Latif, Atif; Scherp, Ansgar; Tochtermann, Klaus

Article

LOD for Library Science: Benefits of Applying Linked Open Data in the Digital Library Setting

KI - Künstliche Intelligenz

Suggested Citation: Latif, Atif; Scherp, Ansgar; Tochtermann, Klaus (2015) : LOD for Library Science: Benefits of Applying Linked Open Data in the Digital Library Setting, KI - Künstliche Intelligenz, ISSN 1610-1987, Springer, Berlin, Vol. 30, Iss. 2, pp. 1-9, http://dx.doi.org/10.1007/s13218-015-0420-x

This Version is available at:
http://hdl.handle.net/11108/225

LOD for Library Science: Benefits of Applying Linked Open Data in the Digital Library Setting

Retrospects and Research Topics

Atif Latif · Ansgar Scherp · Klaus Tochtermann

Abstract Linked Open Data (LOD) has gained widespread adoption by large industries as well as non-profit organizations and governmental organizations. One of the early adopters of LOD technologies are libraries. Since the "early years", libraries have been key use case and innovation driver for LOD and significantly contributed to the adoption of semantic technologies. The first part of this paper presents selected success stories of current activities in the Linked Data Library community. In a nutshell, these studies include i) a conceptualization of the Linked Data Value chain, ii) a case study for consumption of Linked Data in a digital journal environment, and iii) an approach to publish metadata on the Semantic Web from an Open Access repository. These stories reveal a strong relationship between LOD in libraries and research topics addressed in traditional fields of computer science such as artificial intelligence, databases, and knowledge discovery. Thus, in the second part of this paper we systematically review the relation of LOD in digital libraries from a computer science perspective. We discuss current LOD research topics such as data integration and schema integration, distributed data management, and others. These challenges have been discussed with computer scientists at a German national database meetup as well as with librarians from ZBW - Leibniz Information Center for Economics and at international librarians meetup.

ZBW - Leibniz Information Center for Economics
Kiel, Germany
E-mail: a.latif@zbw.eu · a.scherp@zbw.eu · k.tochtermann@zbw.eu

Keywords Digital Libraries · Semantic Web · Linked Open Data

1 Introduction

Digital Libraries are facing up to the challenge of providing access to the huge amounts of data which is hidden, inaccessible, and stored in the data silos since a long period of time. With key developments of web technologies in promoting easy access to heterogeneous data, it is getting more relevant for digital libraries to find optimal ways for the publishing of wall gardened metadata which will subsequently permit library collections to be discovered, linked and accessed by related libraries in a sustainable manner. On the other hand, Linked Open Data [10] provides the best practices in publishing and sharing of information by use of semantic technologies and gives access to a large amount of heterogeneous data which presents exciting opportunities for the application development like: applications on seem-less data integration which in particular aggregate information from multiple data-sources to provide a coherent data view. Interestingly, Linked Data guidelines [1] can help digital libraries to get rid of data silos by publishing their dataset as structured data; and they can bring a lot of value to the libraries. However, for Linked Data to get recognition within library community, it is critical to highlight the Linked Data application value with the help of developed system case studies.

In the first part of this paper, we will retrospect on systems and technologies we have developed and applied to put up the strong case of Linked Data usage in the library community. The scientific impact of these systems and technologies has been published in various research conferences and journals. The major work covers the following three success stories which are extensively explained in Section 2 of the paper: (i) We developed a model to identify the implicit actors which are involved in the Linked Data generation process [18]. This work led us to better conceptualize the dynamics within the Linked Data value generation cycle and how it may impact Linked Data uptake at various levels, e. g., organizational and commercial. (ii) In order to understand Linked Data consumption we developed the "Keyword to URI" technique to retrieve the user queried keyword information from the Linked Data cloud[1] by successfully hiding the complexities of semantic querying [15]. (iii) After understanding the basic mechanics of Linked Data and its links to digital libraries, in our third story we describe how we published our open access repository metadata as linked data[2] and started developing systems helping us to connect to other digital libraries [17].

Following the presentation of the three selected success stories, the paper will discuss general research topics where computer scientists, particularly in the traditional fields of artificial intelligence, databases, and knowledge discovery, can contribute to library sciences. We identify six research topics that surprisingly are well known in the above mentioned fields but where library sciences, in particular in the context of Linked Open Data, brings in a new shed of light and poses interesting research challenges. The research topics are put into the context of the three success stories. In addition, they have been discussed in large groups of researchers and practitioners, both within the computer science community as well as the library science community.

