# Application of LOD Technology in German Libraries and Archives: A Postprint

**Authors:** Dong Jie

**Date:** 2023-10-08T00:00:00+00:00

## Abstract

[ Purpose/Significance ] Linked Open Data (LOD) has been widely applied across numerous industries, non-profit organizations, and government agencies. Libraries and archives represent 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 cases of LOD implementation within these institutions.

[ Method/Process ] This study employs literature review, web-based investigation, and content analysis to examine successful application cases of LOD technology in German libraries and archives.

[ Results/Conclusion ] These cases illuminate the interconnections among research topics in computer science domains, including artificial intelligence, databases, and library and archival science. The practical experiences from Germany are synthesized to provide valuable references for developing related practices in our country.

## Full Text

## Preamble

**ChinaXiv Partner Journal [Academic Exploration]**
Application of LOD Technology in German Libraries and Archives
Library of Harbin University of Commerce, Harbin 150028

## Abstract

[**Purpose/significance**] Linked Open Data (LOD) has been widely applied across numerous industries, non-profit organizations, and government agencies. Libraries and archives represent early adopters of LOD technology, and their implementation has further propelled LOD's development. Germany possesses

a highly advanced library and archive sector, offering many successful cases of LOD application. [**Method/process**] This study employs literature investigation, web surveys, and content analysis to examine successful LOD implementations in German libraries and archives. [**Result/conclusion**] These cases reveal connections among research topics in computer science—including artificial intelligence, databases, and knowledge discovery—and library/archive studies. Summarizing Germany's practical experience provides valuable references for developing similar practices in China.

**Keywords:** Linked Open Data, LOD, Germany, Library, Archives, Application

## 1 Introduction

Germany is home to over 8,000 public libraries and archives, approximately half of which are state or municipal institutions and half church-affiliated, complemented by more than 10,000 private libraries and archives. This density—roughly one library or archive per 4,000 inhabitants—demonstrates Germany's status as a leader in the library and archive industry [1].

Growing numbers of nations and international organizations are prioritizing collaboration among digital libraries and archives. As users increasingly publish data online, a global Web of Data has emerged. Unlike document-centric networks, this structured data network creates more complex relationship webs that facilitate data retrieval and comprehension for both humans and machines. In February 2017 [2], the W3C project released a new Linked Open Data Cloud diagram (LOD Cloud), shown in Figure 1 [Figure 1: see original paper], establishing a novel visual model. The number of open linked datasets has multiplied several-fold to several hundred, spanning publications, cross-domain resources, media, linguistics, geography, user-generated content, government, environment, life sciences, and social networks. LOD integrates these diverse domain resources into a visualized, interconnected network. From an informatics perspective, this represents a new network paradigm following citation and co-authorship knowledge networks [3].

Digital libraries and archives have accelerated information resource sharing, yet face the challenge of providing access to vast quantities of hidden, inaccessible data stored in silos. As Web technologies for heterogeneous data access have matured, LOD enables metadata publication, allowing library and archive collections to be searched, linked, and accessed sustainably [4]. Moreover, LOD offers an optimal approach for publishing and sharing information using semantic technologies, providing access to large volumes of heterogeneous data that can stimulate application development. LOD helps digital libraries and archives escape data silos by publishing their data as structured information, delivering substantial application value [2].

## 2 Successful Cases of German Digital Libraries and Archives

These success cases illustrate diverse information supply needs in digital libraries and archives and demonstrate how relevant data technologies address these requirements, while clarifying LOD's primary advantages in this domain.

### 2.1 Successful Application of the Linked Data Value Chain

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

The German Leibniz Information Centre for Economics (ZBW) applied this value chain to existing BBC3 business cases while testing potential risks. Overall, the linked data value chain helps identify and classify potential risks for engineers to address, establishing a clear methodology for understanding the complete linked data lifecycle. This model readily transfers to other disciplines—including digital libraries and archives, life sciences, and media—facilitating linked data publication and highlighting potential issues in data conversion and linking processes [5].

### 2.2 LOD Technology for Author Information Retrieval in Digital Journals

A key application value of LOD in digital journals lies in linking real-world authors with their digital journal identities. The ZBW digital environment analysis system confronts author name identification and disambiguation challenges when processing personal information. The system extracts relevant details from profiles, including expertise, social media influence, and publication counts. LOD-based analysis systems play crucial roles in organizational personnel allocation, making accurate author information essential for enhancing digital journal visibility and efficiency [6].

