Analysis of the Application Logic of Linked Open Data in Library and Information Services

Authors: Gu Liping

Date: 2016-06-08T00:00:00+00:00

Abstract

Linked Open Data constitutes an important component of open data integration applications within open science. This paper briefly introduces the application requirements and scope of Linked Open Data; focuses on implementation strategies for libraries' application of Linked Open Data, including the transformation of bibliographic data and usage data into open data, as well as the utilization of open data such as social networks and social tags; and systematically analyzes the technical pathways of Linked Open Data for bibliographic quality enhancement, electronic resource management, informetric analysis, patent competitive intelligence, digital curation, embedded subject consultation, and related domains.

Full Text

Preamble

Linked Open Data in Library and Information Services

(National Science Library, Chinese Academy of Sciences, Beijing 100190, China)

Abstract Linked Open Data (LOD) represents a crucial component of open data integration and application within open science. This article briefly introduces the application requirements and scope of LOD, focusing on implementation schemes for libraries using LOD—including how to transform bibliographic and usage data into open data and how to leverage open data such as social networks and social tagging. The author systematically analyzes the technological

roadmap of LOD for improving bibliographic quality, electronic resource management, informetric analysis, patent competitive intelligence, digital archives, and embedded subject consultation.

[**Keywords**] Open Science; Open Data; Linked Data; Semantic Web; Ontology; Social Network; Social Tagging; Social Media; Community Space
[**Classification Number**] G250

---

1. Library Choices in Handling Open Data in the Open Science Era

The reuse of knowledge published in scientific publications and open critique constitute the two foundations of science. Open scientific data is key to enabling effective research operations and allowing society to fully benefit from scientific labor, with its foundation being the effective dissemination of research data with "openness" [1,2]. In the internet era, openness that supports open knowledge, open content, and open services has 11 characteristics [3]: access/redissemination, reuse, no technical restrictions, attribution, integrity, equal treatment of users, equal treatment of fields, authorized dissemination, licensing terms not exclusive to specific products, and licensing terms not restricting the dissemination of other works. Therefore, not all online information possesses these openness characteristics, making the full promotion and utilization of existing open data in digital libraries a critical issue for contemporary scientific development.

On the other hand, network information resources are heterogeneous (different content types and genres), heterogeneous (different data formats and corresponding semantic rules), and distributed, which is unfavorable for traditional processing methods that centralize information resources in one location (such as libraries). As an emerging knowledge organization and discovery method, Linked Data provides a set of low-cost standardized data access mechanisms and can effectively circumvent some complex data rights disputes. However, in the process of publishing existing resources as Linked Data, complexity arises from different service methods, classification systems, data content, and ontologies/vocabularies adopted in different domains [4]. This complexity further leads to issues with interactive interfaces, relationship validity, data fusion mapping, and data licensing rights [5].

As one of the important driving forces for Semantic Web development, Linked Data connects previously unrelated data through the network. The key elements include [6]: data collection, publishing linked data, standardizing "connection points" in links, obtaining data from individual datasets, and client application data. Linked Open Data (LOD) includes six core elements [6]: standards, access, license, identifiers (URIs), data model (such as RDF), ontologies (RDFS, OWL), and query language (SPARQL).

In the networked, digital, and open era of scientific research, library and information services must confront the emerging demands of open data and open science, and LOD is timely and relevant.

2. Formal Logic of Linked Open Data

Let us define: - Dataset D{d , d , ···, d }; metadata describing D: M{m d , m d , ···, m d } - Dataset D' {d' , d' , ···, d' }; metadata describing D' : M' {m' d' , m' d' , ···, m' d' } - Dataset D' '{d' ' , d' ' , ···, d' ' }; metadata describing D' ': M' '{m' ' d' ' , m' ' d' ' , ···, m' ' d' ' }

According to traditional methods: - Extracting m and m from M yields D{d , d } and D' {d' , d' }

However, if we define M = M' , then: - Extracting m and m from M yields D{d , d }, D' {d' , d' }, and D' '{d' ' , d' ' }

If we define M ~ M' where m = m' and m m' , then: - Extracting m and m from M yields D{d , d }, D' {d' , d' }, and D' '{d' ' }

Comparing these scenarios reveals two key properties of this strategy: data that would not previously be found can be discovered through newly defined relationships; adjusting definitions can obscure partial data. This open-close dynamic forms the logical foundation.

