posts using the STW thesaurus as a source. Examples of this term assignment can be found in the "Process workflow" section.
2. *Blog post metadata selection*: Blogs differ in the attributes they employ in their post collection. In our case, besides the usual blog post metadata, such as author, title, content, and publication date, we also retained blog

<sup>11</sup> <http://zbw.eu/stw>

<sup>12</sup> <https://code.google.com/p/maui-indexer/>



post comments for each blog. Some social web features, such as "shares" and "tweets", although present in the data set, were left out of the current research analysis. We have plans to include them in our future work.

### 4.3 Dataset modeling and RDF conversion

This part of the methodology covers three activities that complete the data set modeling and conversion process, including: (i) listing selected vocabularies for the blog post collection, (ii) providing the technical details of mapping the blog post collection from a relational database to RDF<sup>13</sup> via a platform that "wraps" the RDB into a virtual RDF graph; and (iii) combining EconStor and WSJ blog post collections in a single collection, to be used as a unified query space; this activity also provides for query and exploration capabilities via an HTTP server for RDF data.

1. *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 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<sup>14</sup> (and SIOC Types module) and the Dublin Core Metadata Initiative<sup>15</sup>; the former covering especially well the user-generated content, whereas the latter the typical publications metadata. An example representation of a blog post is shown in Figure 1 below, with a single comment and subject keyword ("Marketing", in this case) due to easier readability. One of the key parts of the modeling refer to blog posts and related comments, with SIOC Types `BlogPost` and `Comment` classes used for the blog post and their comments, respectively. The modeling requirements of blog posts did not necessitate developing a new ontology, so we reused already established vocabularies for this process.
2. *Mapping relational data to RDF:* The DL collection is represented via Resource Description Framework (RDF) standards as (RDF) triples, whereas the scientific blog post collection is stored in a relational database. To bring both collections into the same model, we use a specific platform to generate an RDF data dump of the blog post collection. This requires mapping the relational database tables and columns to RDF classes and properties, based on the vocabularies identified beforehand. With this both data sets are combined in a single data model – RDF.

Figure 1 displays the key classes and properties used to model the blog post collection. The modeling is based on three "entities" of a WSJ blog post: the blog post itself, its comments, and the subject it covers (represented

<sup>13</sup> <https://www.w3.org/RDF/>

<sup>14</sup> <http://rdfs.org/sioc/spec/>

<sup>15</sup> <http://dublincore.org/specifications/>

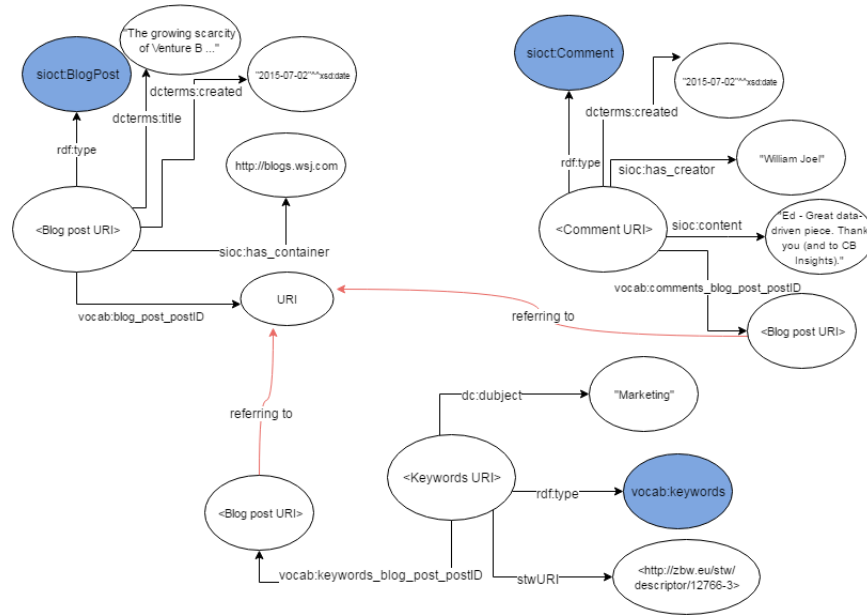


Fig. 1: Classes (colored in blue) and properties modeling a blog post instance.  
Source: Limani et al. [8]

