Linking and Consuming Agricultural Big Data with Linked Data and KOS

Guojian Xian, Ruixue Zhao, Xianxue Meng, Yuantao Kou, and Liang Zhu

Agricultural Information Institution of CAAS, Beijing 100081, P.R. China
{xianguojian,zhaoruixue,mengxianxue,kouyuantan,zhuliang}@caas.cn

Abstract. This paper gives brief introduction about the big data, linked data and knowledge organization systems (KOS) and their relationships. As the authors mainly focus on the variety and value characteristics of big data, the linked data and KOS technologies are used to link and consume the large amounts of literature and scientific data in agricultural research community. The results show that it is a good way to describe, connect, organize, represent, visualize and access to big data effectively and semantically based on the linked data and KOS technologies.

Keywords: Big Data, Linked Data, Knowledge Organization System (KOS), Semantic Web, Scientific Data.

1 Introduction

Big data is now one of the hottest topics. Nowadays, we are generating huge amount of data every day, and the total volume of data would be doubled every 18 months, as for the rise of multimedia, social media, and the Internet of Things. It is true that of our activity, innovation, and growth are more and more based on the big data[1].

When talking about big data, people are likely to focus on technical issues, such as scalability, performance and how to deal with large quantities of heterogeneous data, but pay less attention to the connections, interoperability of data in disparate sources, and how to make sense of all the large data pools either. Actually, there are both latent and actual links, which are worth enriching, utilizing and publishing along with the raw dat. In most cases, it is the connections inside and outside of big data where the real value lies.

The formalized, structured and organized nature of linked data and its specific applications, such as the linked knowledge organization systems (KOS), have the potential to provide a solid semantic foundation for the classification, connection, representation, visualization of big data.

The reminder of this paper will firstly give a conceptual analysis of big data, linked data and KOS. And then we will illustrate how to create and consume semantic big data in agricultural research community, utilizing linked data technologies and knowledge organization systems as new tools for the describing, linking, organization, representation, visualization and access to big data.

D. Li and Y. Chen (Eds.): CCTA 2014, IFIP AICT 452, pp. 546–555, 2015.
DOI: 10.1007/978-3-319-19620-6_61

2 Big Data, Linked Data and KOS

Big data always refers to large, diverse, complex, longitudinal, distributed data sets generated from instruments, sensors, internet transactions, email, video, click streams, and other digital sources. While there is no generally agreed understanding of what exactly is big data, an increasing number of V's has been used to characterize different dimensions and challenges of big data: volume, velocity, variety, value, and veracity[2, 10]. These terms means to the growing volume of different types of structured and unstructured data, the complex and heterogeneous nature and machine-processability. The organization, exploration, management, preservation, visualization and access to and use of these types of data pose technological and computational challenges.

Linked Data defines a set of guidelines, best practices or patterns for exposing, publishing, sharing, and connecting pieces of data, information, and knowledge on the Semantic Web using URIs and RDF[3, 11], which takes the WWW's ideas of global identifiers and links and applies them to (raw) data, not just documents[12-13]. Linked data can be effectively used as a broker, mapping and interconnecting, indexing and feeding real-time information from a variety of sources, and inferring relationships from big data analysis that might otherwise have been discarded, which is made all the more valuable by the connections (links) that tie it all together than the sum of its volume, velocity and variety.

Semantic text analysis, natural language processing, data mining and data visualization are the typical challenges in addressing the management and effective use of big data[4]. The knowledge organization systems, such as thesauri, classifications, subject headings, taxonomies, and folksonomies would increasingly important roles. As W3C's standard, the Simple Knowledge Organization System (SKOS) standard, aims to build a bridge between the world of KOS and the linked data community[17]. SKOS-based linked controlled vocabularies can provide a framework rich in semantics to effectively manage big data through combining, aligning and cross-linking multiple KOSs in order to automatic or semi-automatic analyzing, indexing and organizing text, and develop faceted, categorized or hierarchical views of big data[5].

A typical and successful example of combining the big data, linked data and KOS together is the new semantic web platform version of Agris: OpenAgris[6]. Based on nearly 4 million structured bibliographical records on agricultural science and technology, and AGROVOC vocabularies, alignment between KOS and ontologies, the OpenAgris also aggregates various open access data sources available on the Web, providing much data as possible about a topic or a bibliographical resource[7, 20].

