Shiri, A. (2014). Linked Data Meets Big Data: A Knowledge Organization Systems Perspective.
Advances In Classification Research Online, 24(1). doi:10.7152/acro.v24i1.14672

Linked Data Meets Big Data: A Knowledge Organization Systems Perspective

Ali Shiri

School of Library and Information Studies

University of Alberta, Edmonton, Alberta

ashiri@ualberta.ca

ABSTRACT

The objective of this paper is a) to provide a conceptual analysis of the term big data and b) to introduce linked data applications such as SKOS-based knowledge organization systems as new tools for the analysis, organization, representation, visualization and access to big data.

Keywords

Big data, linked data, Knowledge organization systems, SKOS.

1. INTRODUCTION

The vast volume, variety and complexity of digital data available on the web has resulted in the emergence of what is called 'Big Data'. There are many different terms used in the literature that may refer to or be associated with the phenomenon of 'big data', including such terms as digital data, research data, linked data, open data, web of data and data repositories (Lyon, 2007; Research Data Strategy Working Group, 2011; Borgman et al., 2012; Hodson, 2012;; National Science Foundation, 2012). Examples of big data may include social media data, e-business data, linked data, web usage data, government open data, funded research data, as well as the data created in such diverse contexts as cloud-based computing infrastructures, virtual collaboratoies, e-science, e-humanities and e-social sciences projects. A facet analysis approach to the conceptualization of big data and its various aspects has previously been proposed (Shiri, 2012).

Advances in Classification Research, 2013, November 2, 2013, Montreal, Canada.

2. BIG DATA: DEFINITIONS

A number of definitions have been proposed for big data in the literature. Because of the multifaceted nature of this phenomenon, scientists, information technology managers, information scientists, policy makers and funding agencies have approached it from various perspectives. This is, in part, due to the vague nature of the term big data and what it means to people from various educational and occupational backgrounds. For instance, the National Science Foundation report on *Long-lived Digital Data Collections* (2005) avoids using the term 'big'. Rather it focuses on the longevity and proper management of digital data. The report defines digital data as follows:

> "The term 'data' is used in this report to refer to any information that can be stored in digital form, including text, numbers, images, video or movies, audio, software, algorithms, equations, animations, models, simulations, etc. Such data may be generated by various means including observation, computation, or experiment."

This definition has a clear focus on demonstrating the vast variety of data, its origins and the associated techniques for its analysis and maintenance.

A more technologically and industrially focused definition is offered by Kusnetzky (2011) who defines big data below:

> "In simplest terms, the phrase **[**big data**]** refers to the tools, processes and procedures allowing an organization to create, manipulate, and manage very large data sets and storage facilities."

This definition takes a more pragmatic approach to big data and places emphasis on the volume of data and the challenge of its technical management.

Jacobs (2009) approaches big data from a database technology perspective and notes that the fact that most large datasets have inherent temporal or spatial dimensions, or both, is crucial to understanding one important way that big data can cause performance problems, especially when databases are involved. His meta-definition of big data

ISSN: 2324-9773

stresses the significance of temporal data as a key factor and believes that big data should be defined at any point in time as "data whose size forces us to look beyond the tried-and-true methods that are prevalent at that time." In today's world, it may mean that data is too large to be placed in a relational database and analyzed with the help of a desktop statistics/visualization package—data, perhaps, whose analysis requires massively parallel software running on tens, hundreds, or even thousands of servers.

In line with technological approaches to big data, Warden (2011) provides a particularly useful glossary of big data that provides a listing and description of 60 most recent technological innovations in the area of big data that can help those working with large data sets navigate the large number of new data tools available. These technologies vary from noSQL databases, MapReduce, storage and servers to natural language processing, machine learning, acquisition and visualization.

In the context of the sciences, Borgman (2007) makes use of the term 'data deluge' and refers to the variety of data created, ranging from laboratory and field notebooks, slides from talks and composite objects to graphic visualizations of data. Examples of data in the science may include X-rays, protein structures, spectral surveys, specimens and events and objects. She argues that it is difficult to separate data from software, equipment, documentation and knowledge required to use them. This observation points to the challenges of defining data and data carriers. Borgman (2007) also provides a categorization for the types of data created by social scientists. The first category is data collected by researchers through experiments, interviews, surveys, observations. The second type is data that is collected by other people or institutions usually for purposes other than research. These include government and institutional data such as census figures, economic indicators, demographics and other public records. Other data sources such as mass media content and records of corporations, she notes, can be useful sources of social science data. She suggests that in the area of humanities the distinction between documents and data is the least clear due to the fact that almost any document, physical artifact and any record of human activity can be used to study culture.