Furthermore, if we use function L to represent access permissions for each D, then the extraction process from M to D must incorporate the L function. In reality, licenses take many forms, such as the 11 characteristics of openness. For conceptual understanding, we simplify license L to two forms: - L = 0 // Access not permitted - L = 1 // Access permitted

This leads to the consideration: How can linking occur if data is not open?

3. Three Effects of Linked Open Data

Traditionally, libraries process data as either data or metadata. Data is collected and preserved—narrowly referring to a set of bits in a computer, broadly referring to content carriers such as books, journals, newspapers, and individual electronic files. Metadata is used to describe this data, such as bibliographic records, abstracts, classifications, and catalogs. Based on this foundation, experts can perform information work including acquisition, analysis, and presentation.

Theoretically, the rapidly developing network environment and computing technologies enable free data circulation and free metadata creation, allowing new datasets to be created through recombination, addition, and deletion of datasets. Metadata can be established to create relationships between metadata, enabling discovery of previously unfoundable data from these relationships.

Reframing the above question as an observable research problem: To what extent can openness thresholds be utilized to produce what effects?

If LOD can maximize data openness, three significant effects emerge:

(1) Structuring Unstructured Text on the Network
Although structured data continues to increase online (especially actively published data to promote linking), the network still includes (and is primarily composed of) unstructured data, particularly textual content. Therefore, an important question is: How can people effectively access large amounts of unstructured information online?

Let text information be T{D , D , ···, D }. To satisfy LOD formal logic for D, there are two approaches: - **Data-driven**: Based on identical data d in D {d , d , ···, d } and D {d , d , ···, d }, form a new text dataset TD{d , d , ···, d } - **Inductive (Context-aware)**: Match similar datasets D in text information T{D , D , ···, D } based on known dataset D' {d , d , ···, d }

Combining both approaches—integrating ontologies, machine learning, information retrieval, information extraction, and text mining—is a hot topic in information retrieval [7-9].

(2) Enhancing Information Retrieval Accuracy Based on Structuring
Enhancing keyword-based search through LOD is a new approach to strengthening web navigation systems [10]. According to LOD formal logic, alternately using various M = M' and M ~ M' logics can form a classification mechanism for information retrieval core indexes, avoiding the situation described in case (4) where datasets D, D' , D' ' were previously limited. However, data heterogeneity problems exist, such as different URLs for the same object across different datasets [11]. The solution is to improve object recognition through semantic similarity metrics, relevant algorithms, and information architecture using ontologies [12].

LOD provides entity type information that can serve as prior knowledge for named entity classification. Using LOD methods to extract information (named entity strings and types) to build a new type of knowledge base and adding it to existing classification systems can improve application performance.

(3) Expanding Information Scope Based on Accurate Retrieval
Linked data not only pre-establishes reliable relationships between data objects, but the rapidly developing linked data space also provides powerful resource support for building efficient linked reference services [13]. Traditional search is document-oriented queries based on keywords, whereas LOD spans different data sources, linking web data between type-related entities in a machine-readable manner based on standards such as RDF format and SPARQL query language. It is more similar to distributed or federated databases, but the cooperating data sources independently maintain and update their respective data [14]. For example, the instance-based myCBR system [15], through model generation, data import, similarity model interpretation, and visualization user interfaces, uses Symbolic to extract information from different entity texts for data association and information recommendation.

4. Specific LOD Implementation Schemes for Libraries

Libraries using linked data technology can publish resources as linked datasets, improve retrieval service systems, enhance resource discovery services, achieve data fusion and semantic retrieval services, cross-institutional data access and reuse, promote academic research and exchange, and realize integration between libraries and teaching systems [16,17]. In fact, the National Science Library of the Chinese Academy of Sciences has already published entity relationships from its institutional repository (IR) as linked data formats capable of semantic revelation and performed semantic annotation on IR data [18].

However, satisfying open data requirements for open science involves not only the above linked data technology applications. The following describes specific schemes for how libraries can export bibliographic and usage data as LOD and import social network and social tagging data.

4.1 Converting Closed Data (Inside Library) to Open Data (Outside)

(1) LOD includes two parts: open (standards, access, license) and linked (identifiers, data model, ontology, query language) [19]. Therefore, considering common protocols enables libraries to expand from linked data to linked open data.