The remainder of the paper is structured as follows: In Section 2, we present the success stories and explain how these studies are related to digital libraries and what are their contributions. Section 3 describes in detail the general research topics which arise when dealing with LOD in digital libraries. We stress how these challenges of LOD in digital libraries overlap with research topics in traditional computer science fields, before we conclude the paper.

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

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

2 Success Stories

In this section, we present the success stories of LOD in digital libraries in detail. In general, these stories will depict the varying needs of digital libraries with respect to information supply and will summarize how maturation of Linked Data technologies and systems have happened to counter these needs. Moreover, with the help of established system case studies, this section will also categorically highlight the major benefits which digital libraries can harness by applying the Linked Data technologies.

2.1 Linked Data Value Chain

Since its inception, the Linked Data project has been facilitating the transformation of publicly available Open Data to Linked Data. Still by now the vast majority of data is being generated by research communities and commercial uptake of Linked Data is catching up. From a corporate and business perspective it is very important to conceptualize the Linked Data generation cycle. We introduce the Linked Data Value Chain as a lightweight model for business engineers to support the conceptualization of successful business cases [18]. Thereby, we identified three main concepts as illustrated in Figure 1: Different Entities acting in different roles (i. e., *Raw Data Provider*, *Linked Data Provider*, *Linked Data Application Provider* and *End Users*), both consuming and providing different Types of Data (i. e., *Raw Data*, *Linked Data*, and *Human Readable Data*).

We established that the assignment of roles to these entities, the combination and involvement of roles, the data selected as well as the data transformation process may hold inherent risks. Moreover, we proposed two main areas where pitfalls may arise and grouped them into *Role-Related Pitfalls* and *Data-Related Pitfalls*. In a nutshell, Role-Related Pitfalls are either related to individual roles or to the interaction of different roles, e. g., usage rights, privacy policies, data availability, and role incentives. Whereas Data-Related Pitfalls are either related to the data itself or the data transformation process, e. g., data quality and trust, data

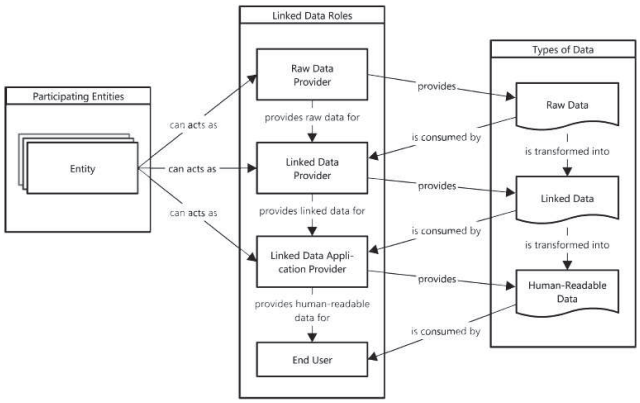

Fig. 1 Linked Data Value Chain (adopted from [18])

provenance, transparent data transformation and interlinking.

For demonstration purpose, we applied the Linked Data Value Chain to an existing business case from the BBC[3] (a pioneer adopting to Linked Data technologies) and highlighted the potential pitfalls along the way. Overall, the Linked Data Value Chain helped to identify and categorize potential pitfalls which have to be considered by business engineers and furthermore, has led us in establishing the ways to a clear understanding of the complete Linked Data generation cycle. This model is easily mappable to other disciplines, e. g., digital libraries, life sciences, and media. It will help in better planning for Linked Data publishing along with clear indication of potential research challenges which may arise during data conversion and data interlinking as covered in Section 3.

2.2 Author Profiles

This study was conducted to highlight the added value which can be drawn by consuming Linked Data for a real world application, i. e., a profiling systems in digital journal environment. In this study, we pointed out the challenges of author name recognition and disambiguation, which we faced during processing of person related information [15]. A profiling system tends to find information about persons on some individual particulars i.e. expertise, influence in social media, and the number of publications. For instance, a well connected digital journal can play a vital role in creating opportunities for collaborations between organizations, institutions, and persons. However, finding correct information about authors (author profiles) is crucial to increase the overall visibility, efficiency, and unprecedented success of a digital journal.