Bizer et al. (2011) argue for the meaningful and semantic use and applications of big data by providing four challenges, namely a) the fact that big data integration is multidisciplinary b) web of data and structured data as part of big data faces processing and integration challenges c) lack of good use cases to provide the opportunity for experimenting with open linked data on the Web and d) demonstrating the value of semantics in data linking and integration.

The idea behind "digging into data challenge", an international grant competition, "was to address how "big data" changes the research landscape for the humanities and social sciences. Now that we have massive databases of materials used by scholars in the humanities and social sciences -- ranging from digitized books, newspapers, and music to transactional data like web searches, sensor data or cell phone records -- what new, computationally-based research methods might we apply? As the world becomes increasingly digital, new techniques will be needed to search, analyze, and understand these everyday materials. 'Digging into Data' initiative challenges the research community to help create the new research infrastructure for 21st century scholarship".

Simon Hodson, the Research manager for JISC Digital infrastructure names a number of areas that deal with the big data issue, namely web archiving, learning analytics, usage statistics and research data. In line with big data in the context of social sciences, JISC has sponsored a project to be conducted by the Oxford Internet Institute titled Big Data: Demonstrating the Value of the UK Web Domain Dataset for Social Science Research, which aims to enhance JISC's UK Web Domain archive, a 30 terabyte archive of the .uk country-code top level domain collected from 1996 to 2010. It will extract link graphs from the data and disseminate social science research using the collection. In his final remarks for the Eduserv Symposium 2012: Big Data, Big Deal? held in London, UK in May 2012, Powell (2012) suggests that there seems to be confusion about open data and big data and that there is a potential confusion between big data and data that happens to be big. He notes that open data is considered to be big data. He also suggests that we need to think carefully about the kinds of questions we need to ask when deal with big data. Hitzler and Janowicz (2013) note: "…it appears to be uncontroversial that Linked Data is part of the Big Data landscape. We would even go a bit further and claim that Linked Data is an ideal testbed for researching some key Big Data challenges."

The National Science Foundation and the National Institutes of Health *Core Techniques and Technologies for Advancing Big Data Science & Engineering (BIGDATA)* solicitation states the aim of big data as:

> " to advance the core scientific and technological means of managing, analyzing, visualizing, and extracting useful information from large, diverse, distributed and heterogeneous data sets so as to: accelerate the progress of scientific discovery and innovation; lead to new fields of inquiry that would not otherwise be possible; encourage the development of new data analytic tools and algorithms; facilitate scalable, accessible, and sustainable data infrastructure; increase understanding of human and social processes and interactions; and promote economic growth and

ISSN: 2324-9773

improved health and quality of life. “

These two organizations emphasize that big data does not exclusively refer to the volume of the data, but also to its variety and velocity. They note that “Big data includes large, diverse, complex, longitudinal, and/or distributed data sets generated from instruments, sensors, Internet transactions, email, video, click streams, and/or all other digital sources. “

A review of above definitions and characteristics of big data demonstrates the complexity and variety of concepts and terms used to identify what constitutes big data. One could argue that research data, open data, linked data and semantic web data can be construed as part of big data. These terms refer to the growing volume of different types of structured and unstructured data, their complex and heterogeneous nature and machine-processability and the challenges they pose for creators and users of big data. The organization, curation, exploration, management, preservation, visualization and access to and use of these types of data pose similar technological and computational challenges.

3. LINKED DATA

To contextualize linked data in relation to big data, one can argue that liked data in itself may be viewed as one major type of data in the universe of big data. The main argument of this paper is that the formalized, structured and organized nature of linked data and its specific applications, such as linked controlled vocabularies and knowledge organization systems, have the potential to provide a solid semantic foundation for the classification, representation, visualization and the organized presentation of big data. Some of the key advantages of linked knowledge organization systems may include their utilization in automatic or semi-automatic analysis of text, assignment of subject metadata and the development of faceted, categorized or hierarchical views of content.