(2) There are multiple ways to output data, such as using a system module, adding a conversion system, or directly copying and transferring partial .dat and .ini documents within the file system. During the process of generating XML-structured files, disproportionate output/input can cause computer resource occupation and work interruptions due to data updates. Therefore, irregular dump updates are recommended.

(3) Raw data publication involving access, licensing, identifiers, and data models requires four steps: loading HTTPServer, describing the published dataset, attaching licenses, and registering the dataset.

(4) Appropriate RDF vocabularies must be selected for data (e.g., bio for bibliographic data) to convert identifiers and data, then map and convert raw data to RDF and write HTTP URIs to describe resources, finally presenting them through SPARQL interfaces and HTML via the Pubby framework.

4.2 Converting Open Data (Outside) to Internal Resources (Inside)

Growing social network data has become an important component of open data. Therefore, new distributed architectures have been built on Semantic Web identification systems (such as RDF(S)/OWL, RDFa, SPARQL, etc.). The new paradigm must not only allow users to own their data but also increase their data through the Semantic Web. Thus, based on centralized annotation and retrieval components like DBpedia, RDF-based infrastructures independent of context semantic relationship graphs and external information sources (such as search engine results and social tagging systems) can be developed as scalable

alternatives to web ranking technology. Using LOD semantic markup generates query-returned approximate results [20,21].

The specific approach uses the DBpedia modeling information retrieval framework, then extracts search history feature vectors from these concepts to establish an LOD model based on user queries [22]. However, this model requires batch import and batch processing maintenance capabilities.

End users typically annotate, comment on, and describe content in social tagging systems for different purposes, such as listing objects appearing in photos, expressing contextual information about entries, or self-referencing and organizing personal content (e.g., opinions, comments). If retrieval content contains a considerable proportion of tags, identifying the personal intentions behind basic tags may improve search and recommendation accuracy [23]. In practice, broader sampling scope and sustainable operation costs should be considered.

4.3 Other Potential Developments

The essence of linked data is linking ontologies and related data resources according to RDF format standards while requiring HTTP URI access and SPARQL query language retrieval. Following this principle, libraries can apply linked data. For example, Germany' s R2R enhances RDF linking capabilities from vocabularies and dataset instances to achieve LOD [24]. However, as datasets continue to grow, object co-reference problems caused by data publication openness, context independence, and generic resource identifiers [25] will trouble library decisions on applying linked data technology. On the other hand, while linked data technology can process library knowledge resource data, user behavior data involves more openness standards, access, and licensing principles—data that is key for libraries to implement LOD for open data and is unique to each library.

5. Future Directions for LOD Application

(1) Supporting Bibliographic Data Quality Assurance
Librarians engaged in cataloging work understand that building and maintaining thesauri is complex and arduous. If tags from large, already-structured text datasets are used to edit vocabularies, creating and maintaining thesauri becomes much easier [26]. Because LOD provides a vast set of structured data, once transformed into a vocabulary, its accuracy is relatively high. Moreover, even experts or professional librarians unfamiliar with Semantic Web techniques can smoothly perform thesaurus work based on the vocabulary.

(2) Supporting Electronic Resource Management Systems
Using Semantic Web technology and linked data principles can overcome limitations such as passively receiving resource lists from electronic content suppliers or operating on specific modules, improving data interoperability of resource list management tools [27]. Currently, all library workflows for electronic resource management must wait for integrators to provide resource lists, convert them

into resource lists for upload and management—this timing depends on content providers' "giving list" speed and librarians' "keying list" work speed. Another method is ordering knowledge bases and "transferring lists" from knowledge bases to electronic resource management resource lists. Applying LOD may accelerate the processes of "taking lists," "giving lists," "transferring lists," and "keying lists."

(3) Supporting Informetric Analysis

Linked open data provides opportunities to enhance data availability and utility. When applied properly, it can comprehensively improve scientific and technological development across all fields. Most published literature is in non-machine-readable formats, making information extraction, data organization, and accuracy improvement hot topics. For example, the developing PatentEye prototype system aims to perform information extraction on chemical reactions in patent literature, capturing data on reactants and product identifiers to accelerate information processing efficiency for scientific research [28]. Systems like PatentEye focus on leveraging linked open data to achieve text mining and information organization functions.