Building on the LOD initiative, we have developed a proof of concept application CAF-SIAL (Concept Aggregation Framework for Structuring Informational Aspects of Linked Open Data)[4]. It can discover and present informational aspects of persons from Linked Data. CAF-SIAL identifies a person's relevant information from DBpedia[5] by employing a set of heuristics, which is extracted by applying a "Keyword to URI" technique [16]. This extracted information is further filtered and integrated with the help of a Concept Aggregation Framework [19] which subsequently is presented as a profile.

To showcase the application utility in the library setting, it was further extended to establish the links between authors of the digital journal with relevant semantic resources from LOD, i. e., DBpedia and DBLP[6]. The underlying approach of this application was able to identify, disambiguate, retrieve and structure relevant information about an author from these data sets. As a final output, the system constructed a comprehensive aspect-oriented author's profile and was helpful in giving insights of authors biography (personal and professional information) and lists his published works. This system was further implemented and integrated with the "Links into Future" feature of the Journal of Universal Computer Science (J.UCS)[7]. For instance, the profile of an author Gio Wiederhold who has published a paper in J.UCS can be viewed by following this link: `http://goo.gl/tJFtgI`.

From our point of view these kinds of systems can be easily re-produced in the broader scholarly communication domain, i. e., open access repositories and subject portals. Moreover, the corpus of search can be further extended to integrated authority files like the Integrated Authority File of the German National Library (GND)[8] and Virtual International Authority File (VIAF)[9] for more and complete results. In general, authority files comprise controlled keywords and descriptors, which are assigned to a publication during the cataloging process. It's goal is to further simplifying

[3] `http://bbc.co.uk`

[4] `http://cafsial.lod-mania.com`

[5] `http://www.dbpedia.org/`

[6] `http://dblp.l3s.de/d2r/`

[7] `http://www.jucs.org/`

[8] `http://www.dnb.de/EN/gnd`

[9] `https://viaf.org/`

the search and retrieval process. In contrast, a name authority file is an authority file for persons.

2.3 Linked Data Publishing - EconStor

In last few years, Open Access repositories have contributed heavily towards the success of Open Data and have become one of the most prominent types of library applications. These repositories are systems for collecting, publishing, disseminating, and archiving digital scientific content. With respect to Open Access publishing, repositories nowadays serve as platforms for acquiring and disseminating scientific content, which before had been almost exclusively released by commercial publishers. Citing the importance of Open Access repositories to today's digital libraries and to provide metadata of scientific working papers from the repository in a machine readable fashion, we published our Open Access repository EconStor[10] as Linked Data [17]. EconStor is the Open Access server of ZBW - the German National Library of Economics and provides a platform for publishing working papers in economics. EconStor currently provides access to working papers from approximately 100 institutions as well as full text access to more than 80,000 full text papers.

In this study, we provided both conceptual and practical insight into the process of converting a legacy relational dataset to machine understandable semantic statements, a.k.a. 'triplification', and provided an overview of the D2RQ framework[11] that can be used for this purpose. Publishing of EconStor repository data as Linked Data is illustrated with with help of the system architecture as depicted in Figure 2. In a first step, repository data as a relational database was acquired. In a second step, major resources within the repository, i. e., communities, collections, and items (publications and authors) were mapped into a D2R Server mapping file by re-using popular vocabularies, i.e., Dublin Core (DC)[12], Friend of a Friend (FOAF)[13], and Semantic Web Conference Ontology (SWC)[14]. In the last step, repository data was transformed by using the D2R

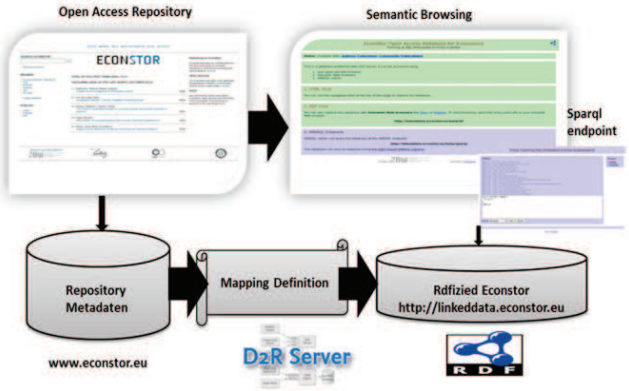

Fig. 2 System Architecture of EconStor Publishing as Linked Data (adopted from [17])

Server and made available as Linked Data along with a SPARQL endpoint for querying[15].