4. LINKED VOCABULARIES FOR DATA MANAGEMENT AND ACCESS

One of the key challenges in addressing the management and effective use of big data is associated with text and natural language. More specifically, big data analytics is one of the areas of growing importance as it provides ways in which one can make sense and effective use of data (Russom, 2011). Among the key areas of analytics technologies are semantic text analysis, natural language processing, data mining and data visualization. Natural language analysis and processing is considered to be particularly useful for supporting semantic search within the context of big data (Suchanek & Weikum, 2013). The Simple Knowledge Organization System (SKOS) standard, developed by the World Wide Web Consortium, aims to build a bridge between the world of knowledge organization systems—including thesauri, classifications, subject headings, taxonomies, and folksonomies—and the linked data community, with the goal of bringing benefits to both. SKOS-based linked data controlled vocabularies can provide a semantically rich framework for the analysis and visualization of big data.

The SKOS standard provides a framework for mapping and connecting multiple thesauri and other types of controlled vocabularies in order to create cross-browsing and cross-searching applications for linked data repositories, open archives, digital libraries, and various search systems and services. Currently, a number of controlled vocabularies have been encoded in the SKOS format that can be used as linked data sources. Examples include: GEneral Multilingual Environmental Thesaurus (GEMET), AGROVOC, MESH, Helping Interdisciplinary Vocabulary Engineering (HIVE), Library of Congress Subject Headings (LCSH), Thesaurus of Economics, and Schools Online Thesaurus. There are a number of web services that are compliant with RDF and linked data standards that offer new opportunities for addressing some of the big data challenges. Zemanta is a service that uses RDF and URIs to deliver relevant tags, links and pictures from unstructured content. It also provides disambiguation features for entities using open linked data sources. CALAIS, provided by Thomson Reuters, is another example of a web service that analyzes unstructured texts and returns RDF formatted results. Using natural language processing and machine learning, it categorizes and links documents with entities. It has a RDF creator service that is used for tagging of blog articles and for the organization of museum collections. A good example of a SKOS-based linked data thesaurus is PoolParty (Schandl and Blumauer 2010), which is capable of linking to many different linked data repositories to facilitate search and retrieval. Méndez and Greenberg (2012) note that vocabularies in their many forms (thesauri, taxonomy, ontology, and discipline, domain and community languages) can be leveraged and made more powerful via RDF/SKOS. Linked Open Vocabularies (LOV) developed and created by Bernard Vatant and Pierre-Yves Vandenbussche is a particularly useful registry of linked data vocabularies that could be used for big data processing and access.

There are a growing number of linked data sources that allow various data providers and publishers to incorporate their linked data resources in semantic web-enabled repositories. General purpose and domain-specific controlled vocabularies published as linked open vocabularies can not only be used to organize and represent structured data such as linked data repositories and semantic web applications, they can also be used to index, organize and analyze unstructured textual information that exists in several big data sources. SKOS-based linked data thesauri can be used to effectively manage, big data through combining, aligning and cross-linking multiple knowledge

ISSN: 2324-9773

organization systems in order to access, index, organize and retrieve big data.

5. USE CASES

Science:

- Agricultural Information Management – using linked data across various Agricultural Informatin Management Standards (AIMS) to improve information sharing within a worldwide network of research institutes, non-profit organization, government and non-government agencies. Example: Using basic descriptive metadata, thesauri, AGROVOC vocabularies, and ontologies, the Food and Agriculture organization (FAO) continually collects, organizes and disseminates information on nutrition, food and agriculture in many formats with the goal to fight world hunger (Baker & Keizer, 2010).

- Personalized healthcare: processing large quantities of patient records against semantically-enabled ontologies built specifically for healthcare to look for patterns, help doctors make decisions and create individualized disease risk profiles. Using collaborative filtering methods used in other settings by recommender systems (Amazon, Netflicks, Lastfm), and combined with formalized SKOS vocabularies and thesauri, we can apply several information management strategies to large banks of health data for more powerful analysis (Chawla & Davis, 2013).