3 Linking Agricultural Big Data

The 4th paradigm of science – data intensive scientific discovery has emerged within the last years, which means scientific innovations and breakthroughs will be powered by advanced computing capabilities that help researchers manipulate and explore massive datasets[8]. As for the several V's characteristics of big data, different disciplines highlight certain dimensions and neglect others. Fox example, people

working on sensor and the internet of things may care more about the velocity, super-computing would be mostly interested in the volume dimension, and the research community pay more attentions to variety and value dimensions of big data.

The Chinese governments always pay high attentions to the agricultural science and technology innovation and the development of modern agriculture. With more investments are input to this sector, a large amount of agricultural big datasets are produced, such as the 3S data, scientific research data, and academic achievements (e.g. papers, books, proceedings, reports). These data have different formats, conceptualizations or data models, temporal and spatial dependencies[9]. Effective usage of these big data is very important to promote and advance the research work in a new round.

The following parts of this paper will explain how to describe, organize, integrate, and consume parts of these agricultural big data with linked data and KOS technologies, as we just focus on the variety and value dimensions of big data.

3.1 Linking Chinese Agricultural Thesaurus to Other KOS

As KOS can play important role to addressing some challenges of the big data, such as text analysis, natural language processing, data mining and data visualization of big data, the priority of our work is to convert the Chinese Agricultural Thesaurus (CAT) as linked data and link to other KOS.

CAT was developed by Agricultural Information Institution of CAAS in early 1990s and kept maintenance all the time. So far, CAT contains more than 60,000 Chinese descriptors and non-descriptors, most have corresponding English translations, and also include about 130,000 semantic relationships, such as UF, BT, NT and RT.

We describe CAT's concepts and their semantic relationships with SKOS and SKOS-XL standards. In addition, we map and link CAT to other well-known KOS such as AGROVOC, NALT, EUROVOC and LCSH, as shown in Fig.1.

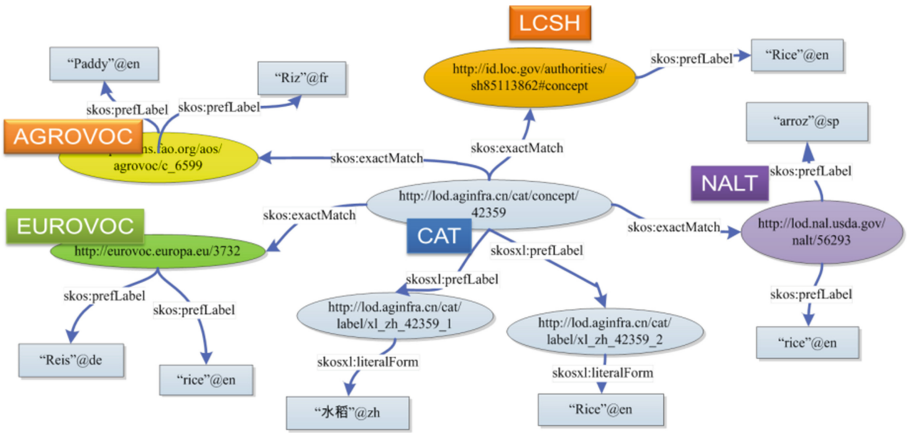

Fig. 1. The SKOS Model of CAT and Link to Other KOSs

We also develop a SKOS-based CAT linked data web system, providing services such as HTTP URI dereference, CAT concepts browsing and navigation, SPARQL query endpoint and RDF Triples Dumps, as shown in Fig.2. This work could greatly improve the CAT's visibility, accessibility and interoperability with other systems, and lay fundamental base to describe, organize and link other agricultural information resources semantically as well.

