Submitted on: 14.11.2018

IFLA Satellite Meeting: *Knowledge Management Section*
Is knowledge management the new library science?
Date: 23 August 2018
Location: Xiamen University Malaysia campus (XMUM), Salak Tinggi, Selangor, Malaysia

Research on Semantic Publishing Model Based on Knowledge Management Ecosystem

Yongjuan Zhang[*]
Department of Library, Information & Archives Shanghai University, China.
Shanghai Information Center for Life Sciences Chinese Academy of Sciences, China
E-mail: zhangyj@sibs.ac.cn

Tao Chen[*]
School of Information Management, Nanjing University, Nanjing, China
Shanghai Library/Institute of scientific & Technical Information of Shanghai, Shanghai, China
E-mail: tchen@libnet.sh.cn

Heng Chen[§]
Shanghai Information Center for Life Sciences Chinese Academy of Sciences, China
E-mail: chenheng@sibs.ac.cn

Jianrong Yu[§]
Shanghai Information Center for Life Sciences Chinese Academy of Sciences, China
E-mail: jryu@sibs.ac.cn

Wei Liu[§]
Shanghai Library/Institute of scientific & Technical Information of Shanghai, Shanghai, China
E-mail: kevenlw@gmail.com

[*] Yongjuan Zhang and Tao Chen are co-first authors.
§ Wei Liu, Jianrong Yu,and Heng Chen are co-ccorresponding authors.

Abstract:

With change of the scientific research paradigm and development of the semantic technology, the traditional publication is facing the need change of the users of scientific research, so it begins to transform into a knowledge service model, which traces back to the source of the scientific research innovation process, is based on the mass content resources of the authoritative systems, and provides diversified, three-dimensional, customized services and solutions for the researchers and organizations dealing with research project selection, literature retrieval and analysis, experiments, scholarly communication and exchange, accomplishment publication and scholarly evaluation. The knowledge management system is the foundation of knowledge service. The new technologies such as semantic web, artificial intelligence and knowledge mapping etc. also provide support and possibility for evolution of the knowledge management system.

On basis of the above new technologies, a semantic publishing model is built in this study based on the knowledge management ecosystem. The knowledge management ecosystem using the machine readable RDF triple model as the metadata structure surpasses the restriction of the system or platform, makes it possible to understand each other between machines, and builds a cross-domain linking relationship which can be understood by the machine based on linked data and knowledge mapping so as to support and drive the transformation of science & technology publication into knowledge service, finally forming a machine-readable, data-driven, multidisciplinary collaboration, demand-oriented, resource deconstruction granularity, and gradually growing semantic publishing model.

In the further study, a semantic publishing model was established on the basis of the knowledge management ecosystem in the field of life sciences and basic medicine. The resource of the model system included researchers, institutes, literature, dataset, patents, funds, honors and evaluation and so on, which were stored in RDF triples and published in form of the linked data, and combined with natural language processing (NLP) and neural network based on the deep learning technology so as to realize reasoning and linking of the new knowledge as well as complete knowledge growth with help of the machine intelligence. This model system could facilitate seamless integration between different databases and also promote cross-domain integration of the system with external system resources such as Pubmed, and construct the service modules such as knowledge acquisition, data mining, internalization, sharing, evaluation, and externalization and so on, in order to provide support guarantee for cooperation among multiple different agencies.

Keywords: Semantic Publishing; RDF; Linked Data; Knowledge Management Ecosystem.

1 INTRODUCTION

Data-intensive research based on data or big data is becoming the "four paradigm" for the scientific research, at the same time, the massive research data has accelerated the transition from academic publishing to knowledge service, and has also accelerated the engineering and innovation of data warehousing technology, however, the traditional semantic publishing platform, knowledge management system and database possess these shortcomings such as a low degree of structuring, unfinished internal association of literature and the automatic recognition of machine so that they cannot meet the new demand of academic publishing users [1].

