LOD Technology in German Libraries and Archives: Postprint

Authors: Dong Jie

Date: 2017-12-25T00:00:00+00:00

Abstract

[Purpose/Significance] Linked Open Data (LOD) has been widely applied across various industries, non-profit organizations, and government agencies. Libraries and archives are among the early adopters of LOD technology, which has in turn facilitated the advancement of LOD itself. Germany, possessing a highly developed library and archive sector, offers numerous successful implementations of LOD within these institutions. [Method/Process] This study employs literature review, web-based investigation, and content analysis to examine successful cases of LOD technology deployment in German libraries and archives. [Results/Conclusion] The analyzed cases illuminate the interconnections among research topics in computer science domains, including artificial intelligence, databases, and library and archive studies. The practical experiences gleaned from Germany are synthesized to provide valuable references for developing analogous practices in our country.

Full Text

Application of LOD Technology in German Libraries and Archives

Dong Jie
Library of Harbin University of Commerce, Harbin 150028, China
ORCID: 0000-0001-5758-0139
E-mail: 124348423@qq.com

Abstract

[Purpose/Significance] Linked Open Data (LOD) has been widely adopted across numerous industries, non-profit organizations, and government agencies. Libraries and archives represent early adopters of LOD technology and have actively promoted its development. Germany stands as a developed nation in

the library and archive sector, with many successful cases demonstrating LOD applications in libraries and archives. [**Method/Process**] This study employs literature investigation, network survey, and content analysis methods to examine successful LOD application cases in German libraries and archives. [**Result/Conclusion**] These cases reveal the interconnections among research topics spanning computer science domains—such as artificial intelligence, databases, and knowledge discovery—and library/archive science research. Summarizing the characteristics and practical experience of German implementations can provide valuable references for developing relevant practices in China.

Keywords: Linked Open Data; LOD; Germany; library; archives; application
Classification: G250

Fund Projects: This research is supported by the National Natural Science Foundation of China "Control Strategy Research on Algae-Derived Organic Matter Removal by Adsorption/Membrane Method Based on Adsorption Mechanism and Membrane Fouling Control" (Project No. 51408169), the Doctoral Scientific Research Startup Fund of Harbin University of Commerce (Project No. 14LG15), and the Natural Science Foundation of Heilongjiang Province "Research on Human Resource Management Model and Demonstration System for Universities in Heilongjiang Province Based on Multi-Plane Management" (Project No. F201217).

Germany is home to over 8,000 public libraries and archives, approximately half of which are state or municipal institutions and half are church-affiliated facilities, complemented by more than 10,000 private libraries and archives. This translates to roughly one library or archive for every 4,000 people, underscoring Germany' s status as a leading nation in the library and archive industry [1].

Increasingly, countries and international organizations emphasize collaboration among digital libraries and archives. As more users publish data online, a global Web of Data has emerged. Compared to document networks, this structured data network forms more complex relationship webs, enabling easier retrieval and comprehension of Web data by both humans and machines. In February 2017, the W3C project released a new Linked Open Data Cloud visualization [Figure 1: see original paper], establishing a novel visual model. The number of open linked datasets has grown exponentially to several hundred, encompassing publications, cross-domain resources, media, linguistics, geography, user-generated content, government data, environmental information, life sciences, and social networks. LOD integrates these domain-specific open data resources into a visualized interconnected network. From an informatics perspective, this represents a new network paradigm beyond citation and co-authorship knowledge networks [3].

In recent years, digital libraries and archives have further promoted information resource sharing. A key challenge lies in providing access services for massive amounts of hidden, inaccessible data stored in data silos. With advancements in

Web technologies for heterogeneous data access, LOD enables metadata publishing, allowing library and archive collections to be searched, linked, and accessed sustainably [4]. Moreover, LOD represents an optimal method for publishing and sharing information using semantic technologies, providing access to vast amounts of heterogeneous data and stimulating application development. LOD helps digital libraries and archives escape data silos by publishing their data as structured information, delivering significant application value [2].