Humanities:

- DPLA and Europeana are both examples of Big Data in the Arts and Humanities in that they function as data banks (through open API services that they provide) and as large data-processing organization because of their carefully planned and constantly-improving data model. Both are keenly aware of the role of metadata, and specifically, semantically-oriented metadata to allow users more intuitive, smarter access to their digital collections. (Mitchell, 2013)

6. APPLICATIONS OF LINKED DATA VOCABULARIES FOR BIG DATA

Some of the advantages and applications of SKOS-based linked open vocabularies for big data management and access include:

- Support for cross-searching and cross-browsing of open data repositories and big data sources
- Connecting linked vocabularies to structured, semi-structured or unstructured data
- General-purpose and domain-specific natural language processing of big data
- Visualization of subject metadata
- Development of semantically-enhanced search engines and recommender systems
- Term/ query auto-completion mechanism s
- Interactive and automatic query expansion
- Development of faceted and categorized browsing
- Analysis and processing of text in digital humanities projects

Figure 1 shows the potential applications of SKOS-based vocabularies for addressing the challenges associated with the organization, management and representation of big data.

7. CONCLUSION

Following the emergence of search engines, digital libraries and various types of information repositories in the 1990s and 2000s, big data is gradually finding its way into our new digital information environment. SKOS is a vocabulary model to represent knowledge organization systems on the semantic web in a simple way. Networking controlled vocabularies in SKOS allows us to align and match concepts to develop a broad and high level analytical framework for managing and representing big data. SKOS and linked data vocabularies provide new opportunities and approaches for big data analysis, management, representation, and visualization. The complex and vague nature of big data poses challenges for information and computer science research as well as for knowledge organization scholars. Big data organization, representation and visualization will be among the emerging areas or research that information organization research will have to address.

REFERENCES

Baker, T., & Keizer, J. (2010). Linked Data for Fighting Global Hunger: Experiences in setting standards for Agricultural Information Management. Linking Enterprise Data, 177. doi:10.1007/978-1-4419-7665-9_9

Bizer, C., Boncz, P. A., Brodie, M. L., Erling, O. (2011) The Meaningful Use of Big Data: Four Perspectives - Four challenges. SIGMOD Record 40(4): 56-60.

Borgman, C. (2007). Scholarship in the Digital Age: Information, Infrastructure, and the Internet. Cambridge, MA: MIT Press.

Borgman, C. L. (2012). The conundrum of sharing research data. Journal of the American Society for Information Science and Technology, 63(6), 1059-1078.

CALAIS: http://www.opencalais.com/

Chawla, N. V., & Davis, D. A. (2013). Bringing big data to personalized healthcare: A patient-centered framework.

Journal of general internal medicine, 28(3), 660-665.

Hitzler, P., & Janowicz, K. (2013). Linked Data, Big Data, and the 4th Paradigm. *Semantic Web*, 4(3), 233 -235.

Hodson, S. (2012). JISC and Big Data. Eduserv Symposium 2012: Big Data, Big Deal? May 10, 2012, London, UK.

Kuznetsky, D. (2010) What is "Big Data?"ZDNet February 16, 2010. Accessed July 18, 2012: http://www.zdnet.com/blog/virtualization/what-is-big-data/1708

Jacobs, A. (2009) The Pathologies of Big Data. Queue 7(6), pp. 1-12.

Lyon, L. (2007). Dealing with Data: Roles, Rights, Responsibilities and Relationships. Consultancy Report. June 19, 2007. Accessed July 18, 2013: http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/dealing_with_data_report-final.pdf

Méndez, E., & Greenberg, J. (2012). Linked data for open vocabularies and HIVE's global framework. El profesional de la información, 21(3), 236-244.

National Science Foundation. (2012). Core Techniques and Technologies for Advancing Big Data Science & Engineering (BIGDATA) Solicitation. Accessed July 18, 2013: http://grants.nih.gov/grants/guide/notice-files/NOT-GM-12-109.html

Mitchell, E. T. (2013). Three Case Studies in Linked Open Data. Library Technology Reports, 49(5), 26-43.

Powell, A. (2012) Final Remarks. Eduserv Symposium 2012: Big Data, Big Deal? May 10, 2012, London, UK.