One important outcome of this effort was that a repository's content can be straightforwardly published as Linked Open Data. Another result was the ability to link to valuable external datasets which enabled a repository's data to become more contextualized and 'meaningful'. Below, we discuss the envisioned goals which we have achieved after publishing the EconStor Open Access repository as Linked Data[16]:

- Published scientific working papers into the Semantic Web and thereby supported the publishing process and dissemination of current research results in Economics.
- Successfully opened up content from typical repository systems like DSpace for the Semantic Web and integrated it into the mainstream of Linked Data by data publishing.
- Created new possibilities for querying distributed research information over the EconStor dataset in form of SPARQL queries. For example, a query can be "Show me all articles on 'Financial Crisis', which have been published by European research institutes after 2012".
- The publishing of EconStor as Linked Data has created the potential for the development of mashup applications which can curate data from different relevant Linked Data stores.

From both practical and software engineering points of view, this study describes an approach towards pub-

[10] http://econstor.eu
[11] http://d2rq.org/
[12] http://purl.org/dc/terms/
[13] http://xmlns.com/foaf/0.1/
[14] http://data.semanticweb.org/ns/swc/ontology#
[15] http://linkeddata.econstor.eu/beta/snorql/
[16] http://linkeddata.econstor.eu/beta/

lishing a repository's content as Linked Open Data. This can be of great interest to librarians, repository managers, and software developers who work for libraries.

3 Research Topics of LOD in Library Sciences

Motivated by the success stories of Linked Open Data in library sciences documented above, the authors set back and thought about what are the general research topics for computer science when dealing with LOD in digital libraries. Following the Linked Data value data chain described in Section 2.1, there emerged six research topics for LOD in library sciences. Interestingly, all these topics refer to challenges in traditional areas of computer science such as artificial intelligence, databases, and knowledge discovery.

In summary, the six research questions are: First, regarding the conversion of "raw" data such as semi-structured metadata of scientific publications (author names, titles, publishers provided as strings), we consider the research topic of *Entity Resolution*. This topic was addressed, e. g., by the success story of author profiles described in Section 2.2. Second, related with entity resolution is the topic of *Schema Matching*. Third, regarding the enhancement of content, we find the topic of *Automated Indexing*. Unlike indexing in the database community, the topic of automated indexing in this context refers to the machine learning task of multi-label classification such as assigning a set of descriptors obtained from a thesaurus to a scientific publication. Fourth, while predominantly the task of automated indexing in library sciences is applied to text content, e. g., PDF documents, it is increasingly also important to automatically *Indexing Non-textual Content* such as social media and audio-visual material. Fifth, referring to the phase of consumption in the Linked Data value data chain outlined in Section 2.1, we find the challenges of *Distributed Data Management* that targets to retrieve and aggregate different information such as bibliographic records from various data sources. Finally, when merging data from different sources, such as query results but also results from automated indexing services, it is essential to track the provenance of the data to show the users where certain information comes from and to demonstrate its trustworthiness.

This list of research topics was discussed with the computer science community in form of an invited talk at a meet-up of the German database community[17]. In addition, the topics were also discussed in a 90 minutes session with about 50 domain experts in the broader context of library sciences during an interactive presentation at ZBW in August 2014. Finally, the topics were also presented and discussed in a talk at the 104th German National library meetup on May 26th, 2015.[18] Below, we briefly describe and discuss the nature of the different research topics along the structure outlined above. In particular, we highlight the role of Linked Open Data that brings in a new shed of light and poses interesting research questions to computer science.

3.1 Entity Resolution

Entity resolution refers to the problem of identifying whether two resources of Linked Open Data refer to the same real world entity. This is a challenging task, as the resources do not have any identity of their own but the semantics is only defined by considering the properties that are used to describe and connect the resources [7]. One approach to deal with the problem is manual alignment as it is conducted in the Linked Open Data project of the German National Library[19]. The Integrated Authority File of the German National Library contains among others information about the authors that published in Germany and is connected with DBpedia and others. Here, the challenge is to discriminate famous authors like the former German chancellor Helmut Kohl (available through: `http://d-nb.info/gnd/118564595`) from his namesake who also publishes.