here based on the STW terms automatically assigned to the post). The blog post entity and comment entities are mapped to SIOC Types BlogPost and Comment classes, respectively; the last class that of blog post terms uses a D2RQ platform-internal vocabulary. The properties used to describe this entity, i.e. the subject it covers and the link to the STW URI, were enough to support our use cases; we did not find it necessary to use a vocabulary to specifically model the blog post terms beyond the one that D2RQ platform provided during mapping.

3. *Unified query space over data sets:* At this point both collections use the STW thesaurus to describe their resources; this implies a closer connection of the data sets and renders them more integrated than just two data sets sharing the same representation (RDF, in this case). Rather, this enables us to address the combined data sets as being part of a single "information space". In this way, we can solely rely on the STW terms to search resources in both data sets.

## 5 Tying it all: A process workflow

After discussing the methodology followed in this work, in this section we turn to a concrete implementation represented as a process workflow which is depicted in Figure 2. In following, we provide technical details and rationale for every process

of this workflow. Concretely, we show the complete process of making scientific blog posts available to the DL environment, and provide technical details and information about the tools used to achieve this goal. We believe that this could help anyone planning to apply or extend the workflow according to their needs for their DL environment or in a completely new project. We provide some potential extension examples as we go through the process.

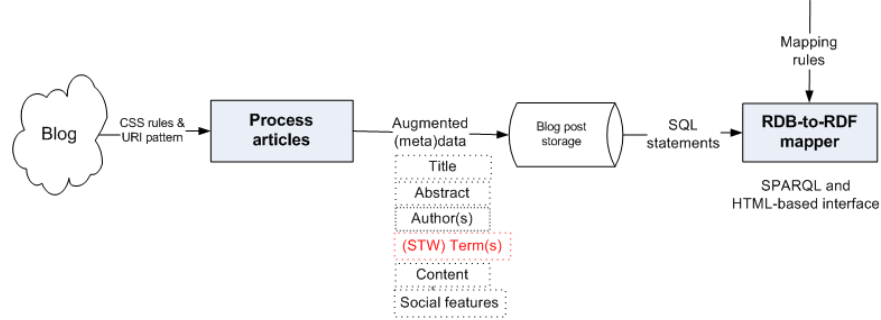


Fig. 2: Process workflow: from "raw" blog posts to access within DL

1. *Data set retrieval*: EconStor is already available as RDF and regularly updated; furthermore, it is also available for download (as RDF data dump). The WSJ blog post collection, on the other hand, needed some effort. Based on the URL pattern of the WSJ blogs, we retrieved all blog posts from the different categories that WSJ supports. Blog post collections are more dynamic in nature, as they get new posts and updates (comments, shares, etc.) to existing posts. We retrieved posts of a certain period, and that suits our experimentation purposes.

Potential extension for this step: Depending on the datasets of interest, the retrieval process could entail different approach, such as using an API that makes resources of interest easily available.

2. *(Pre)processing blog posts*: After retrieving a single blog post, we extracted elements of importance to our experiment (as described earlier in the paper) by relying on the CSS description that WSJ employs for the different post elements. Sometimes the author names are part of the post contents. We relied on NLP operations (Apache OpenNLP<sup>16</sup> in this case) to identify (and remove) author names from post content. A final operation for this process adds to the list of elements by (automatically) assigning STW thesaurus terms to every blog post; MAUI was configured to assign up to 5 terms from the STW to every post, and the result is stored in a RDB.

Just to afford the idea of this last operation, Table 1 lists few examples from the process of automatic term assignment with MAUI. The first column

<sup>16</sup> <https://opennlp.apache.org/>

stores the title of the blog post; with it, we can guess to some extent what the topic of the post is. The second column stores the terms (or tags) that are used to describe the post; these tags are specific to WSJ blogs, and are used by its authors and readers to find posts of interest. The third column represents the terms that MAUI automatically assigned to every blog by choosing the terms from the STW thesaurus.