In the “Resource Development Plan for Data-Intensive Research in Social Sciences, Behavioral Sciences, and Economic Sciences”, the National Science Foundation also pointed out [2] (2017) that a new generation of digital repository with large-scale structured data and analytical functions was produced for the model transformation of data-intensive, interdisciplinary, collaborative, and problem-driven approaches. These digital warehouses were integrated with the Research Information Management (RIM) system [3]. OCLC also defined the role of libraries in the process and how RIM supported the Digital Scholarship.

Linked Data, as a best practice for opening data from the database to the semantic network of the web, has an increasingly widespread impact on semantic technology and business development. It is also becoming the research mainstream and important development directions of a semantic related discipline such as computer science and data science. .LD Knowledge Bases and LD Vocabularies continue to emerge. C. Sarven et al. proposed the opening principle of linked data so as to encourage the openness of related scientific research knowledge [4]. The Linked Open Research Cloud (LORC) [5] was opened in August 2017. LORC is a new knowledge management ecosystem who’s aims are to increase awareness, discovery, and reuse of online academic resources in the form of opening linked data for laying the foundation of the semantic publishing, which can solve the those existing problems of the data warehousing technology dealt with academic publication, such as structured RDF data, more granular knowledge units, deep semantic association and semantic enhancement in order to drive the transformation of semantic publishing model and its rapid development.

Therefore, the semantic publishing model based on the new Knowledge Management Ecosystem Linkage Open Research Cloud (LORC) is becoming the focus of the research institutes such as major academic publishing organizations and libraries, and also becoming one of the important directions of academic research open scientific research and digital academic investigation.

2 RELATED RESEARCH

2.1 REVIEW AND RESEARCH PROGRESS OF ACADEMIC HISTORY OF THE INTERNATIONAL RESEARCH

(1) The rapid development of LD Knowledge Bases and LD Vocabularies

The linked data had been put forward for more than ten years since 2006, the associated data repository has continuously emerged. There are representative databases such as DBpedia, GeoNames, etc., and ontologies and vocabularies for organizing the contents of knowledge database.

(2) The library's traditional bibliographic data is transiting towards LD Vocabularies, and the traditional Institutional Repository is also transforming into LD Knowledge Bases.

With the promotion of the linked data semantic technologies, the library bibliographic data firstly became the object of publication with its own closed characteristics. The Swedish National Library, the USA Library of Congress, the British Library and also on, began to act, and the BIBframe model allowed the publication of the bibliography to possess global standards, which accelerated the development of library-linked data vocabularies. The digital library has been undergoing decades of digitization of collection resources, the construction of institutional repositories based on metadata specifications, and long-term preservation and retrieval functions. Under the drive of the scientific research paradigm, the digital library gradually concentres toward resource-oriented content granularity and toward Digital Repository transformation of objects or things [6]. LD Knowledge Bases are best practices for digital warehousing.

With the emergence of the digital warehouse such as the publication of bibliography and associated data repository, the digital warehouse is also changing the way of library services, making library an important area for digital academic research such as semantic publishing, e-Science and digital humanities.

(3) The development of the associated open integration platform (Portal) and the proposal of the Linked Open Research Cloud (LORC) initiative.

With the constant emergence of individual LD Knowledge Bases and LD Vocabularies in the fields of library, e-Science, and digital humanities, the related open and converged platforms (Portal) have also emerged, such as Governments, Bio2RDF, and Europeana and so on. W3C's Linked Open Data (LOD) [7] (Linking Open Data, LOD) interconnects different linked data sources to form a huge dataset; while Linked Open Vocabularies (LOV) [8] aggregates the ontology and vocabulary of the linked data; the SPAR ontology set for semantic publication of academic literature [9] includes eight core ontologies for describing bibliographic features of bibliography and references, supplemented by FRBR and FOAF.

In the 10th LDOW 2017 (Linked Data on the Web) symposium, L. Jens et al. [10] proposed establishing an initiative of Pioneering the Linked Open Research Cloud, which created a new knowledge management ecosystem. In the next decade, linked data will play a greater role in academic exchanges, and will also become an important cornerstone for the formation of a new digital publishing ecosystem.