German scientists developed the CAF-SIAL platform (see Figure 3 [Figure 3: see original paper]) to search and provide personnel information from linked data (http://cafsial.lod-mania.com). Using heuristic techniques, CAF-SIAL identifies individual information from DBpedia by applying keyword-based URI extraction. This information undergoes further filtering and integration into a conceptual aggregation framework presented as a profile [7].

In library and archive environments, DBpedia and DBLP demonstrate practical utility by extending connections between digital journal authors and LOD

semantic resources. The system identifies, disambiguates, retrieves, and constructs comprehensive author profiles containing personal and professional information alongside academic achievements (http://dblp.l3s.de/d2r/). Such systems apply to broader scholarly communication domains, with search capabilities extendable to integrated authority files like 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/) for more complete results. Authority file keywords and descriptors assigned during cataloging further streamline search and retrieval processes.

### 2.3 LOD Technology in Linked Data Publishing

LOD has significantly advanced data openness in recent years, becoming a critical repository application. These repositories collect, publish, disseminate, and archive digital scientific content. For digital library and archive applications, EconStor provides machine-readable metadata for scientific papers (http://econstor.eu). As the German National Library of Economics' open access server, EconStor offers a platform for publishing economics research, currently providing full-text access to papers from nearly 100 institutions and over 80,000 complete documents [8].

The D2RQ framework converts relational datasets into understandable statements, publishing EconStor repository data as linked data (http://d2rq.org/) (see Figure 4 [Figure 4: see original paper]). The process involves: (1) treating the open repository as a relational database; (2) mapping publications and authors to D2R server mapping files using vocabularies; and (3) converting repository data via D2R server for querying as linked data and SPARQL endpoints (http://linkeddata.econstor.eu/beta/snorql/). Repository content can be directly published as Linked Open Data and linked to valuable external datasets, contextualizing and enriching the data.

Publishing EconStor as linked data achieves several goals: disseminating current research via semantic web publication; transforming typical repository systems (e.g., DSpace) into semantic web open content integrated into the linked data flow; and enabling distributed research information queries through SPARQL patterns, such as retrieving all articles about the financial crisis published by European institutions after 2012.

This approach has implications for mashup application development that manage data from multiple linked data stores. From a software engineering perspective, the 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 identifies whether two Linked Open Data resources refer to the same real-world entity. This challenging task arises because resources lack inherent identity; their meaning derives solely from semantic descriptions and connecting attributes. Manual adjustment offers one solution, as seen in the German National Library's Integrated Authority File containing author information linked to DBpedia [9].

However, manual adjustment is labor-intensive and impractical for large dataset integration. DBpedia contains 364,000 entries, the German National Library Authority database holds 1,797,911 entries, the Library of Congress database includes 3,800,000 entries, and VIAF comprises approximately 10 million entries (combining multiple national library authority files). For such vast databases, entity resolution based solely on name, collaborators, title, and location proves insufficient [10].

### 3.2 Schema Matching

Schema matching faces challenges similar to entity resolution. Linked Open Data aims to define and publish self-describing vocabularies by referencing existing concepts, yet integrating different vocabularies and their described data remains critical, even for similarly structured databases. Schema matching quality requirements are stringent when improving library and archive services through schema integration [11]. Manual thesaurus adjustment has been used for schema matching across works; for instance, ZBW manually created thousands of mappings between the economics thesaurus STW (http://zbw.eu/stw/versions/latest/about) and other thesauri like TheSoz in social sciences (http://lod.gesis.org/pubby/page/thesoz/) during 2004-2005. Mappings are typically described using Simple Knowledge Organization System (SKOS) vocabulary (http://www.w3.org/2004/02/skos/).

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

### 3.3 Distributed Data Management

LOD data is inherently distributed, with VIAF exemplifying this characteristic through collaboration among dozens of international organizations building a distributed library and archive resource network encompassing publishers, individuals, and organizations. Accessing distributed data requires federated query technology to identify data sources and storage formats.

Semantic web researchers have developed various technologies for querying linked open distributed data, streaming linked open data, and searching service data and sources. However, the optimal method for accessing distributed data remains unclear [12]. Additionally, providing library and archive search services requires result ranking to meet user needs; like web search, users perceive first links as more important or relevant. To address this, ZBW's DFG project developed LibRank (http://www.librank.info/).