Potential extension for this step: Any other pre-processing activity that one could need, such as: using a different (automatic) indexer and/or CV to assign terms to resources; applying different enrichment activity, such as linking a resource to an external KB, for example; using any NLP-based operation that the new domain/application could requirement; etc.

Table 1: Several blog posts as described by their tags and the STW thesaurus after automatic indexation

<b>Blog post title</b>	<b>Original terms</b>	<b>Auto-assigned terms from STW</b>
The Growing Scarcity of Series B Venture Rounds	Culture; Investors/Raising capital; Investor	Marketing; Market; Seed; Enterprise
Fueling the Next Generation of Innovation	Business model; Culture; Entrepreneur	Enterprise, Needs, Innovation, Fuel
Understand the Risks of Going Public Before You Ring the Bell	Business model; Culture; Investor	Risk; Listed company; Publication; IPO
Why Design Matters More than Moore	Culture; Customer acquisition; Investor; Weekend read	Designers; Engineering; Technology; Metalloid
Dont Fall in Love with Your First Idea	Business model; Culture; Entrepreneur; Getting to Product Market Fit	Emotion; Marketing; Marketing; Joint production

3. *Access and query:* The D2RQ platform<sup>17</sup> "wraps" a RDB and provides access to it as if it were a graph DB. In our case, since the blog post collection was stored in a RDB, we decided to use D2RQ. One prerequisite to this option is to provide a mapping file that D2RQ will use to map RDB columns to graph representation. This way, we can have a "direct" access to the blog post collection as if it were a graph, or generate a RDF data dump to be used later on. We choose the latter to have the complete blog post collection as RDF data set, later to be combined with the DL collection (already available as RDF).

<sup>17</sup> <http://d2rq.org/>

Potential extension for this step: The datasets could be directly represented as RDF, thus eliminating the requirement for any intermediary tools for access and query of datasets. For example, having the datasets stored as RDF in a triplestore solution, as opposed to "extending" RDB solution's feature set to provide SPARQL access. Moreover, in case other/new vocabularies are applicable for a different domain or a new application, one can always change the mapping rules in the D2RQ platform to accommodate for this requirement. In this case, as one would expect, changes introduced should be accounted for during the query process.

4. *Availability and storage of RDF data sets:* Now that both data sets are represented in the same way (RDF graphs), and use the same term vocabulary (STW thesaurus), we needed a way to integrate them in a single data set and demonstrate the use cases envisioned for this work. We choose Apache Jena Fuseki SPARQL server<sup>18</sup> for this purpose.

We loaded both data sets as a single data "unit", with 2 different named graphs in order to provide flexibility when implementing use cases, such as when required to access or query only one of the data sets. Such was the case in one of our use cases where we wanted to select relevant blog posts after the DL user has accessed a publication from the DL collection. In the use cases section you can see more on the usability of the separate graphs for the two data sets.

Potential extension for this step: Depending on the technical choices from the previous activities of the workflow, one can choose a solution for this step, such as graph, NoSQL, or any other storage solution that suits their needs.

5. *Vocabulary bridging: Any ontology alignment/matching?:* One of the key decisions planned in our methodology and implemented via our process workflow is the one regarding the challenge stemming from having to integrate different vocabularies, DL and blog posts in this case. For example, EconStor uses a thesaurus to describe and base information retrieval services on it, whereas the WSJ uses its own categories and terms to describe posts. Furthermore, even for blogs themselves, there is a difference in terms of vocabularies used. This makes it a real challenge to try to align or map vocabularies used in a DL with all the different vocabularies that could be employed by blogs.

This is where the automatic indexer comes into play: by using the STW thesaurus to automatically assign terms to all blog posts (currently, we are only using WSJ posts, but a DL will most likely include more blogs), we effectively bridge the terminology gap, thus eliminate the need to pursue other alternatives, such as ontology alignment or mapping between the the STW thesaurus, in our example, with the vocabularies adopted by the blogs.