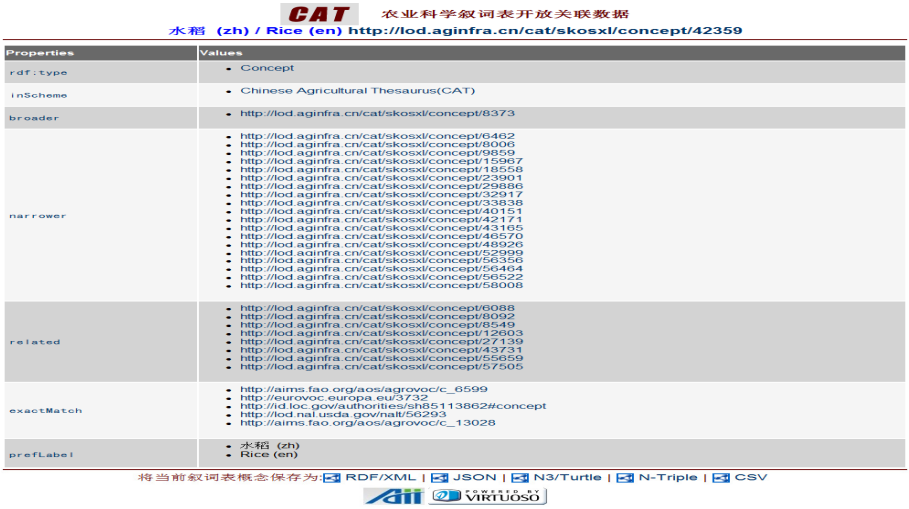

Fig. 2. The Linked Open data of SKOS-Based CAT

3.2 Linking Agricultural Literature and Scientific Research Data

As the current global research data is highly fragmented, by disciplines or by domains, from oceanography, life sciences and health, to agriculture, space and climate[21]. When it comes to cross-disciplinary activities, building specific "data bridges" are becoming accepted metaphors for approaching the data complexity and enable data sharing.

The millions of books, journals, proceedings and bibliographic records of the China National Agricultural Library, and also over 700 scientific datasets holds in the National Agricultural Scientific Data Sharing Platform, are the most valuable data materials for agricultural research communities. What we want to do most here is connect these literature and research data together in a light-weighted semantically way. Fig.3 shows the available and linkable data resources we could access to.

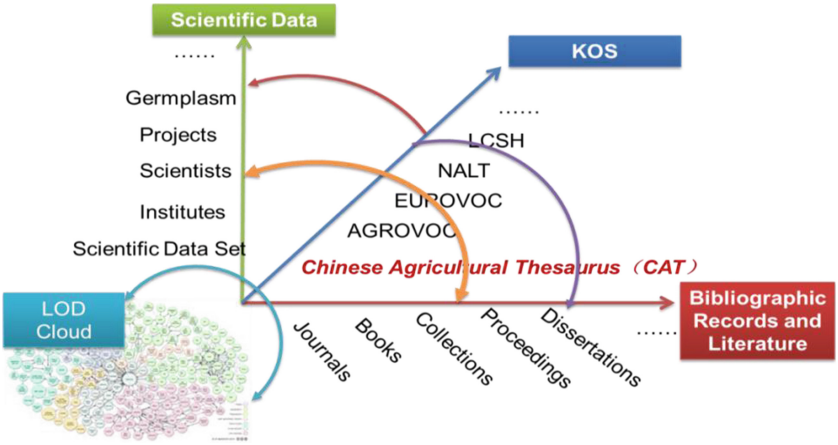

Fig. 3. Available and Linkable Agricultural Data Resources

We analyze and abstract the main classes and properties from the literature and bibliographic records of the China National Agricultural Library, and also reuse the widely used vocabularies and ontologies such as DCMI[15], BIBO[16], etc., to formally describe and model these classes, properties and their semantic relationships. The following table shows the mapping result of the journal article (bibo:AcademicArticle) and its properties to common vocabularies and ontologies.

Table 1. The Core Properties of Journal Article

Property	Type	Available Vacobulary
title	DataProperty	dc:title、swrc:title
alternateTitle	DataProperty	dcterms:alternative、prism:alternateTitle
author	ObjectProperty	dc:creator、foaf:maker、swrc:creator
keywords	ObjectProperty/ DataProperty	dc:subject、swrc:keywords、prism:keyword
abstract	DataProperty	bibo:abstract、dcterms:abstract、swrc:abstract
language	DataProperty	dc:language、swrc:language
startPage	DataProperty	bibo:pageStart、prism:startingPage
endPage	DataProperty	bibo:pageEnd、prism:endingPage
totalPage	DataProperty	bibo:numPages、prism:pageCount
DOI	ObjectProperty	bibo:doi、prism:doi

We also describe and connect the widely used literature, such as books, journals, collections, proceedings together, and the concepts and semantic relationships of CAT are taken into consideration while designing the describing and linking model (Fig. 4).