(4) Supporting Patent Competitive Intelligence

International patent classification in the World Intellectual Property Organization is crucial for patent searching, typically serving as an entry point in search processes. If scientific portal classification methods from Wikipedia are used, different categories can be assigned due to their different classification approaches. LOD can be used to build patent classification integration and interaction frameworks, enabling dataset integration that allows different patent ontologies to interact extensively [29]. Currently, this is frontier research in patentometrics and patent competitive intelligence: using linked open data to mine professional discipline classifications from specialized discipline portals, then comparing them with international patent classifications to collect and analyze relevant patents.

(5) Supporting Digital Archive Construction Capacity

Traditionally, digital archives mean scanning, filing, backing up, and limited access. However, to expand digital archive construction capacity beyond funding and manpower, technology-driven innovation can be employed. For example, applying innovation theory to cultural heritage open data can create a comprehensive semantic knowledge base based on Semantic Web technology and standards, forming integrated applications similar to museum open clouds through LOD [30]. LOD reduces the need for metadata editing staff, transforming data quality control into more manageable data quality assessment methods and converting high-quality data into open data to provide external services.

(6) Supporting Embedded Subject Service Models

Industrial Ecology (IE) is an emerging research field requiring community-driven data collection, processing, preservation, and sharing, as well as data and knowledge sharing mechanisms. Due to the technical and standard types involved being beyond many industrial ecologists' normal work scope, numerous young

scholars discuss this online [31]. However, people rarely notice that library science has extensive successful and unsuccessful experiences in data management that can help avoid detours. Research libraries have professional subject librarians and system institutional talent that can provide support. In interdisciplinary, emerging, and experimental disciplines, there are many communities —aware or unaware—needing linked open data services, representing potential service targets for library and information service teams.

6. Library LOD Applications Toward Open Science

LOD still faces other issues. For example, in network environments, people often encounter questionable or even contradictory information. Determining the authenticity of such information requires data provenance. Currently, creating provenance information typically uses annotation or query-based methods, with multiple provenance models existing. However, ensuring complete reproducibility of provenance information for scientific workflows remains a challenge [32]. When considering LOD, progress on other related issues must also be considered to fully present the complete picture of library information graph evolution, enabling libraries to better grasp open data processing capabilities and be more proactive and creative in the open science era. Based on practical implementation experience and theoretical foundations, this article aims to stimulate further discussion.

(Acknowledgments: Thanks to reviewers and editorial staff for their corrections.)

References

[1] Panton Principles. Principles for Open Data in Science [EB/OL]. [2012-08-16]. http://pantonprinciples.org/. [2] The Royal Society. Science as an Open Enterprise [EB/OL]. [2012-08-16]. http://royalsociety.org/policy/projects/science-public-enterprise/report/. [3] Open Definition. Open Definition Version 1.1 [EL/OL]. [2012-08-16]. http://opendefinition.org/okd/. [4] Shen Zhihong, Zhang Xiaolin. Linked Data and Its Applications: An Overview [J]. *New Technology of Library and Information Service*, 2010(11): 1-9. [5] Huang Yongwen. Research on Linked Data-driven Web Applications [J]. *Library Journal*, 2010, 29(7): 55-59. [6] Omitola T, Koumenides CL, Popov IO, et al. Putting in Your Postcode, Out Comes the Data: A Case Study [C]. In: *Proceedings of the 7th Extended Semantic Web Conference*. 2010: 318-332. [7] Horrocks I. Ontologies and the Semantic Web [J]. *Communications of the ACM*, 2008, 51(12): 58-67. [8] Krummenacher R, Norton B, Marte A. Towards Linked Open Services and Processes [C]. In: *Proceedings of the 3rd Future Internet Conference*. Berlin, Heidelberg: Springer-Verlag, 2010: 68-77. [9] Bloehdorn S, Blohm S, Cimiano P, et al. Combining Data-driven and Semantic Approaches for Text Mining [C]. In: *Proceedings of Foundations for the Web of Information and Services*. Springer, 2011: 115-142. [10] Waitelonis J, Sack H. Towards Exploratory Video