Potential extension for this step: One can choose to develop a vocabulary that specifically covers the domain of blogs (as some of the related work

---

<sup>18</sup> <https://jena.apache.org/documentation/fuseki2/index.html>

presented in the paper have chosen to do), and then try the arsenal of approaches of ontology alignment/matching to bridge the vocabulary gap.

6. *SPARQLing for more insight*: One benefit from our methodology, especially the dataset representation in RDF, is the capability for more insightful queries. For example, one could be interested to know the female authors active with blog posts during a certain period of time on the topic of technology transfer; or some of the most commented (shared, or liked, etc.) blog posts during the financial crisis of 2008, and similar queries. The fact that it is the query language for Linked Data, opportunities for more perceptive querying are great.

The following SPARQL code listing shows an example that lists WSJ blog posts and EconStor publications related to the term "Technology transfer" published since 2014/2015. The code queries the named graphs – `econstor` and `wsj` – with the same query, and "combines" the results (note the UNION operand) into a single result list; namespace definitions are omitted for brevity.

Potential extension for this step: The query can be refined by specifying narrower or broader terms from the STW thesaurus, in case there are many or few results from the query; filtering results based on a number of parameters, such as comments, tweets, shares, etc.

Listing 1.1: A SPARQL example querying heterogeneous datasets

```

SELECT ?publ ?title
From Named <http://localhost:3030/EconstorWSJ/data/econstor>
From Named <http://localhost:3030/EconstorWSJ/data/wsj>
WHERE {{
  Graph ?g {
    ?publ a ?o;
    econstorDC:issued ?date;
    econstorDC:title ?title;
    econstorDC:keyword "Technology_transfer" .
    Values ?o {swc:Paper}
    Filter(?date >= "2015"^^xsd:gYear) }}
UNION {
  Graph ?g{
    ?publ a ?o;
    dcterms:created ?date;
    dcterms:title ?title .

    ?term a vocab:keywords;
    dc:subject "Technology_transfer";
    vocab:keywords_blog_post_postID ?publ .

    Values ?o {sioc:BlogPost}
    Filter(?date >= "2014-01-01"^^xsd:date) }}}

```

Our proposed methodology, and the process workflow implementation driven by it, can be adopted/adapted for similar work by other DLs that seek to integrate emerging scientific publications into their collections. In our case we dealt with a DL that relies on a domain-specific thesaurus, and plans to integrate blog posts from that same domain. However, new use cases and possibly completely new applications, thus new extensions to the workflow are more than encouraged to pursue by interested parties.

## 6 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.

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 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). While SPARQL can support insightful queries of the data, we would like to also mention at this point that the queries we show are constrained by the metadata in both dataset. For example, we could explore the contribution of female bloggers from a certain region, during the "financial crisis" period, for example, had he had the (meta)data in our datasets.

1. *Search across the "unified query space" (EconStor and WSJ datasets):* 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 see a result of publications from corresponding datasets (treating the datasets as if they are one source of information). The following listing shows the SPARQL query that implements this use case scenario.
2. *Retrieve relevant blog posts for a DL publication:* 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 onward. 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 "QA: 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's reading experience. 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.
3. *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 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:

- (a) 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.
- (b) 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.
- (c) 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.

## 7 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:

1. *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.
2. *Integration of heterogeneous collections in a 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.

3. *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.

## 8 Conclusion and Future work

In this section we present our final thoughts on the research endeavor, as well as present its follow up steps for the near future.

- *Summary of conclusions:* With our undertaking we have addressed a DL requirement for integrating non-library resources a scientific blog post collection and making it available to its users as a complementary or additional content in their search operations. In doing so, we have pre-processed the non-library resources in order to bring them up to par with 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.
- *Future work:* One of the follow up efforts is developing a prototype enabling evaluation scenarios with the users. Furthermore, we would also like to compare our algorithmic approach (i.e. for generating recommendations of blog posts) with state of the art similarity measures from Graph-based and conventional Machine Learning approaches. We are of view that this comparative evaluation will help us to improve the accuracy and recall of our generated recommendations. 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.

