CLASSIFYING THE LOD CLOUD: DIGGING INTO THE KNOWLEDGE GRAPH

Daniel Martínez Ávila (1) Richard P. Smiraglia (2) Rick Szostak (3) Andrea Scharnhorst (4) Wouter Beek (5) Ronald Siebes (6) Laura Ridenour (7) Vanessa Schlais (8)

(1) São Paulo State University (UNESP), Brazil, martinez.avila@unesp.br. (2) School of Information Studies, University of Wisconsin-Milwaukee, USA, smiragli@uwm.edu (3) University of Alberta, Canada, rick.szostak@ualberta.ca (4) Data Archiving and Networked Services, Royal Netherlands Academy of the Arts and Sciences, The Hague, The Netherlands, andrea.scharnhorst@dans.knaw.nl (5) Vrije Universiteit, Amsterdam, The Netherlands, w.g.j.beek@vu.nl (6) Data Archiving and Networked Services, Royal Netherlands Academy of the Arts and Sciences, The Hague, The Netherlands, ronald.siebes@dans.knaw.nl (7) School of Information Studies, University of Wisconsin-Milwaukee, USA, ridenour@uwm.edu (8), vschlais@uwm.edu

Abstract

Massive amounts of data from different contexts and producers are collected and connected relying often solely on statistical techniques. Problems to the acclaimed value of data lie in the precise definition of data and associated contexts as well as the problem that data are not always published in meaningful and open ways. The Linked Data paradigm offers a solution to the limitations of simple keywords by having unique, resolvable and shared identifiers instead of strings This paper reports on a three-year research project "Digging Into the Knowledge Graph," funded as part of the 2016 Round Four Digging Into Data Challenge (https://diggingintodata.org/awards/2016/project/digging-knowledge-graph). Our project involves comparing terminology employed within the LOD cloud with terminology employed within two general but different KOSs – Universal Decimal Classification and Basic Concepts Classification. We are exploring whether these classifications can encourage greater consistency in LOD terminology and linking the largely distinct scholarly literatures that address LOD and KOSs. Our project is an attempt to connect the Linked Open Data community, which has tended to be centered in computer science, and the KO community, with members from linguistics, metaphysics, library and information science. We focus on the shared challenges related to Big Data between both communities.

Keywords: Linked Open Data; Knowledge Organisation Systems; Big Data; Knowledge Graph

1 Introduction

In the age of big data, massive amounts of data from different contexts and producers are collected and connected relying often solely on statistical techniques. There are different problems to the acclaimed value of data. First, what are data and what value data have is context dependent, and cannot be defined in an abstract, generic way (Borgman, 2015). Still, data are automatically extracted, indexed and published on the web, and correspondingly data search engines are developed treating data as defined elements in scholarly communication (Gregory et al., 2018). Which brings us to the second problem, that data are not always published in meaningful and open ways, so their interconnections have been often related to commercial and private interests (forgetting public interests and the value of data for science and education). The concept of Linked Open Data takes the issue of data to another level. The Linked Data paradigm offers a solution to the limitations of simple keywords, like homonyms, synonyms, spelling mistakes, language variations, and the meaning of unknown terms by adding relational information connecting it to known terms/concepts/classes. By having unique, resolvable and shared identifiers instead of strings, many of the difficult problems of mapping, understanding and querying are reduced, if not solved. Instead of indexing web resources (documents), now the content within the web resources is indexed (Berners-Lee et al., 2001). Publishing data in a Linked Open Data (LOD) format is a big step forward to free data from data silos and make them available to be further interlinked and so enriched. What is often less addressed in this discourse, is that publishing data as LOD is only the necessary first, but not yet the sufficient step towards data creating meaning by creating context by creating links. The philosophy of the semantic web, based on the creation of meaningful machine-readable data by different communities, has led currently to isolated information systems, which come with their own domain-specific knowledge organization systems, limiting potential interoperability. They can be in principle linked - the technology for that is there but linking them requires expert knowledge (supported by machines) and does not happen automatically. The Linked Open Data cloud requires interoperable vocabularies scaled up for better organization of large data clusters. The promise of the web-based Linked Open Data (LOD) Cloud is to free up data, metadata and information to a large extent