However, manual alignment is very expensive and not possible when merging large datasets. For example, in the context of library sciences there are data sources providing information about persons of sizes ranging from 364,000 in DBpedia, 1,797,911 in the German National Library Authority File, 3,800,000 in the Library of Congress, and 10 million in the Virtual International Authority File[20] (VIAF) [23]. VIAF combines multiple name authority files of different national libraries. However, a particular problem here is that en-

[17] "Databases meets LOD", Frühjahrstreffen der GI Fachgruppe Datenbanken, March 20-21, 2014, Brunswick, Germany. URL: `http://www.ifis.cs.tu-bs.de/node/2911`

[18] 104. Bibliothekartag, May 26-29, 2015, Nuremberg, Germany, URL: `http://www.bibliothekartag2015.de/`

[19] `http://www.d-nb.de/`

[20] `https://viaf.org/`

tity resolution over name, co-authors, title, and venue is often not sufficient [11].

3.2 Schema Matching

Ontology matching [3] or more general schema matching [26] is similar to the challenge of entity resolution and refers to the question of data integration. The goal of Linked Open Data is to define and publish vocabularies that become self-descriptive by referring to definitions of concepts and properties of other existing vocabularies. However, the integration of different vocabularies and thus the data they describe is far from trivial, even for databases with similar schemata [26]. For example, the property foaf:name of the famous Friend-of-a-Friend (FOAF) vocabulary[21] is quite similar to vcard:family-name of the vCard ontology[22]. However, the foaf:name property is more general and can take more than just the surname as in the case of the vcard:family-name property.

While schema integration is desired to improve library services, at the same time the library science community demands a very high quality in schema matching. As a consequence, different works on schema matching in library science have been carried out in the past by manually aligning thesauri. For example, ZBW's thesaurus for economics STW[23] is aligned with other thesauri such as TheSoz[24] in social sciences where a couple of thousand manually created mappings were build in 2004 to 2005. For describing the mappings, the relations between keywords are typically represented using the Simple Knowledge Organisation System[25] (SKOS) vocabulary. For example, related keywords are expressed using skos:related.

Due to the size of the thesauri of usually a couple of thousand or even ten-thousand descriptors plus corresponding synonyms, automated approaches for schema matching are needed. Thus, since 2012 there is a Library Track on ontology matching in the Ontology Alignment Evaluation Initiative[26] (OAEI). The OAEI aims to compare different schema matching techniques and to establish a consensus regarding the evaluation of methods for ontology matching.

[21] http://xmlns.com/foaf/spec/
[22] http://www.w3.org/TR/vcard-rdf/
[23] http://zbw.eu/stw/versions/latest/about
[24] http://lod.gesis.org/pubby/page/thesoz/
[25] http://www.w3.org/2004/02/skos/
[26] http://oaei.ontologymatching.org/

3.3 Distributed Data Management

Particular characteristics of Linked Open Data is its highly distributed fashion of publishing and interlinking data on the web. The VIAF is a good example where a couple dozen international organisations collaborate in building a distributed network of library resources, not only traditional records (i. e., publications) but also persons and organisations. A central data storage and search over this data is neither a desired nor a feasible solution. To access the data published in such a highly distributed fashion, technologies for federated querying are needed and index-structures that store information about which data is provided by which source.

In the past, the Semantic Web community has developed various different techniques for distributed querying of Linked Open Data [4,9,8] as well as stream-processing of Linked Open Data for providing a lookup service what data is provided by which source [14,5]. However, so far it is not clear which approach is the most appropriate for accessing the distributed data, e. g., a distributed set of (SPARQL-based) endpoints [4] vs. traversal-based querying of Linked Open Data [9].

In addition, one needs to think about ranking of results in order to accommodate the user expectations when using library search services. Like in web search, users of library search services consider the first hits implicitly more important and relevant than others. To address this challenge, the DFG project LibRank[27] at ZBW investigates the integration of journal rankings in the computation of the search result order.

3.4 Automated Indexing

In contrast to the notion of indexing in the database community, indexing in library sciences refers to the task of selecting multiple labels for the classification of documents such as scientific publications. One approach for indexing is the manual labeling of scientific publications that has been conducted by library scientists of ZBW for more than 1.6 million economic documents in the past using the STW. On average, each

[27] http://www.librank.info/

of the 1.6 million scientific publications has been annotated with five STW descriptors. Another example is the publication server EconStor (see Section 2.3), which performs auto-completions on author keywords regarding STW and other thesauri. Here, the author confirms a keyword by selecting a suggested STW term, i. e., the keyword is matched with the semantic concept. A particular challenge here is the quality of the provided annotations, which is, unlike the annotations by library scientists, of lower quality and does not necessarily refers to a semantic concept from the STW.