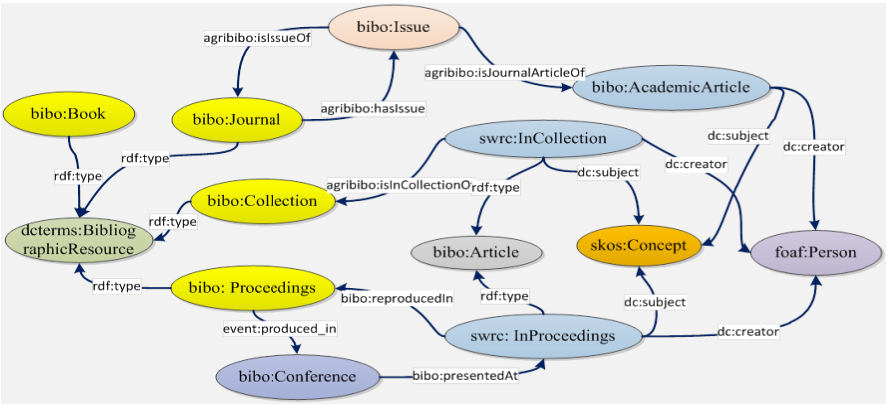

Fig. 4. Bibliographic Records and Literature Linking Model

The most interested and meaningful work we done is integrating and linking the SKOS-based CAT and several kinds of literature to the scientific research datasets. So far, we have modeled and linked about 700 core metadata of the scientific research datasets, and also some particular datasets hold the data about the institutes, researchers and projects of agricultural related domain, by reusing the well-known vocabularies or ontologies (e.g. VIVO, SWRC, FOAF)[19].The multidimensional semantic linking model covering the scientific data, literatures, and thesaurus we designed is as shown in Fig 5.

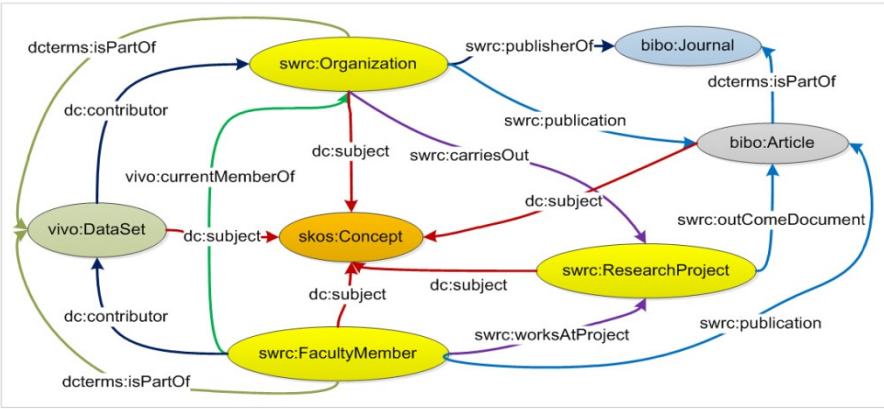

Fig. 5. Multidimensional Linking Model Covering the Scientific Data, Literatures and KOS

In order to process these resource based on the SKOS-based CAT, we developed an automatic text analyzing and indexing tool based on some open source tools, such as Lucene, IKAnalyzer, etc. The tool we realized could tag the concepts and semantic links of CAT into these literatures and scientific datasets. That is a very important step to make semantic connections between several kinds of data from different sources, with the professional knowledge of KOS.

As for most of the data resources mentioned above are stored in relational database, so we construct and publish the semantic agricultural linked data based on the open source software D2R Server[14].