from prior "data silos" (meaning database systems). The LOD Cloud promises to deliver machine-readable Knowledge Organization Systems (KOSs) and their implementation in a way that enables easy cross-linking. For example, the platform GeoNames (http://www.geonames.org) publishes about eleven billion place names in machine readable format and has been used by many other services to relate a term like "Manaus" to a specific geographic reference, which in turn enables other services to link other names to this location, e.g., colloquial and historical alternative names such as "Barra do Rio Negro." Similarly, interdisciplinarity in science has also come with terminological problems that affects both knowledge organization systems and communication among scientific communities. The Knowledge Organization domain is beginning to grapple with these problems, as exemplified by the studies on the approaches to interdisciplinarity represented by the synthetic and faceted discipline-based Universal Decimal Classification (UDC) and that of the phenomenon-based Basic Concepts Classification (BCC) (Smiraglia and Szostak, 2018).

This paper reports on a three-year research project "Digging Into the Knowledge Graph," funded as part of the 2016 Round Four Digging Into Data Challenge (https://diggingintodata.org/awards/2016/project/digging-knowledge-graph). Our project involves comparing terminology employed within the LOD cloud with terminology employed within these two general but different KOSs - UDC and BCC. We are exploring whether these classifications can encourage greater consistency in LOD terminology and linking the largely distinct scholarly literatures that address LOD and KOSs. The outlines of our project have been described in Szostak et al. (2018). The primary task in the first year was to convert the BCC to LOD. The UDC summaries were converted to LOD in 2016; our group, working together with the UDC Consortium, will undertake conversion of an abridged version of the UDC as a next step.

Our first use case comes from musicology and focuses our efforts on problems in the digital humanities. Computerized Mensural Music Editing (http://cmme.org/), or CMME, emerged from the University of Utrecht under the leadership of Theodor Dumitrescu and Marnix van Berchum. Guided by an international team of musicologists, CMME is a web-based repository for a corpus of high-quality edited scores from the era of Renaissance polyphony. The repository has a simple yet elegant structure incorporating rich metadata together with scores that can be viewed in modern or mensural notation. Content originates from edited published sources, and is indexed by composer, work, and manuscript source. Analysis of an SQL data dump demonstrated approximately 50% of the content is mensural music metadata, indicating the richness of the content. Contents include 3671 works by 221 composers; 14086 terms include 4586 unique musical terms occurring in 12961 multi-word phrases.

CMME is a good use-case for DiKG because it covers many aspects of humanistic scholarship, it has stable open access content and has been deposited at DANS. It is amenable to enrichment as LOD by linking composers, works and manuscript sources to the Virtual Internet Authority File (http://viaf.org/viaf/data/), by linking metadata for musical forms, media, notation, texts and liturgical functions to available LOD controlled vocabularies. However, work to date has shown that less than 75% of the composer-work authorized access points from CMME occur in VIAF; thus, our project will have to create authority data for those and find a way to enter them into VIAF. Similarly, LOD form and genre tools for musicology are evolving and few of the appropriate terms from CMME metadata can be linked at this time. Finally, the conceptual content also presents challenges for representation in any general classification but particularly in UDC and BCC, because most classifications of musical works are not concept-based but rather are medium and document-based.

2 The semantic problems of big data

In the "Web of Big Data," massive amounts of diverse kinds of information are constantly collected, processed, mixed, and statistically analyzed in hopes of finding correlations that reveal meaning. In this context, the production of knowledge is also a constant process that more often than not goes beyond the control of its producers and raises privacy concerns (Mai 2016). The model of big data has been compared to crowdsourcing, as it requires human labor to improve the predictive power of its algorithms (Ibekwe-SanJuan and Bowker, 2017). However, in this bottom-up approach to the production and search for meaning, the participation of people varies in quality and often lacks documentation on provenance (who did what, why, when and where), and thus there is not always agency in the definition of semantics. It is an art in itself to make implicit knowledge explicit. For the providing expert to disseminate knowledge, she must be able to also address the contextual assumptions, customs and limitations. Only then an 'outsider' has a good chance to understand what is made explicit in the same way the 'insider' intended. Contrary to the ideal of the semantic web in which "information is given well-defined meaning, better enabling computers and people to work in cooperation" (Berners-Lee et al., 2001), in the Web of Big Data, information does not always have well-defined meaning and people are not always able to consciously cooperate. In other words, Big data may be shared in a digital way but without having a shared understanding for those parties consuming it.