(4) Linked Data Portal and LORC Drive New Emergence of Semantic Publishing Models.

C. Sarven et al. proposed linked research principle in order to encourage the openness of associated scientific research [4]. Linked Open Research Cloud (LORC) [5] was opened in August 2017 for increasing the awareness, discovery, and reuse of online academic resources in the form of open, connected data. It is a new knowledge management ecosystem, which lays the foundation for semantic publishing. Springer Nature integrates information on scientific publications, literature, projects, funding, conferences, institutions, etc., launches the SciGraph associated open data platform [11], and takes the first step towards semantic publishing. Deconstructing full texts and forming open scientific knowledge maps will become its next stage of work.

2.2 REVIEW AND RESEARCH PROGRESS OF ACADEMIC HISTORY OF DOMESTIC RELATED RESEARCH

After many years of theoretical and case studies, the domestic research on connected data began in depth into bibliographic data distribution and knowledge base construction and other practical applications. Some breakthroughs have been made in the construction of LD Knowledge Bases and LD Vocabularies, such as Shanghai. Library's family tree, ancient books and other literature knowledge databases, combined bibliographic data based on BIBframe, Sinopedia's knowledge database of Chinese encyclopaedias, and Linked Brain

Data (LBD) of the Institute of Automation, Chinese Academy of Sciences [12]; Shanghai Institute of Life Sciences, Chinese Academy of Sciences The independent knowledge databases and knowledge management systems such as the "Semantic Knowledge Database of Life Science and Basic Medical Evaluation (π Index)" of the hospital, as well as related data portals, such as the Open Chinese Knowledge Map (OpenKG) initiated by Zhejiang University [13], with 87 RDF data sets with various topics being combined, but some parts of data sets could not be accessed, some of them were not published as format of the linked data, and therefore a complete knowledge management ecosystem was not formed.

In summary, the foreign countries have not quickly developed in the semantic publication field since David Shotton proposed the concept and model of semantic publishing in 2006. However, with the development of LD Knowledge Bases and LD Vocabularies, when a breakthrough was made in the semantic publishing related technology, driven by the relevant data, the area of semantic publishing has rapidly developed. There has been an advocacy plan for developing associating open scientific research clouds. With semantic publishing attempts and the emergence of the prototype SciGragh, the semantic structure of the document-oriented structure has become more mature, and the content-oriented semantic model has become the focus of research [14]. Since the semantic model of the Journal3.0 which has broken through the single article-by-unit information organization has been published, its application level is not fast and there are few application cases. The main reasons for analysis of this case are the concept precedence, while technology lag, which impacts development of from the model to application. At present, there are breakthroughs in the construction of LD Knowledge Bases and LD Vocabularies in China, but they are still scattered. There is no associated open integration platform (Portal) that can serve semantic publication, or a knowledge management ecosystem represented by LORC. Semantic publishing platforms based on this have not been reported. Therefore, a knowledge management system based on the associated Open Research Cloud (LORC) established in China for completing the growth and evolution from point to line to surface and laying the foundation of structured semantic knowledge associated with the development of the semantic publishing application platform has become a problem to be solved.

3 MODEL RESEARCH

3.1 RESEARCH OBJECTS

This study aims to demonstrate the construction of a semantic publishing platform based on the Linked Open Research Cloud (LORC) in the area of "Life Sciences and Basic Medicine", at the same time, an English journal "Cell Research (CR)" sponsored by the Shanghai Institute of Life Sciences, Chinese Academy of Sciences and the Chinese Journal of Chinese Cell Biology are used as carriers for studying data standards, related technologies, and model structures built on the LORC-based semantic publishing platform, and until the unified data standard for a set of structures based on E-RDF will be established. ; a set of aggregated LD Knowledge Bases and LD Vocabularies docking Chinese and English semantic publishing ontology, structure, and fine-grained full-text refactoring technology packages will be also built; and a complete LOCL-based semantic publishing model with publishing, integration and knowledge service functions suitable for promotion in various fields will be also created.