2 Successful Cases of German Digital Libraries and Archives

German digital library and archive success cases illustrate diverse information supply requirements and demonstrate how relevant data technologies address these needs. Furthermore, they clarify the primary advantages of LOD technology in digital library and archive applications.

2.1 Successful Application of the Linked Data Value Chain

German digital library and archive research projects convert publicly available data into linked data, with the vast majority generated by research institutions. Introducing the linked data value chain (see Figure 2 [Figure 2: see original paper]) into commercial engineering models enables conceptualization of successful business cases, defining role allocation, composition, and participation. However, inherent risks may exist in selected data and conversion processes, including usage rights, privacy policies, data availability, role incentives, data quality and credibility, data provenance, transparent data transformation, and interconnection.

The Leibniz Information Centre for Economics (ZBW) applied the linked data value chain to existing business cases at the BBC, testing potential risks during the process. Overall, the linked data value chain helps identify and classify potential risks that corresponding engineers can address, while establishing a clear methodology for understanding the complete linked data generation cycle. This model is easily applicable across other disciplines, including digital libraries and archives, life sciences, and media, facilitating linked data publication and highlighting potential issues that may arise during data conversion and linking processes [5].

2.2 LOD Technology for Author Information Retrieval in Digital Journals

One application value of LOD technology in digital journals involves linking real-world authors with digital journal authors through linked data. In ZBW' s digital environment analysis system, author name identification and disambiguation pose significant challenges when processing personal information. The analysis system identifies relevant personal details in profiles, such as expertise, social media influence, and publication volume. LOD-based analysis systems

play a crucial role in personnel allocation within organizations and institutions. Therefore, obtaining accurate author information is essential for enhancing the overall visibility and efficiency of digital journals [6].

Based on LOD, German scientists developed the CAF-SIAL platform (see Figure 3 [Figure 3: see original paper]), which searches for and provides personal information from linked data (http://cafsial.lod-mania.com). The CAF-SIAL platform employs heuristic techniques to identify relevant personal information from DBpedia by applying a "keyword" to the "URI" technology. This extracted information is further filtered and integrated under a conceptual aggregation framework, subsequently presented as a profile [7].

In library and archive environments, DBpedia and DBLP demonstrate application utility, further extending connections between digital journal authors and relevant semantic resources in LOD. This application can identify, disambiguate, retrieve, and structure relevant author information from these datasets. The system constructs a comprehensive author profile database, providing personal and professional information and listing academic achievements (http://dblp.l3s.de/d2r/).

Such systems can be applied across broader academic communication fields. Search subjects can be extended to integrated authority files, such as the German National Library' s Integrated Authority File (GND) (http://www.dnb.de/EN/gnd) and the Virtual International Authority File (VIAF) (https://viaf.org/), to obtain more complete results. Keywords and descriptors contained in authority files, assigned to publications during cataloging, can further simplify search and retrieval processes.

2.3 LOD Technology for Linked Data Publishing

In recent years, LOD has played a major role in data openness and become one of the most important library-like applications. These repositories are systems for collecting, publishing, disseminating, and archiving digital scientific content. Regarding digital library and archive applications, EconStor makes metadata of scientific papers in repositories machine-readable (http://econstor.eu). EconStor serves as Germany' s open access server for the National Library of Economics, providing a platform for publishing economics research papers. Currently, EconStor offers full-text access to scientific papers from nearly 100 institutions and over 80,000 complete text documents [8].

The D2RQ platform is used (http://d2rq.org/) (see Figure 4 [Figure 4: see original paper]) with the following steps: First, the open repository serves as a relational database; second, publications and authors are mapped to D2R server conversion mapping files using vocabularies; finally, repository data is converted using the D2R server and published as linked data with a SPARQL endpoint (http://linkeddata.econstor.eu/beta/snorql/). Repository content can be directly published as Linked Open Data and linked to valuable external datasets, contextualizing repository data and imbuing it with meaning. Pub-

lishing EconStor as linked data achieves the following objectives: enabling publication and dissemination of current research by posting scientific papers on the Semantic Web; successfully transforming typical repository systems (such as DSpace) into Semantic Web open content and integrating them into linked data flows; enabling distributed research information queries through SPARQL query patterns, such as retrieving all articles about financial crises published by European research institutions after 2012.