Perhaps the Knowledge Organization (KO) community has not been involved enough in the organization of big

data. Perhaps some of the characteristics of big data requires us to reflect on the scope of KO if we want it to be relevant in the era of big data. As KO, as a field, "deals with the classification of existing knowledge accumulated over thousands of years of scientific inquiry" (Ibekwe-SanJuan and Bowker, 2017, p. 194) and "it is important for KO to be concerned with scientific theories" (Ibekwe-SanJuan and Bowker, 2017, p. 190; Hjørland, 2015), it is important to reflect on the scientific validity and nature of those data and knowledge to be organized (in relation to the role that theories and hypotheses play in knowledge organization systems and in a supposedly pure data-driven analysis, (Mayer-Schönberger and Cukier, 2013, p. 50-72; Mazzocchi, 2015; Martínez-Ávila, 2018). Knowledge Organization is said to be about "describing, representing, filing and organizing documents and document representations as well as subjects and concepts both by humans and by computer programs" (Hjørland, 2016, p. 475). However, since the methods of collection of Big data make data difficult to reflect theories and knowledge, as they are often taken out of context, in the web of (big) data, it seems that the emphasis does not have to be put as much on documents and document representations as on the concepts and meanings. This would be also aligned with the activity and interest of Linked Data: "Simply put, Linked Data takes the World Wide Web's ideas of global identifiers and links and applies them to (raw) data, not just documents" (Hitzler and Janowicz, 2013, p. 233). In this sense, Jens-Erik Mai (2016, p. 193) also points out that the datafication of contents of the big data universe, for instance in projects such as Google Books, dismantles the traditional data–information–knowledge–wisdom (DIKW) pyramid ("When something is datafied there is no distinction between what data are and what information is: These are all elements that can be analyzed for patterns and correlations"). (Lack of) wisdom has also been the object of study in the context of big data (Pauleen et al., 2016). In this context, can we say that KO is still relevant for the organization of big data? In spite of the dismissal of the document as the main carrier of knowledge, it is also true that these data still have to be organized to infer knowledge. In this sense, we can say, as Mai (2016, p. 192) puts it, that "in the age of big data we need be concerned not only about the collection of data but equally about the way we process data in order to generate new information and knowledge."

All The relevance of KO and knowledge organization systems (KOSs) for big data has been discussed by Fidelia Ibekwe-SanJuan and Geoffrey C. Bowker in relation to the scalability of the KOSs: "big data algorithms raise the question of the relevance of humanly constructed KOSs and their capacity to keep up with the ever-increasing size of available data on specific topics and domains" (2016, p. 188). Drawing on the work of Hjørland, they also listed some of the problems that affect the KO field in the digital age and in relation to big data: "1) the possible obsolescence of universal bibliographic classification schemes; 2) the neglect of subject knowledge by library classifiers; and, 3) the reluctance of the KO community to leverage data analysis techniques as an alternative to manually constructed KOSs." In relation to 2, while Ibekwe-SanJuan and Bowker recognize that most of Hjørland's criticism is focused on the Universal Decimal Classification (UDC) and the Dewey Decimal Classification (DDC), they also claim that other kinds of KOSs such as thesauri, ontologies, and specialized classification schemes "are all domain-dependent knowledge artefacts that make no claim to universalism and should therefore be amenable to more frequent updates" (p.189). However, the UDC (as well as other universal classifications such as the Basic Concepts Classification, BCC), as they deal with interdisciplinarity, might also present some advantages for the organization of vocabularies. For instance, we believe these systems can be helpful in aspects related to authority control, establishment of relationships, and mappings of Linked Open Data (LOD), as well as being used as reference systems to develop generic principles of indexing (see Szostak et al., 2018).

Several studies have highlighted the relationship between Big data and Linked Data (e.g., Hitzler and Janowicz 2013), and also in relation to KO (Shiri, 2014). It seems to be well-accepted that Linked Data is part of the Big data realm. Shiri (p.18) also states that Linked Data can be considered a major type of data in the universe of Big data ("research data, open data, linked data and semantic web data can be construed as part of big data"). In this sense, Hitzler and Janowicz (2013, p. 234) claim that "Linked Data is an ideal testbed for researching some key Big Data challenges and to experience the 4th paradigm in action." The fourth paradigm (Hey et al.; 2009) refers to exploration and data intensive computing and is understood by Hitzler and Janowicz as "the scientific view on how Big Data changes the very fabric of science" (p. 233). Two points are important here: first, again, the problem of the interpretability of raw data and the role of theories and hypotheses in making science (something that also affects the development of KOSs); and second, the importance of semantics for processing big data. Hitzler and Janowicz (2013, p. 234) also say: "Indeed, Linked Data reduces Big Data variability by some of the scientifically less interesting dimensions. […] In this sense, Linked Data is a bit like Big Data in a laboratory setting, where certain variables are under control and thus can be ignored in the development of solutions or at least a deeper understanding of the issues. And once we have learned how to deal with the remaining variety dimensions in Linked Data, we are in a much better position to take further steps towards tackling Big Data at large." Indeed, Bizer et al (2011), while reporting on the 2011 STI Semantic Summit in