3.2 OVERALL FRAMEWORK

There are four modules in the open related semantics publishing model based on the linked open research cloud LORC (Figure 1):

Module 1: Unstructured data such as periodicals and scientific data are converted into structured RDF triples and stored in non-relationship databases.
Module 2: Screening and Integration of Linked Data Repository and Linked Data Vocabulary.
Module 3: The formation of the associated open scientific research cloud and the supported semantic publishing platform
Module 4: Service Editing, Peer Review, Open Research, Digital Academic (e-Science and Digital Humanities)

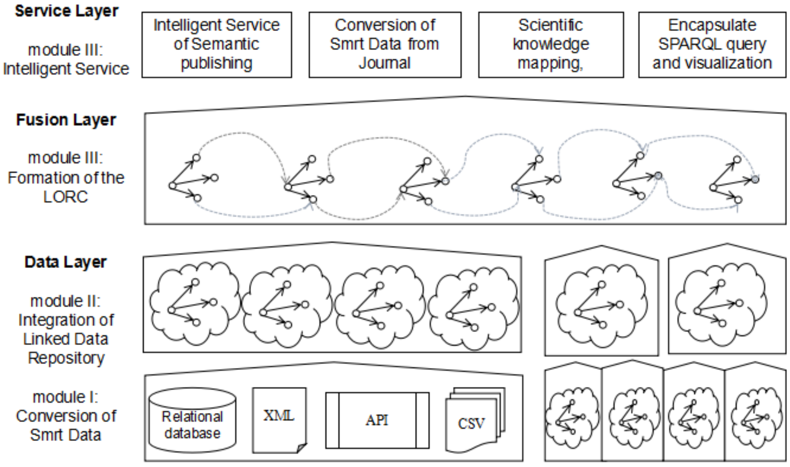
Service Layer
module III: Intelligent Service
Intelligent Service of Semantic publishing
Conversion of Smrt Data from Journal
Scientific knowledge mapping,
Encapsulate SPARQL query and visualization
Fusion Layer
module III: Formation of the LORC
Data Layer
module II: Integration of Linked Data Repository
module I: Conversion of Smrt Data
Relational database
XML
API
CSV

Figure 1 Open Association Relevance Semantic Publishing Model Based on Linked Open Cloud Research (LORC)

3.3 IMPLEMENTATION PROCESS

(1) Construction of Linked Data Publishing Platform to unify data standards and reduce technical barriers, so as to provide publishers with convenient data releasing and publishing models

Build a linked data publishing platform for magazines and scientific research workers, and provide a Linked Data Service solution that allows users to publish unstructured data to be released into linked data form in a streamlined, automated, and standardized way.

The platform combines the characteristics of a relational database (ER diagram) and the characteristics of RDF triples to establish a set of simple and universal standard E-RDF. This standard improves, optimizes, and develops the "five-star data" standard, known as The 5 +E-level data standard which is characterized with standardization, distributed, and machine readable. It is a high-quality semantic data in terms of graphic rendering and unique URI.

The core of the entire platform is the “3+1” technology components, namely the streamer, the manager, and the cocurator plus the integrator, to support the entire data circulation ecology.

As for the technology layout, a search portal + multiple technical sites operation mode have been proposed to build a high-quality global data sharing ecosystem, aiming to better build an integrated technology architecture.

(2) Building an Open Research Cloud Linked Open Research Cloud (LORC) HUB Aggregation Center

Based on the evaluation of the maturity of semantic data in SPARQL endpoint, a high-quality linked data set covering English and Chinese is selected. Many semantic data warehousing (both Chinese and English) opened in libraries, digital humanities, natural sciences, and semantic webs at home and abroad, including self-built semantic data knowledge bases, form a semantic publishing infrastructure together, but the data quality is uneven. There are 9960 associated datasets in the LOD, among which there are 6971 datasets with problems, accounting for 70.1%. It is necessary to screen high quality data.