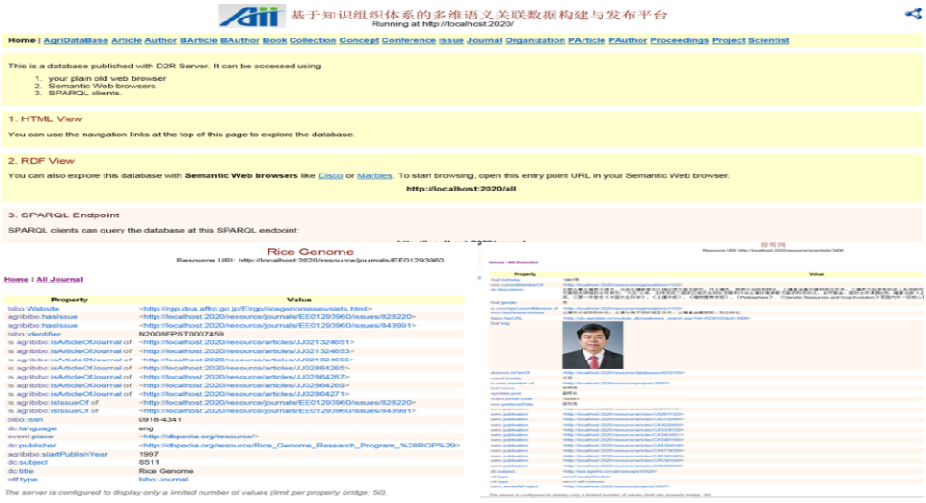

Fig. 6. The Linked Data Publishing Platform Based on D2R

4 Consuming Linked Agricultural Big Data

One effective way to explore and mine the linked big data is following the linked data rules and utilizing some semantic web technologies, here we do not care too much about the volume of big data and neither the computing performance. We pay more attentions to the variety and value of the big data.

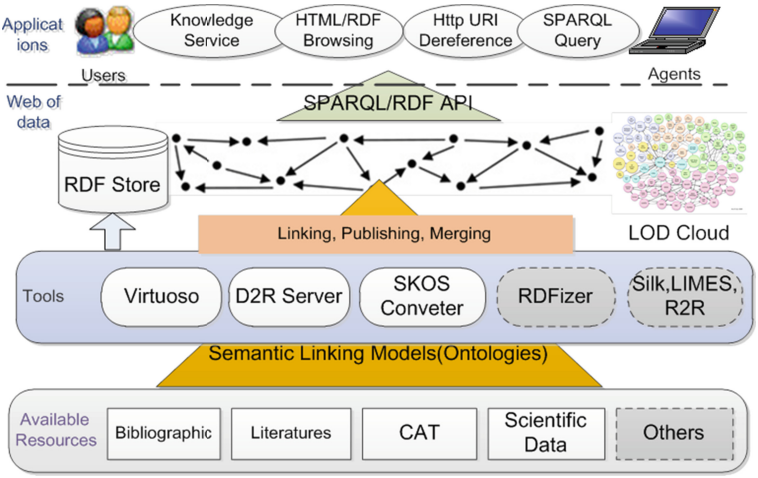

Fig. 7. The Architecture of Knowledge Service System Driven by Linked Big Data

The architecture and function models of domain knowledge service system have been designed, which totally driven by the linked data, as shown in Fig. 7. A prototype system was realized based on some key technologies such as SPARQL, Virtuoso[17] and so forth, see Fig.8. Some import service functions have been provided in this system, such as integrated browsing and discovery of domain knowledge, dynamic facet navigation and searching, SPARQL query endpoint, HTTP URI dereferencing, downloading RDF triples.

Fig. 8. The Domain Knowledge Service Prototype System Driven by Linked Data

5 Conclusions

The results of this study proves that it is one of the best practices to applying the ideas, principles and methodologies of linked data, to describe, organize and merge the huge amount of agricultural information resources in a more fine described, formally structured and semantically linked way. Linked data would play a great role to increase the popularity, visibility, accessibility and value of agricultural big data resources. The knowledge organization systems in SKOS formats would help us to align and match concepts to develop a broad and high level analytical framework for managing, representing and mining big data.

Further research work are needed because we at present just now focus on the variety and value dimensions of big data, we should actually address the volume, velocity issues in practice based on the cloud computing and other efficient big data infrastructure and technologies.

As massive amounts of data are available, linked and identifiable via URIs, big data, linked data and KOS would be an integral part of the future web infrastructure, the web of data, global data space and semantic web is beginning to take shape.