Search Using Linked Data [C]. In: *Proceedings of the 11th IEEE International Symposium on Multimedia*. 2009: 540-545. [11] Noessner J, Niepert M, Meilicke C, et al. Leveraging Terminological Structure for Object Reconciliation [C]. In: *Proceedings of the 7th Extended Semantic Web Conference (ESWC2010)*, Heraklion, Crete, Greece. Berlin, Heidelberg: Springer-Verlag, 2010: 334-348. [12] Ni Y, Zhang L, Qiu ZM, et al. Enhancing the Open-Domain Classification of Named Entity Using Linked Open Data [C]. In: *Proceedings of the 9th International Semantic Web Conference*. 2010: 566-581. [13] Liu Yuanyuan, Li Chunwang, Huang Yongwen. Study on Building the Service of Relevance Reference Based on LOD [J]. *New Technology of Library and Information Service*, 2011(6): 66-71. [14] Gorlitz O, Staab S. Federated Data Management and Query Optimization for Linked Open Data [A]. In: *New Directions in Web Data Management 1*. Berlin, Heidelberg: Springer-Verlag, 2011: 109-137. [15] Roth-Berghofer T, Adrian B, Dengel A. Case Acquisition from Text: Ontology-based Information Extraction with SCOOBIE for myCBR [C]. In: *Proceedings of the 18th International Conference on Case-Based Reasoning*. 2010: 451-464. [16] Huang Yongwen. Overview of Linked Data Applications in Libraries [J]. *New Technology of Library and Information Service*, 2010(5): 1-7. [17] Huang Yongwen, Yue Xiao, Liu Jianhua. Architecture on Linked Data Based Applications and Some Suggestions for Building Linked Data Applications [J]. *New Technology of Library and Information Service*, 2011(9): 7-13. [18] Wang Sili, Zhu Zhongming. Study on the Semantic Expansion of Institutional Repository Based on Linked Data [J]. *New Technology of Library and Information Service*, 2011(11): 17-23. [19] Schomburg S. Publishing Aleph Data as Linked Open Data [OL]. [2011-09-22]. http://igelu.org/wp-content/uploads/2011/09/Schomburg_2011_igelu_final.pdf. [20] Passant A, Breslin JG, Decker S. Rethinking Microblogging: Open, Distributed, Semantic [C]. In: *Proceedings of the 10th International Conference on Web Engineering*. 2010: 263-277. [21] Mirizzi R, Ragoné A, Di Noia T, et al. Ranking the Linked Data: The Case of DBpedia [C]. In: *Proceedings of the 10th International Conference on Web Engineering*. 2010: 337-354. [22] Meij E, Bron M, Hollink L, et al. Mapping Queries to the Linking Open Data Cloud: A Case Study Using DBpedia [J]. *Journal of Web Semantics*, 2011, 9(4): 418-433. [23] Cantador I, Konstas I, Jose JM. Categorising Social Tags to Improve Folksonomy-based Recommendations [J]. *Journal of Web Semantics*, 2011, 9(1): 1-15. [24] Tao Jun, Sun Tan, Liu Zheng. Linked Dataset Mapping Language: R2R [J]. *Journal of Library Science in China*, 2012, 38(3): 97-106. [25] Liu Yuanyuan, Li Chunwang. Studies on Object Co-reference in Linked Data [J]. *Information Studies: Theory & Application*, 2012, 35(2): 46-51. [26] Schandl T, Blumaue A. PoolParty: SKOS Thesaurus Management Utilizing Linked Data [C]. In: *Proceedings of the 7th Extended Semantic Web Conference (ESWC2010)*. 2010: 421-425. [27] Clarke C. A Resource List Management Tool for Undergraduate Students Based on Linked Open Data Principles [C]. In: *Proceedings of the 6th European Semantic Web Conference*. 2009: 697-711. [28] Jessop DM, Adams SE, Murray-Rust P. Mining Chemical Information from Open Patents [J]. *Journal of Cheminformatics*, 2011, 3(1): 40. [29] Pesenhoffer A, Berger

H, Dittenbach M. Offering New Insights by Harmonizing Patents, Taxonomies and Linked Data [A]. In: *Current Challenges in Patent Information Retrieval*. Springer, 2011: 357-371. [30] Damova M, Dannells D. Reason-able View of Linked Data for Cultural Heritage [C]. In: *Proceedings of the 3rd International Conference on Software, Services and Semantic Technologies*. 2011: 17-24. [31] Davis C, Nikolic I, Dijkema GPJ. Industrial Ecology 2.0 [J]. *Journal of Industrial Ecology*, 2010, 14(5): 707. [32] Shen Zhihong, Zhang Xiaolin. Data Provenance Model in Semantic Web Environment: An Overview [J]. *New Technology of Library and Information Service*, 2011(4): 1-8.

(Author E-mail: gulp@mail.las.ac.cn)

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