### 3.4 Automatic Indexing

Unlike database indexing, library and archive indexing assigns multiple tags to classify documents such as scientific publications and archives. Manual tagging represents one approach; German scientists have used STW to tag over 1.6 million economics publications, averaging five STW subject terms per publication. The EconStor publishing server also enables automatic publishing of authors and keywords for STW and other thesauri.

The German National Library and Archives' annual electronic publication volume has increased significantly, necessitating automated indexing methods. Automated PDF classification approaches have been developed, such as the PETRUS project using support vector machines to classify 100 categories (Sachgruppen). The DFG-funded GERHARD project in the 1990s studied automatic indexing methods for scientific web content, using the Universal Decimal Classification (UDC) to automatically index approximately 1 million documents in three languages (German, English, French) with Oracle's full-text indexing (ConText). Automatic scientific literature indexing remains an active research field [10].

Recent ZBW projects apply Linked Open Data to automatic scientific document indexing, using kNN classifiers, entity detection, and HITS algorithms to evaluate STW term matching for specific documents. ZBW's automatic indexing experiments offer the advantage of requiring no expensive training [13].

Although "automatic indexing" implies no human involvement, these technologies require human intervention for accuracy. In practice, library and archive professionals must continuously monitor the quality of automatically suggested descriptors using their expertise to ensure proper subject representation.

### 3.5 Indexing Non-textual Content

Beyond textual content like PDF publications and indexed websites, substantial non-textual content exists, including social media and audiovisual materials encompassing traditional scientific content mappings, social media profiles, and research data. ZBW addressed indexing these non-textual contents in the EU EEXCESS project (http://eexcess.eu/), which automatically combines structured scientific content (metadata, full text, paragraphs, citations) with informal, ephemeral social media content to associate topics, objects (textual and non-textual resources), and users.

Challenges persist in entity resolution, multi-schema indexing, and cross-media retrieval. To address multi-modal retrieval, ZBW developed a channel to better understand charts in scientific publications by automatically extracting textual information from charts through combined methods like data mining and computer vision, enabling textual chart searches integrated with publication text [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 reduces costs and increases utility, yet cross-border and cross-language scenarios raise new questions: How to track data/metadata (re)use? How should Library A reference metadata when using (part of) Library B's records? How to evaluate the credibility of merged data/metadata?

To address provenance tracking, the library and archive science community has developed complex resource description models. The FRBR model describes variants of the same resource, such as different printings or language translations (http://www.ifla.org/publications/functional-requirements-for-bibliographic-records), applying to any resource type. The RDA model describes all content types including online media and allows attaching information sources to data (http://www.rda-jsc.org/rda.html). The Europeana Data Model queries sources of both metadata creators and resources (http://www.europeana.eu/portal/).

However, reliable metadata provenance verification methods remain lacking. A. Kasten et al.'s framework for digitally signing graphic data tracks metadata provenance by marking graphics with digital signatures published on networks like Linked Open Data, establishing a "web of trust" [15]. Applications like the semantic search engine Sig.ma support LOD entity searching with source-based filtering, though the project has been discontinued [16].

Table 1 summarizes LOD technology applications in German digital libraries and archives and their limitations, revealing future research directions.

## 4 Implications for China

Digital information collection, storage, application, and long-term preservation are inseparable from digital and network technology development. Since 1998, German libraries and archives have participated in EU initiatives like the "European Networked Deposit Library," researching digital preservation, building network platforms, developing multimedia transmission systems, and studying migration and emulation technologies. Many LOD-based technologies developed by German libraries and archives exhibit universality and applicability, with some contributing significantly to global digital library development. Germany's commitment to scientific and technological excellence is evident in library and archive technology, where its LOD applications hold important international status.

As LOD datasets grow rapidly, LOD technology applications in library and archive information services are expanding. China's library and archive sector still shows deficiencies in LOD application, with some research remaining theoretical rather than operational. These technologies can provide fundamental support for future digital library applications with broad utility. Comparing German LOD applications (see Table 1) clarifies research directions for library and archive practices. Introducing LOD technology into China's libraries and archives is urgent; by learning from German experience and building LOD-based linked application platforms with existing resources, we can apply established methods and tools to solve practical problems, generating new services through these technologies.

## References

[1] 王永丹. 德国公共图书馆服务初探 [J]. 图书馆理论与实践, 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] 夏立新, 谭荧. LOD 的网络结构分析与可视化 [J]. 现代图书情报技术, 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: 1-8.
[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: 37–44.
[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] KONRAT 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 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.*