In addition, the tremendously increasing amount of electronically delivered publications per year (compared to print publications which remains stable) at the German National Library shows that automated approaches for indexing literature are needed [21]. In the past, automated approaches for the classification of PDFs have been developed. For example, the PETRUS project at the German National Library uses Support Vector Machines to classify documents along 100 classes (so called "Sachgruppen"). However, the automated indexing of scientific literature has already been investigated by libraries since a long time ago. For example, the DFG-funded project GERHARD in the nineties investigated methods for automatically indexing scientific Web content.

About 1 million documents were crawled and automatically indexed using about 10.000 hierarchically organized concepts from the Universal Decimal Classification (UDC) [22] system. The indexing was performed using the UDC in three language (German, English, French). The infrastructure was a single server machine using the Oracle relational database management system with full-text indices ConText (today: Oracle Text). At the time being, the GERHARD project of course neither use Semantic Web technologies nor refer to data published on the web for the annotations. Despite these early developments, the automated indexing of scientific literature remains a very active research field until today.

A promising work towards the automated indexing of scientific documents using Linked Open Data is developed in a recent ZBW project for multi-labeling scientific documents using the STW. It uses the kNN classifier in combination with entity detection and the HITS algorithm [13] for assessing the importance of STW concepts for a specific document [6]. Experiments over a large corpus of about 62,000 open access documents from ZBW's EconBiz[28] literature search portal showed an average recall of .40 (SD:.32) and an average precision of .40 (SD:.32), resulting in a F-measure of .39 (SD:.31). By this, the technique outperforms today's approaches for multi-labeling such as Maui [20] using decision trees (average F-measure of .36 on the same dataset [24]). The solution developed at ZBW is to the best of our knowledge the by far largest experiment for automated indexing carried out. In addition, it has the advantage that is does not require expensive training phases that are required with approaches like Support Vector Machines where samples need to be manually selected in order to train the machine classifier [24].

Please note, although the term "automated indexing" denotes that there is no human in the loop, the above mentioned technique is not designed to be run without human intervention. In fact, the expertise of library scientists is needed to constantly monitor the quality of the automatically suggested descriptors and adapt the thesauri like the STW to reflect new trends and topics. Thus, a particular challenge besides the multi-labeling task itself is the integration and use of the machine learning results in the context of a real-world application and the organizational integration.

3.5 Indexing Non-textual Content

Besides textual content such as scientific publications provided as PDF and websites relevant for being indexed by libraries, there is also a large amount of non-textual content such as social media and audio-visual material. Specific challenges here are the mapping of traditional scientific content with social media but also with research data, which is addressed by ZBW in the EU project EEXCESS[29]. The idea is to automatically combine structured scientific content (metadata, full-texts, paragraphs, citations, and others) with informal and hasty content from social media channels in order to link topics, objects (the textual and non-textual resources), and the users. Challenges are entity resolution and indexing over multiple modalities, but also the cross-media retrieval of content.

In order to address the challenge of multimodal retrieval, we developed a novel pipeline for better understanding information graphics that are typically contained in scientific publications. The pipeline allows

[28] http://www.econbiz.de/

[29] http://eexcess.eu/

for the automated extraction of multi-oriented text elements from information graphics by a novel combination of different methods from data mining and computer vision [2]. This allows for textual search over the information graphics and combining it with the textual content of the scientific publications.

3.6 Data Provenance

The Virtual International Authority File (VIAF) mentioned above aims to facilitate an inter-organizational and cross-border and thus cross-lingual linkage of bibliographic records. The goal is to lower costs and increase utility of library authority files by matching and linking widely-used authority files and making these links available on the Web. However, in such a multi-national setting particular challenges arise:

- How to track data/metadata (re)use?
- How to refer to original data/metadata when library A uses a (part of) record from library B?
- How to assess the trustworthiness of data/metadata incorporated into one's systems?

In order to (partially) address these challenges of provenance, the library sciences community developed in the past sophisticated models for describing library resources. The Functional Requirements for Bibliographic Records (FRBR)[30] is a quite powerful model to describe different variants of the same library resource, e. g., different prints and translations of the same book into different languages. Thus, it is not only applicable to books but to any kinds of resources. The concepts of FRBR are incorporated in the new cataloging code Resource Description and Access (RDA)[31] to describe any kind of content, including online media. RDA also allows to attach provenance information to the different concepts. The data model of Europeana[32] foresees the attachment of provenance to the metadata, i. e., who created the metadata record, as well as the provenance of the resource itself, e. g., Leonardo da Vinci as painter of Mona Lisa. A standard for describing the provenance of data on the web is the W3C PROV ontology[33].