Publishing EconStor as linked data potentially impacts mashup application development (applications that manage data from different relevant linked data stores). From a software engineering perspective, this research provides methods for publishing repository content as Linked Open Data, generating significant interest among librarians, repository managers, and software developers.

3 Technical Challenges and Solutions

3.1 Entity Resolution

"Entity resolution" refers to identifying whether two resources in Linked Open Data point to the same real-world entity. This represents a challenging task because resources lack inherent identity; their meaning is defined solely through semantic descriptions and properties linking resources. One solution involves manual alignment. The German National Library' s Integrated Authority File contains author information linked to DBpedia [9]. However, manual alignment is extremely labor-intensive and impractical for merging large datasets. For instance, DBpedia contains 364,000 entries, the German National Library authority database contains 1,797,911 entries, the Library of Congress database contains 3,800,000 entries, and VIAF contains approximately 10 million entries (VIAF combines multiple name authority files from different national libraries). These databases are enormous, making entity resolution based solely on names, collaborators, titles, and locations typically insufficient [10].

3.2 Schema Matching

Schema matching faces similar challenges to entity resolution. The goal of Linked Open Data is defining and publishing self-describing vocabularies by referencing concepts from existing vocabularies. However, integrating different vocabularies and their described data is crucial, even for databases with similar schemas. Schema matching quality requirements are extremely high when applying schema integration to improve library and archive services [11]. Therefore, manual thesaurus alignment methods are used for schema matching across different works. For example, ZBW manually created thousands of mappings between the economics thesaurus STW (http://zbw.eu/stw/versions/latest/about) and other thesauri (such as TheSoz in social sciences, http://lod.gesis.org/pubby/page/thesoz/) during 2004-2005. The Simple Knowledge Organization System (SKOS) vocabulary typically describes relationships between keywords (http://www.w3.org/2004/02/skos/).

Since thesauri often contain thousands or even tens of thousands of subject terms and corresponding synonyms, automatic schema matching methods are necessary. Consequently, ZBW launched the Ontology Alignment Evaluation Initiative (OAEI) in 2012, which aims to compare different schema matching techniques and establish consensus on ontology matching method evaluation (http://oaei.ontologymatching.org/).

3.3 Distributed Data Management

LOD data is inherently distributed, with VIAF serving as an excellent example. Over a dozen international organizations collaborate to build a distributed library and archive resource network, including not only publishers but also individuals and organizations. Accessing distributed data requires federated query techniques to identify data sources and storage formats.

In the Semantic Web, researchers have developed various technologies, such as query techniques for linked distributed data, stream processing for Linked Open Data, and technologies for searching service data and data sources. However, it remains unclear which approach best suits distributed data access [12].

Furthermore, providing library and archive search services requires considering search result ranking to meet user needs. Like web search, users perceive the first link in search results as more important or relevant than others. To address this, ZBW' s DFG (German Research Foundation) project developed LibRank (http://www.librank.info/).

3.4 Automatic Indexing

Contrary to the indexing concept in database communities, indexing in libraries and archives refers to assigning multiple tags to classify documents such as scientific publications and archives. One indexing method involves manual tagging. German scientists have used STW to tag over 1.6 million economics publications, with an average of 5 STW subject terms per publication. Additionally, the EconStor publishing server enables automatic publishing of authors and keywords from STW and other thesauri.

Moreover, the number of electronic publications issued annually by the German National Library has increased significantly, necessitating automated literature indexing methods. Automated PDF classification methods have been developed accordingly. For example, the PETRUS project at the German National Library uses support vector machines to classify 100 categories (Sach-gruppen). The DFG-funded GERHARD project in the 1990s studied methods for automatic indexing of scientific web content.

Researchers have automatically indexed approximately 1 million documents using the Universal Decimal Classification (UDC). UDC indexing employs three languages (German, English, French). Oracle relational database management

systems enable full-text indexing (ConText). Automatic indexing of scientific literature remains a highly active research area [10].