Research Data Strategy Working Group. (2011). Mapping the Data Landscape: Report of the 2011 Canadian Research Data Summit. December 2011. http://rds-sdr.cisti-icist.nrc-cnrc.gc.ca/eng/events/data_summit_2011/index.html

Russom, P. (2011). Big data analytics. TDWI Best Practices Report, Fourth Quarter.

Schandl, T., & Blumauer, A. (2010). PoolParty: SKOS thesaurus management utilizing linked data. In The Semantic Web: Research and Applications (pp. 421-425). Springer Berlin Heidelberg.

Schools Online Thesaurus: http://scot.curriculum.edu.au/

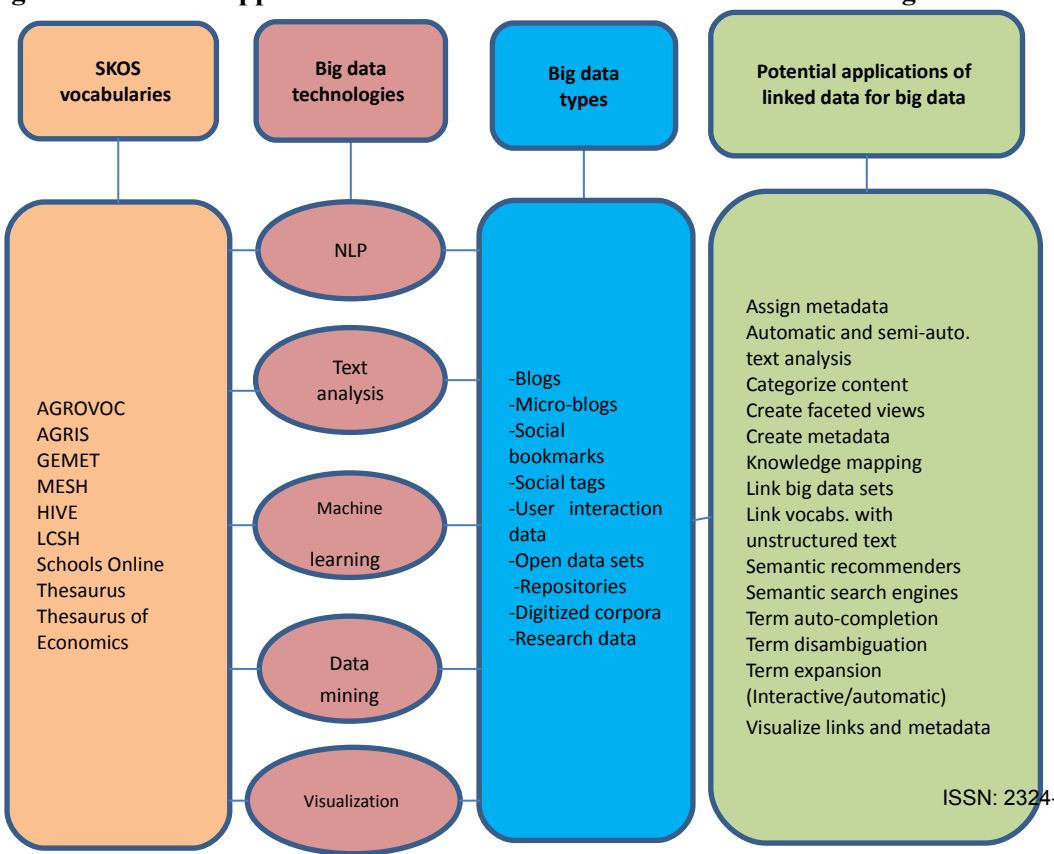
Shiri, Ali (2012). Typology and Analysis of Big Data: An Information Science Perspective. Presented at the Internet, Politics, Policy 2012: Big Data, Big Challenges? Oxford Internet Institute. University of Oxford, September 20-21, 2012, St. Anne's College, Oxford, UK.

Suchanek, F., & Weikum, G. (2013). Knowledge harvesting in the big-data era. In Proceedings of the 2013 international conference on Management of data (pp. 933-938). ACM.

Warden, P. (2011) Big Data Glossary. Sebastopol, CA : O'Reilly.

Zemanta: http://www.zemanta.com/

Figure 1. Potential applications of SKOS-based linked vocabularies for big data



ISSN: 2324-9773