However, what is still missing is an approach to reliably verify the provenance of metadata published as Linked Open Data on the web. One very promising first approach to track reuse of metadata is the framework for digitally signing graph data developed by Kasten et al. [12]. It allows to sign arbitrary graph data such as Linked Open Data by attaching a digital signature to it and publishing the data together with the signature on the web. This allows to track the provenance of metadata and build a "network of trust".

In addition, provenance-aware applications are missing that make use of such information. Applications like the semantic search engine Sig.ma [25] were capable of providing support for entity search over Linked Open Data and filtering results based on the provenance. Unfortunately, the project was discontinued. Search engines such as Sig.ma may, proof very valuable in the international setting of searching for relevant (scientific) literature and information (including social media channels) from diverse and distributed sources on the web based on provenance information.

4 Conclusion

Linked Library Data can be seen as innovation driver and libraries as early adopters of Semantic Web technologies. In this paper, we have presented selected success stories of LOD in library sciences. At the same time, we reflected on different topics and challenges that are relevant for computer scientists and that can be well motivated from library sciences.

At ZBW - Leibniz Information Center for Economics, we are addressing these challenges of LOD in library sciences not only from a technological perspective as described by the success stories in Section 2 and the challenges in Section 3, but in an interdisciplinary setting [24]. For example, in the project on automated indexing [6,2] an interdisciplinary team of domain experts collaborates with computer scientists where new research ideas are discussed and reflected from a practitioners' perspective in order to achieve both high-quality research outcomes as well as improved services for ZBW's customers.

Please note, we focused in this article on discussing technological success stories and research topics of LOD in libraries. Out of scope here (but equally important) are data quality management (e. g., for automated indexing), legal aspects of text and data mining, as well as educating data scientists and the job market for LOD in libraries.

[30] http://www.ifla.org/publications/functional-requirements-for-bibliographic-records
[31] http://www.rda-jsc.org/rda.html
[32] http://www.europeana.eu/portal/
[33] http://www.w3.org/TR/prov-o/