Riga, Latvia, a meeting to discuss the opportunities and challenges posed by big data for the Semantic Web, semantic technologies, and database communities, concluded that "the greatest shared challenge was not only engineering Big Data, but also doing so meaningfully" (p. 56). In this sense, Christian Bizer highlighted some unique characteristics of the Web of Data (i.e., the Web of structured data according to the Linked Data principles) that are relevant for the research on data integration and big data processing (i.e., organization). These characteristics are the use of widely-used vocabularies vs. proprietary vocabularies, correspondences between data using identity links and vocabulary links (e.g., owl:sameAs and SKOS:exactMatch), and the varying degree of data quality. Here is where we believe that KO can contribute to the LOD discourse in the definition of semantics, a view that seems to be shared by Shiri (2014, p. 18): "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."

On the other hand, we believe that KOSs can also benefit from the LOD universe. More specifically, LOD clusters can provide literary warrant for extending enumeration and clarifying the KOS. This would partially address Ibekwe-SanJuan and Bowker's question of literary warrant in the age of big data ("How will literary warrant be construed given that the available size of data from which such warrants can be drawn has grown exponentially and will continue to do so, and also that the said data is constantly changing?" p. 194) and also their third criticism in relation to the capacity of KOSs to keep up with the current state of knowledge and scientific findings. Ibekwe-SanJuan and Bowker related the possible solutions to the challenges of KO in the age of Big data to the possibilities of crowdsourcing. We agree that the maximization of the participation and the number of contributions of crowdsourcing (as it is also one of the basic characteristics of the Semantic Web and of the World Wide Web too) can be an answer to the problems of semantics and interoperability, but only if the contributions are done in a meaningful and structured way. Bizer et al. (2011, p. 60) summarized this problem and the hints for solutions as follows: "when we move to a more diverse field like data integration, which indubitably is the core question of big data, we need more stakeholder involvement." In this sense, we need producers and users to be able to provide and use richer semantics and structures in the vocabularies that are used to describe and publish data. In our view, this situation could also benefit from a greater application and involvement of KO principles.

3 Conclusions

Big data has been characterized by several V's that, according to Hitzler and Janowicz (2013), correspond with different scientific disciplines/discourse communities: volume, whose problem has been addressed by supercomputing; velocity, addressed by researchers working on sensor webs and the internet of things; veracity and value, both of the interest of the social sciences and humanities; and variety, studied by the Semantic Web (as well as the fields of databases, artificial intelligence, and cognitive science as a generalization of the problem of semantic heterogeneity). Where does the KO community fall in this division? The scope of KO, as an interdisciplinary field, seems to be in-between the social sciences and humanities, interested in aspects related to veracity and value, and the Semantic Web Community. Several studies within the KO community (e.g., Smiraglia, 2012; Martínez-Ávila et al., 2014; Martínez-Ávila, 2015) have identified LOD as an emerging and key trend for the future of KO. Shiri (2014, p. 19) also claimed that "Big data organization, representation and visualization will be among the emerging areas or research that information organization research will have to address." However, in spite of these recognitions, there are still many lessons and insights from the Linked (Big) Data which are applicable to the process of developing and maintaining (universal) classification schemas (e.g., Ibekwe-SanJuan and Bowker, 2017, also echoing Soergel 2015). In our current project, we try to connect these two related but rarely collaborating research communities: The Linked Open Data community, which has tended to be centered in computer science, and the KO community, with members from linguistics, metaphysics, library and information science. We focus on the shared challenges related to Big Data between both communities.

References

Berners-Lee, Tim; Hendler, James; Lassila, Ora (2001). The Semantic Web. // Scientific American. (May 2001) 1-4.

Borgman, Christine L. (2016). Big data, little data, no data: scholarship in the networked world. Cambridge, MA: The MIT Press, 2016.