Based on this, evaluation the semantic maturity of the data itself has been carried out, and each semantic data set with SPARQL endpoint has been proposed to be used as the evaluation unit to evaluate the basic characteristics, semantic level and use value of the data. -The concept of RDF and e-index are put forward and used to evaluate the semantic level of data. Finally, the LD Knowledge Bases and LD Vocabularies data sets with high semantic maturity are collected in data hub, to achieve seamless integration of these data sets which can be retrieved in unified way. This will well facilitate and guide users to discover and use semantic data sets with high maturity.

(3) Construct a semantic publication vocabulary and ontology HUB in line with China's national conditions, to realize the standardization, mapping and integration of Chinese and English ontology

To achieve the integration of Chinese-linked data and global data, it is necessary to integrate and align Chinese and English ontology. First of all, the ontology of semantic publishing in LOV is taken as a survey. A general ontology of semantic publishing can be established to reuse, correlate, and integrate into an ontology that conforms to China's national conditions.

In fact, there is no separation of Chinese and English ontology. In designing, there is no need for Chinese and English sets. As long as the URI is set with the international language English, the ontology can be used universally. Classes can be tagged in Chinese, or translated into Chinese and English.

According to this method, the first is to construct a semantic ontology model for document structure, including the description of chapters and other document structure components, literature titles, and bibliographic information; the second is to construct a semantic model for document content, specifically for description of research purposes, assumptions, arguments, methodologies, experiments and conclusion in academic literatures, namely the scientific discourse ontology.

(4) Deconstruct the full text based on the associated open scientific research cloud, vocabulary, and ontology to do X-ray for the full text and construct the scientific research knowledge map

Based on (1) a unified repository of associated data standards, Linked Open Research Cloud (LORC) built in (2), and Chinese and English ontology and mining algorithms built in (3), the original data organization structure system has been broken to realize data correlation, interoperability, data mining and other functions. Describe the publication from a conceptual perspective, do X-rays for the full text, construct knowledge maps based on fine-grained knowledge units to achieve knowledge services based on semantic publishing in a real sense. Publication semantic description has important significance for data interoperability and data association.

(5) Model construction: Semantic publishing model based on Linked Open Research Cloud (LORC) — Open relational semantic publishing platform with data publishing, fusion, reuse, and mining functions.

Build a third-party semantic publishing platform to provide a unified linked data conversion portal, which can realize the interoperability of data between different publishers, reduce costs, and adopt a unified data model and a unified vocabulary under the same discipline by different publishers, to achieve the relevance of publications.

(6) Empirical test: "Cell Research" (CR), an English journal sponsored by the Shanghai Institute of Life Sciences, Chinese Academy of Sciences, and the Chinese Journal "Chinese Cell Biology", are used as a DEMO and different semantic databases as well as different ontologies are used to verify the model built in (5)

In this project, an English publication and a Chinese edition of the Chinese Academy of Sciences are selected as subjects for empirical research. The influence factors of the "Cell Research" in the English version have surpassed that of some internationally renowned academic journals such as "Nature Structural & Molecular Biology" and "Molecular Cell'', where for academic exchanges between Chinese and foreign scientists are very active. The Chinese Journal of "Chinese Cell Biology" is the core Chinese journal of Peking University.
First, based on the Linked Data Conversion Module, the data of the two journals within five years (from 2012 to 2017) are converted into related data, stored in non-relational data. And then based on structural ontology (funds, authors, institutions, domains) and domain ontology (genomic ontology, protein ontology, disease ontology), the full text is reconstructed and knowledge is fine-grained, to form a scientific knowledge map, and then serve researchers.

4 SUMMARY

Linked Open Research Cloud (LORC), as a new knowledge management ecosystem, aims to increase the awareness, discovery, and reuse of online academic resources in the form of open and connected data, laying the foundation for semantic publishing. It's a good solution to the existing problems of academic publishing data warehousing technology, and it will drive transformation and rapid development of semantic publishing mode through structured RDF data, more granular knowledge units, deeper semantic associations and semantic enhancements.