There is also work planned regarding the evaluation of our work. Some of the preliminary research questions are to be directed towards establishing the complementarity of blog posts to DL resources, with the underlying hypothesis being that blog posts add value to and provide the serendipity effect for DL users. The prototype should be able to answer these initial questions, as well as raise new questions for the final evaluation (handling information overload could be such a question, that could potentially hamper or cancel the role of blog posts).

- *Limitations:* During this phase, we are relying on SPARQL to explore the resulting EconStor and WSJ blog post datasets, although a typical DL user does not and should not have to have any knowledge of SPARQL or, at a more general level, Semantic Web technologies to access and use DL services. A solution involving a graphical user interface would enable a more comfortable environment for users, and enable them reap the benefits of our research without the higher technological barrier that Semantic Web technologies represent for common users.

## References

1. Auer, S., Feigenbaum, L., Miranker, D., Fogarolli, A., and Sequeda, J.: Use Cases and Requirements for Mapping Relational Databases to RDF. Working draft, The World Wide Web Consortium (W3C) (2010) <https://www.w3.org/TR/rdb2rdf-ucr/> Accessed: 14 June 2016
2. Burgelman, J-C, Osimo, D., and Bogdanowicz, M.: Science 2.0 (change will happen.). First Monday, 15(7) (2010). Accessed: 28 June 2016.
3. Holgersen, R., Preminger, M., and Massey, D.: Using semantic web technologies to collaboratively collect and share user-generated content in order to enrich the presentation of bibliographic records. Code4Lib Journal 17 (2012). <http://journal.code4lib.org/articles/6695> Accessed: 20 June 2016
4. Hu, Y., and Janowicz, K., and McKenzie, G., Sengupta, K., and Hitzler, P.: A Linked-Data-driven and Semantically-enabled Journal Portal for Scientometrics. 12th International Semantic Web Conference (2013). Springer Berlin Heidelberg
5. Kalb, H., Lazaridou, P., Trier, M.: Establishing Interoperability of a Blog Archive through Linked Open Data. In GI-Jahrestagung, pp. 1931-1936 (2013).
6. Latif, A., and Borst, T., and Tochtermann, K.: Exposing Data from an Open Access Repository for Economics As Linked Data. D-Lib Magazine Vol. 20, 9(10) (2014) <http://www.dlib.org/dlib/september14/latif/09latif.html> Accessed: 7 June 2016
7. Limani, F., and Latif, A., and Tochtermann, K.: Scientific Social Publications for Digital Libraries. 20th International Conference on Theory and Practice of Digital Libraries (2016)
8. Limani, F., Latif, A., and Tochtermann, K.: Bringing Scientific Blogs to Digital Libraries. Proceedings of the 13th International Conference on Web Information Systems and Technologies (WebIST), Porto, Portugal, April 25-27, pp. 284–290 (2017), DOI: 10.5220/0006295702840290
9. Mahrt, M. and Puschmann, C.: Science blogging: an exploratory study of motives, styles, and audience reactions. Journal of Science Communication 13(3) (2014)

10. Powell, J., and Collins, L., and Martinez, M.: Semantically Enhancing Collections of Library and Non-library Content. *D-Lib Magazine* Vol. 16, 7(8) (2010) <http://www.dlib.org/dlib/july10/powell/07powell.html> Accessed: 10 June 2016
11. Passant, A., and Laublet, P., Breslin, G., and Decker, S.: *SemSLATES: Weaving Enterprise 2.0 into the Semantic Web* (2010)
12. Papadokostaki, K., Charitakis, S., Vavoulas, G., Panou, S., Piperaki, P., Papakonstantinou, A., ... Kondylakis, H.: *News Articles Platform: Semantic Tools and Services for Aggregating and Exploring News Articles*. *International Conference on Integrated Information* (2016)
13. Spanos, D.-E., Stavrou, P., and Mitrou, N.: *Bringing Relational Databases into the Semantic Web: A Survey*. *Semantic Web* 3(2), 169–209 (2012)
14. Yoose, B., and Perkins, J.: *The LOD landscape in libraries and beyond*. *Journal of Library Metadata* Vol. 13, 2(3), 197–211 (2013)