Recent ZBW projects apply Linked Open Data for automatic indexing of scientific documents. kNN classifiers, entity detection, and HITS algorithms evaluate STW suitability for specific documents. The advantage of ZBW' s automatic indexing experiments is that they require no expensive training [13].

Although many consider "automatic indexing" as a process without human involvement, these technologies require human intervention for accurate operation. In practice, library and archive professionals must continuously monitor the quality of automatically suggested descriptors using their expertise to ensure correct subject representation.

3.5 Indexing Non-Textual Content

Beyond textual content such as PDF scientific publications and indexed websites, substantial non-textual content exists, including social media and audiovisual materials. These materials encompass mappings of traditional scientific content, social media profiles, and research data. ZBW addressed indexing challenges for these non-textual contents in the EU project EEXCESS (http://eexcess.eu/). The concept involves automatically combining structured scientific content (metadata, full text, paragraphs, citations, and other content) with informal and ephemeral content from social media channels to link subjects, objects (textual and non-textual resources), and users. Challenges remain in entity resolution, multi-schema indexing, and cross-media content retrieval.

To address multi-modal retrieval issues, ZBW developed a novel channel to better understand charts contained in scientific publications. This channel automatically extracts multiple text information items from charts through various methods (such as combinations of data mining and computer vision techniques). This enables textual searching of information graphics and their integration with scientific publication text content [14].

3.6 Data Provenance

The Virtual International Authority File (VIAF) enables bibliographic record retrieval across organizations, borders, and languages. Matching and linking open authority files can reduce costs and increase utility. However, cross-border and cross-language scenarios introduce new challenges: How to track data/metadata (re)use? How should Library A reference metadata when using (partial) records from Library B? How to evaluate the credibility of data/metadata merged into systems?

To address data provenance tracking, the library and archive science community has developed complex models for describing library and archive resources. The FRBR model can describe different variants of the same library resource, such as different printings of the same book or different language translations

(http://www.ifla.org/publications/functional-requirements-for-bibliographic-records). Thus, it applies not only to books but to any resource. Additionally, the RDA model can describe any content type, including online media, and allows attaching information sources to different data (http://www.rda-jsc.org/rda.html). The Europeana Data Model can query the provenance of both persons creating metadata records and the resources themselves (http://www.europeana.eu/portal/).

However, reliable methods for verifying metadata provenance remain lacking. The digital signature framework for graph data developed by A. Kasten et al. can track metadata provenance by digitally signing graphs and publishing data with signatures on the web, such as Linked Open Data, thereby establishing a "web of trust" [15].

Additionally, applications like the semantic search engine Sig.ma support entity searches in LOD and filter results by source. Unfortunately, this project has been discontinued [16].

summarizes specific LOD technology applications in German digital libraries and archives and their limitations, revealing further research directions for LOD technology in library and archive science.

Table 1. Comparative Analysis of LOD Research in Library and Archive Science

Application Area	Description	Limitations
Entity Resolution	Identifying whether two Linked Open Data resources point to the same real-world entity	Requires manual alignment, very expensive, cannot merge large datasets
Schema Matching	Defining and publishing self-descriptive vocabularies to improve library/archive services through schema integration	Requires manual thesaurus alignment for different works
Distributed Data Management	Querying highly distributed data on the web	Unclear optimal methods for accessing distributed data; requires result ranking considerations

Application Area	Description	Limitations
Automatic Indexing of Scientific Documents	Automated indexing using classifiers and algorithms	Requires continuous quality monitoring by library/archive scientists
Indexing Non-Textual Content	Mapping traditional scientific content, social media profiles, and research data	Entity resolution and cross-media retrieval remain challenging
Data Provenance	FRBR concepts integrated into RDA to describe any content; Europeana Data Model queries provenance	Lacks applications for perceiving information sources

4 Implications for China

The collection, storage, application, and long-term preservation of digital information are inseparable from digital and network technology development. Since 1998, German libraries and archives have participated in multiple EU projects, including the "European Networked Deposit Library," focusing on digital resource preservation and application technologies, building foundational network platforms, developing multimedia transmission systems, and researching migration and emulation techniques for information reproduction. To date, many technologies developed by German libraries and archives based on LOD technology exhibit universality and applicability, with some making positive contributions to global digital library and archive development. The German quality of scientific and technological spirit is also evident in library and archive technology, with its LOD applications holding an extremely important international position.