In this project, the data standards, related technologies, and model structures built on the LORC-based semantic publishing platform are studied as the objects., to finally establish a set of structured and easily organized unified data standard based on E-RDF, a group of LD Knowledge Bases and LD Vocabularies, publishing ontology connecting Chinese and English semantic, structured and fine-grained full-text refactoring technology packages and a complete LORC-based semantic publishing model with publishing, integration, and knowledge service capabilities functions which can be promoted in various fields.

Based on the LORC-based data conversion module in the Chinese-English semantic publishing model, the E-RDF-based 5+e star data standard is provided for the semantic publishing industry. Through the conversion in this module, the data has passed 5 star data level, and it has advanced to 6 star and 7 star level. This unified data standard with a global vision enriches the domestic academic publishing technical standards system and lays a foundation for creating a unified and open and globally interconnected semantic publishing model.

The project proposes general data standards, conversion methods, and fusion technologies for the LORC-based Chinese and English semantic publishing model, making it possible to transform science and technology publishing into knowledge service orientation. And through the platform. Semantic and internationalization of domestic publishing and interaction and sharing of data can be achieved, to enable coordinated development of multi-institutions around the world. Study of this project is of a certain significance in demonstrating and promoting the construction of semantic publishing platforms in other fields.

References

[1] Fang Qing et al. Gate of technology has been opened: Analysis of overseas academic publishing hotspots in 2016[J]. Science Technology and Publishing, 2017(2):15-19

[2] NSF. (2017). Resource implementation for data intensive research in the Social, Behavioral and Economic Sciences (RIDIR). https://www.nsf.gov/pubs/2018/nsf18517

[3] Bryant, R., Clements, A., Feltes, C., Groenewegen, D., Huggard, S., Mercer, H., Missingham, R., Oxnam, M., Rauh, A., & Wright, J. (2017). Research Information Management: Defining RIM and the Library's Role. Dublin, OH: OCLC Research. doi:10.25333/C3NK88

[4] Linked Research [EB/OL]. [2018-03-03]. https://linkedresearch.org/

[5] Linked Open Research Cloud [EB/OL]. [2018-03-03]. https://linkedresearch.org/cloud

[6] Qin, Jian. (2017). Transformation from Digital Libraries to Digital Repositories: New Requirements for Digital Scholarship Services in Libraries. Presentation given at the Advanced Digital Library Seminar, Fuzhou,China,Dec.4-5,2017. http://jianqin.metadataetc.org/wp-content/uploads/2017/12/Digital-repository-transition.pdf

[7] The Linking Open Data Cloud Diagram[EB/OL]. [2018-03-03].http://lod-cloud.net/

[8] Linked Open Vocabularies(LOV)[EB/OL]. [2018-03-03].http://lov.okfn.org/dataset/lov/

[9] SPAR-semantic publishing and referencing [EB/OL]. [2018-03-03]. http://sempublishing.sourceforge.net/

[10] Lenmann J. et al. LDOW2017:10th workshop on linked data on the Web [EB/OL]. [2018-03-03]. http://events.linkeddata.org/ldow2017/ldow-10th-workshop.pdf

[11] Springer Nature SciGraph-A Linked Open Data platform for the scholarly domain.[EB/OL]. [2018-03-03].https://www.springernature.com/gp/researchers/scigraph

[12] Linked Brain Data(LBD). [EB/OL]. [2018-03-03].http://www.linked-brain-data.org/

[13] Open Chinese Knowledge Atlas OpenKG. [EB/OL]. [2018-03-03]. http://openkg.cn/home

[14] Li Nan et al. Research on Semantic Publishing Technology for Academic Literature [J] Publishing Science, 2015, 23(6): 85-92

[15] Peng Xiyu et al. Digital development trend of international academic journals [J]. Chinese Sci-tech Journal Studies, 2013, 24(6): 1033-1038