Bizer, Christian; Boncz, Peter; Brodie, Michael L.; Erling, Orri. (2011). The Meaningful Use of Big Data: Four Perspectives - Four challenges. // SIGMOD. 40:4 (2011). 56-60.

Gregory, Kathleen; Cousijn, Helena; Groth, Paul; Scharnhorst, Andrea; Wyatt, Sally. (2018). Understanding Data Retrieval Practices: A Social Informatics Perspective. Preprint, Retrieved from http://arxiv.org/abs/1801.04971

Hey, Tony, Stewart Tansley and Kristin Tolle. (2009). The Fourth Paradigm: Data-Intensive Scientific Discovery. Redmond: Microsoft research, 2009.

Hitzler, Pascal and Krzysztof Janowicz. (2013). Linked Data, Big Data, and the 4th Paradigm. // Semantic Web. 4:3 (2013) 233-235.

Hjørland, Birger. (2015). Theories are Knowledge Organizing Systems (KOS). // Knowledge Organization. 42:2 (2015) 113-128.

Hjørland, Birger. (2016). Knowledge Organization (KO). // Knowledge Organization. 43:6 (2016) 475-484.

Ibekwe-SanJuan, Fidelia and Bowker, Geoffrey C. (2017). Implications of Big Data for Knowledge Organization. // Knowledge Organization. 44:3 (2017) 187-198.

Mai, Jens-Erik. (2016). Big data privacy: The datafication of personal information. // The Information Society. 32:3 (2016) 192-199.

Martínez-Ávila, Daniel. (2015). Knowledge Organization in the Intersection with Information Technologies. // Knowledge Organization 42:7 (2015) 486-498.

Martínez-Ávila, Daniel. (2018). Hacía una base teórica social de la Ciencia de la Información. // Anuario ThinkEPI. 12 (2018) 83-89.

Martínez-Ávila, Daniel; San Segundo, Rosa; Zurian, Francisco A. (2014). Retos y oportunidades en organización del conocimiento en la intersección con las tecnologías de la información. // Revista Española de Documentación Científica. 37:3 e053 (2014). DOI: http://dx.doi.org/10.3989/redc.2014.3.1112.

Mayer-Schönberger, Viktor; Cukier, Kenneth. (2013). Big data: A revolution that will transform how we live, work, and think. New York: Houghton Mifflin Harcourt, 2013.

Mazzocchi, Fulvio. (2015). Could Big Data be the end of theory in science? A few remarks on the epistemology of data-driven science. // EMBO reports. 16:10 (2015). 1250-1255.

Pauleen, David J.; Rooney, David; Intezari, Ali. (2016). Big data, little wisdom: trouble brewing? Ethical implications for the information systems discipline. // Social Epistemology, DOI: 10.1080/02691728.2016.1249436

Shiri, Ali. (2014). Linked Data Meets Big Data: A Knowledge Organization Systems Perspective. // Advances in Classification Research Online 24: 16-20. DOI:10.7152/acro.v24i1.14672

Smiraglia, Richard P. (2012). Knowledge Organization: Some Trends in an Emergent Domain. // El Profesional de la Información. 21:3 (2012). 225-227.

Smiraglia, Richard P.; Szostak, Rick. (2018). Converting UDC to BCC: Comparative Approaches to Interdisciplinarity. // Challenges and Opportunities for Knowledge Organization in the Digital Age: Proceedings of the Fifteenth International ISKO Conference, 9-11 July 2018, Porto, Portugal, ed. Fernanda Ribeiro, Maria Elisa Cerveira. Advances in Knowledge Organization 16. Würzburg: Ergon Verlag, 530-38.

Soergel, Dagobert. (2015). Unleashing the Power of Data through Organization: Structure and Connections for Meaning, Learning and Discovery. // Knowledge Organization 42:6 (2015). 401-427.

Szostak, Rick; Scharnhorst, Andrea; Beek, Wouter; Richard P. Smiraglia, Richard P. (2018). Connecting KOSs and the LOD Cloud. // Challenges and Opportunities for Knowledge Organization in the Digital Age: Proceedings of the Fifteenth International ISKO Conference, 9-11 July 2018, Porto, Portugal, ed. Fernanda Ribeiro, Maria Elisa Cerveira. Advances in Knowledge Organization 16. Würzburg: Ergon Verlag, 521-29.

Received:2018-10-18 Accepted: 2018-12-06