As LOD datasets continue growing rapidly, LOD technology applications in library and archive information services become increasingly widespread. LOD applications in China' s libraries and archives still exhibit shortcomings, with some research remaining theoretical rather than becoming operational applied technologies. These technologies provide fundamental technical support for future digital libraries and archives with broad applicability. Through comparative analysis of LOD technology applications in German libraries and archives (see Table 1), many practical research directions for library and archive work can be identified. Introducing LOD technology into China' s libraries and archives is imperative. By learning from German experience and building upon existing conditions to establish LOD-based linked application platforms, we can apply existing methods and tools to solve relevant problems in practice. Libraries and archives utilizing these new technologies will generate novel services.

References

[1] WANG Yongdan. A preliminary study on German public library services[J]. Library Theory and Practice, 2016(2): 8-11.

[2] BERNERS-LEE T. Linked-data design issues. W3C design issue document[EB/OL]. [2017-01-20]. http://www.w3.org/DesignIssue/LinkedData.html.

[3] XIA Lixin, TAN Ying. Analysis and visualization of LOD network structure[J]. New Technology of Library and Information Service, 2016(1): 65-72.

[4] HEATH T, BIZER C. Linked data: evolving the web into a global data space[M]//Synthesis Lectures on the Semantic Web: theory and technology. San Rafael: Morgan and Claypool, 2011: 1-136.

[5] LATIF A, SAEED A U, HO¨FLER P, et al. The linked data value chain: a lightweight model for business engineers[C]// 5th international conference on semantic systems. Graz: Graz Technical University Press, 2009: 568-575.

[6] LATIF A, AFZAL M T, HELIC D, et al. Discovery and construction of authors' profile from linked data (a case study for open digital journal)[C]//CEUR workshop proceedings. Raleigh: LDOW, 2010: 628.

[7] LATIF A, AFZAL M T, HOFLER P, et al. Turning keywords into URIs: simplified user interfaces for exploring linked data[C]// Proceedings of the 2nd international conference on interaction sciences: information technology, culture and human. Seoul: Int. Conf. Interaction Sciences, 2009: 76–81.

[8] LATIF A, BORST T, TOCHTERMANN K. Exposing data from an open access repository for economics as linked data[J]. D-Lib magazine, 2014, 20(9): 9-10.

[9] HALPIN H, PRESUTTI V. An ontology of resources: solving the identity crisis[C]//European semantic Web conference. Heraklion: Lecture notes in computer science, 2009: 521–534.

[10] NEUBERT J, TOCHTERMANN K. Linked library data: offering a backbone for the semantic web[C]// Third knowledge technology week. Kajang: CCIS, 2011: 37–45.

[11] WICK M L, ROHANIMANESH K, SCHULTZ K, et al. A unified approach for schema matching, coreference and canonicalization[C]//Proceeding of the 14th ACM SIGKDD, international conference on knowledge discovery and data mining. New York: ACM, 2008: 722–730.

[12] KONRATH M, GOTTRON T, STAAB S, et al. Schemex—efficient construction of a data catalogue by stream-based indexing of linked data[J]. Journal of Web semantics: preprint server, 2012(16): 52-58.

[13] 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[J]. IEEE STC social networking, 2015, 3(1): 149-157.

[14] BOSCHEN F, SCHERP A. Multi-oriented text extraction from information graphics[C]// Symposium on document engineering (DocEng). Lausanne: ACM, 2015.

[15] KASTEN A, SCHERP A, SCHAUB P. A framework for iterative signing of graph data on the web[C]// The semantic Web: trends and challenges proceedings. ESWC 2014. Lecture Notes in Computer Science. Anissaras: Springer, 2014: 146–160.

[16] TUMMARELLO G, CYGANIAK R, CATASTA M, et al. Sig.ma: live views on the Web of data[J]. Web Semantics, 2010, 8(4): 355-364.

Note: Figure translations are in progress. See original paper for figures.