Acknowledgements. The related research work of this paper was support by the National Key Technology R&D Program "Knowledge Services Application and Demonstration Based on STKOS" (2011BAH10B06) during the Twelfth Five-year Plan Period.

References

1. Manyika, J., Chui, M., Brown, B., et al.: Big data: The next frontier for innovation, competition, and productivity[R/OL](May 01, 2011), http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation (June 08, 2014)
2. Hitzle, P., Janowicz, K.: Linked Data, Big Data, and the 4th Paradigm. Semantic Web 4(3), 233–235 (2013)
3. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space, 1st edn. Synthesis Lectures on the Semantic Web: Theory and Technology, vol. 1(1), pp. 1–136. Morgan & Claypool (2011)
4. Suchanek, F., Weikum, G.: Knowledge harvesting in the big-data era. In: Proceedings of the 2013 International Conference on Management of Data, pp. 933–938. ACM (2013)
5. Shiri, A.: Linked Data Meets Big Data: A Knowledge Organization Systems Perspective. Advances In Classification Research Online 24(1) (2014), doi:10.7152/acro.v24i1.14672
6. Anibaldi, S., Jaques, Y., Celli, F., et al.: Migrating bibliographic datasets to the Semantic Web: The AGRIS case (2013)
7. Celli, F., Jaques, Y., Anibaldi, S., Keizer, J.: Pushing, Pulling, Harvesting, Linking: Rethinking Bibliographic Workflows for the Semantic Web. In: EFITA-WCCA-CIGR Conference "Sustainable Agriculture through ICT Innovation", Turin, Italy, June 24-27 (2013)
8. Hey, A.J., Tansley, S., Tolle, K.M., et al.: The fourth paradigm: data-intensive scientific discovery. Microsoft Research Redmond, WA (2009)
9. Janowicz, K., Hitzler, P.: The digital earth as knowledge engine. Semantic Web 3(3), 213–221 (2012)
10. Bizer, C., Boncz, P.A., Brodie, M.L., et al.: The Meaningful Use of Big Data: Four Perspectives-Four challenges. SIGMOD Record 40(4), 56–60 (2011)
11. Bizer, C., Heath, T., Lee, T.B.: Linked Data - The Story So Far. International Journal on Semantic Web & Information Systems 5(3), 1–22 (2009)
12. Linked Data - Connect Distributed Data across the Web [EB/OL] (June 18, 2012), http://linkeddata.org/ (May 28, 2014)
13. Mike. The Rise of the Data Web [EB/OL] (June 18, 2012), http://www.dataspora.com/2009/08/the-rise-of-the-data-web/ (May 28, 2014)
14. D2R Server-Publishing Relational Databases on the Semantic Web. [EB/OL] (February 16, 2010), http://www4.wiwiss.fu-berlin.de/bizer/d2r-server/ (June 15, 2014)

15. DCMI Metadata Terms [EB/OL] (June 14, 2012), http://dublincore.org/documents/2012/06/14/dcmi-terms/ (February 03, 2013)
16. Darcus, B., Giasson, F.: Bibliographic ontology specification [EB/OL] (November 04, 2009), http://purl.org/ontology/bibo/ (June 12, 2014)
17. Dolan-Gavitt, B., Leek, T., Zhivich, M., et al.: Virtuoso: Narrowing the Semantic Gap in Virtual Machine Introspection. In: 2011 IEEE Symposium on Security and Privacy (SP), pp. 297–312. IEEE, Berkeley (2011)
18. Isaac, A., Summers, E.: SKOS simple knowledge organization system primer [EB/OL] (February 21, 2008), http://www.w3.org/TR/skos-primer/ (June 04, 2014)
19. Sure, Y., Bloehdorn, S., Haase, P., Hartmann, J., Oberle, D.: The SWRC ontology - Semantic Web for research communities. In: Bento, C., Cardoso, A., Dias, G. (eds.) EPIA 2005. LNCS (LNAI), vol. 3808, pp. 218–231. Springer, Heidelberg (2005)
20. AGROVOC Linked Open Data [EB/OL] (March 20, 2013), http://aims.fao.org/standards/agrovoc/linked-open-data (June 09, 2014)
21. Research Data Alliance [EB/OL] (June 20, 2014), https://rd-alliance.org/ (June 22, 2014)