References

1. Berners-Lee, T.: Linked-data design issues. W3C design issue document (2009). Http://www.w3.org/DesignIssue/LinkedData.html
2. Böschen, F., Scherp, A.: Multi-oriented text extraction from information graphics. In: Symposium on Document Engineering (DocEng); Lausanne, Schwitzerland. ACM (2015)
3. Euzenat, J., Shvaiko, P.: Ontology Matching, Second Edition. Springer (2013)
4. Görlitz, O., Staab, S.: Federated data management and query optimization for linked open data. In: A. Vakali, L.C. Jain (eds.) New Directions in Web Data Management 1, *Studies in Computational Intelligence*, vol. 331, pp. 109–137 (2011)
5. Gottron, T., Scherp, A., Krayer, B., Peters, A.: Lodatio: using a schema-level index to support users infinding relevant sources of linked data. In: V.R. Benjamins, M. d'Aquin, A. Gordon (eds.) Proceedings of the 7th International Conference on Knowledge Capture, K-CAP 2013, Banff, Canada, June 23-26, 2013, pp. 105–108. ACM (2013)
6. Große-Bölting, G., Nishioka, C., Scherp, A.: A comparison of different strategies for automated semantic document annotation. In: International Conference on Knowledge Capture (KCAP); Palisades, NY, USA. ACM (2015)
7. Halpin, H., Presutti, V.: An ontology of resources: Solving the identity crisis. In: European Semantic Web Conference, pp. 521–534 (2009)
8. Harth, A., Umbrich, J., Hogan, A., Decker, S.: YARS2: A federated repository for querying graph structured data from the web. In: K. Aberer, K. Choi, N.F. Noy, D. Allemang, K. Lee, L.J.B. Nixon, J. Golbeck, P. Mika, D. Maynard, R. Mizoguchi, G. Schreiber, P. Cudré-Mauroux (eds.) The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007., *Lecture Notes in Computer Science*, vol. 4825, pp. 211–224. Springer (2007)
9. Hartig, O., Bizer, C., Freytag, J.C.: Executing SPARQL queries over the web of linked data. In: A. Bernstein, D.R. Karger, T. Heath, L. Feigenbaum, D. Maynard, E. Motta, K. Thirunarayan (eds.) The Semantic Web - ISWC 2009, 8th International Semantic Web Conference, ISWC 2009, Chantilly, VA, USA, October 25-29, 2009. Proceedings, *Lecture Notes in Computer Science*, vol. 5823, pp. 293–309. Springer (2009)
10. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space, 1st edn. Morgan & Claypool (2011)
11. Kanani, P., McCallum, A., Pal, C.: Improving author coreference by resource-bounded information gathering from the web. In: Conference on Artifical intelligence, pp. 429–434. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2007)
12. Kasten, A., Scherp, A., Schauß, P.: A framework for iterative signing of graph data on the web. In: V. Presutti, C. d'Amato, F. Gandon, M. d'Aquin, S. Staab, A. Tordai (eds.) The Semantic Web: Trends and Challenges - 11th International Conference, ESWC 2014, Anissaras, Crete, Greece, May 25-29, 2014. Proceedings, *Lecture Notes in Computer Science*, vol. 8465, pp. 146–160. Springer (2014)
13. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. J. ACM **46**(5), 604–632 (1999)
14. Konrath, M., Gottron, T., Staab, S., Scherp, A.: Schemex - efficient construction of a data catalogue by stream-based indexing of linked data. J. Web Sem. **16**, 52–58 (2012)
15. Latif, A., Afzal, M.T., Helic, D., Tochtermann, K., Maurer, H.A.: Discovery and construction of authors' profile from linked data (A case study for open digital journal). In: C. Bizer, T. Heath, T. Berners-Lee, M. Hausenblas (eds.) Proceedings of the WWW2010 Workshop on Linked Data on the Web, LDOW 2010, Raleigh, USA, April 27, 2010, *CEUR Workshop Proceedings*, vol. 628. CEUR-WS.org (2010)
16. Latif, A., Afzal, M.T., Höfler, P., Saeed, A.U., Tochtermann, K.: Turning keywords into uris: simplified user interfaces for exploring linked data. In: Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human 2009, Seoul, Korea, 24-26 November 2009, pp. 76–81 (2009)
17. Latif, A., Borst, T., Tochtermann, K.: Exposing data from an open access repository for economics as linked data. D-Lib Magazine **20**(9/10) (2014)
18. Latif, A., Saeed, A.U., Höfler, P., Stocker, A., Wagner, C.: The linked data value chain: A lightweight model for business engineers. In: A. Paschke, H. Weigand, W. Behrendt, K. Tochtermann, T. Pellegrini (eds.) 5th International Conference on Semantic Systems, Graz, Austria, September 2-4, 2009. Proceedings, pp. 568–575. Verlag der Technischen Universität Graz (2009)
19. Latif, A., Saeed, A.U., Höfler, P., Tochtermann, K., Afzal, M.T.: Harvesting pertinent resources from linked open data. JDIM **8**(3), 205 (2010)
20. Medelyan, O., Frank, E., Witten, I.H.: Human-competitive tagging using automatic keyphrase extraction. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, 6-7 August 2009, Singapore, A meeting of SIGDAT, a Special Interest Group of the ACL, pp. 1318–1327. ACL (2009)
21. Mödden, E.: Zukunftsfähige Inhaltserschließung Strategien und Perspektiven in der Deutschen Nationalbibliothek (2013). URL: http://www.dnb.de/SharedDocs/Downloads/DE/DNB/wir/petrus/petrusZukunftsfaehigeInhaltserschlie%C3-%9Fung.pdf
22. Möller, G., Carstensen, K.U., Diekmann, B.: GERHARD (German Harvest Automated Retrieval and Directory). KI - Künstliche Intelligenz (2000)
23. Neubert, J., Tochtermann, K.: Linked Library Data: Offering a Backbone for the Semantic Web. Communications in Computer and Information Science Volume 295 (2012)
24. Peters, I., Scherp, A., Tochtermann, K.: Science 2.0 and Libraries: Convergence of two sides of the same coin at ZBW Leibniz Information Centre for Economics. IEEE STCSN E-Letter on Science 2.0 **3**(1) (2015). URL: http://stcsn.ieee.net/e-letter/stcsn-e-letter-vol-3-no-1

25. Tummarello, G., Cyganiak, R., Catasta, M., Danielczyk, S., Delbru, R., Decker, S.: Sig.ma: Live views on the web of data. J. Web Sem. **8**(4), 355–364 (2010)
26. Wick, M.L., Rohanimanesh, K., Schultz, K., McCallum, A.: A unified approach for schema matching, coreference and canonicalization. In: KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 722–730. ACM, New York, NY, USA (2008)