Guest editorial

New directions, opportunities, and challenges in organizing information and knowledge in the big data environment

The explosion of information that has become big data offers both opportunities and challenges for providing efficient access. For example, knowledge graphs – used increasingly in education (Wang *et al.*, 2018), law (Herring *et al.*, 2018) and medicine (Deng *et al.*, 2019) as well as other domains – is one example of a tool resulting from organizing big data. As quality knowledge graphs help produce valuable discoveries, examination of knowledge graph quality is gaining importance (Paulheim, 2017). With the growing availability of digital data on scholarly activities, there is an increased need for semantic analysis to help ensure that machines can understand the concepts behind academic language and mine valuable knowledge. Semantic role labeling (SRL) – a natural language processing task that locates arguments for a given predicate in a sentence and labels them with semantic roles – helps to meet this need and supports information retrieval, information extraction, question answering and machine translation. The need for SRL to focus on scholarly big data in a variety of domains has been identified (Dahlmeier and Ng, 2010).

The continuous development of information technologies greatly affects the development of digital libraries. The big data environment presents challenges to organizing digital and non-digital information for access; for example, in the digital humanities field (Tomasi, 2018). Digital libraries realize the need to develop a more solid understanding of research data and to overcome the "cyberinfrastructural challenge" to be able to adequately support curation, sharing and reuse of data generated by data-intensive research (Salo, 2010; Xie and Fox, 2017). The role of libraries in big data has been assessed in several recent studies (Zhan and Widén, 2018).

The synergy between library and information science (LIS) and data science emerges to address these big data challenges and opportunities. Some of the most fruitful areas of such collaboration have been identified within the theory and practice of information organization and knowledge organization (as defined in the LIS fields). For example, "big metadata, smart metadata", and leveraging the "metadata capital" have been named as important areas of research (Greenberg, 2017). Major conferences, such as the Annual Meeting of the Association for Information Science and Technology (2018) and the ACM/IEEE Joint Conference on Digital Libraries (2019, 2020) have held workshops focusing on knowledge organization and information organization (with metadata in particular) in the big data and the semantic web environments.

This special issue sought to explore solutions to facilitating access to data, information and knowledge on a large scale, through organizing information and knowledge. Articles published in this issue are based on the research and implementation projects whose selected or preliminary results were presented by the participants of the ACM/IEEE international workshop "Organizing Big Data, Information and Knowledge" held in August of 2020. The issue covers a variety of areas of interest summarized above: knowledge graphs, SRL for scholarly data, metadata and more, from both data science and information science perspectives. Two papers focus on information and knowledge organization as widely practiced in libraries, archives and museums and defined by the International Society for Knowledge Organization (www.isko.org/cyclo/origins): in particular, discovery metadata. Phillips and Tarver examine the novel computational approach of a network-theory-based metadata record graph to establish, visualize, and analyze connections between information objects held by cultural heritage institutions. The authors test the metadata graph development

The Electronic Library
Vol. 39 No. 3, 2021
pp. 389-391

0264-0473
DOI 10.1108/EL-06-2021-333

in the digital library environment, evaluate resulting graphs and discuss implications for metadata quality assurance and knowledge discovery. Zavalin and Miksa address the metadata theme from a somewhat different angle. They discuss the technological challenges that arise (and propose methodological solutions) for researchers and practitioners aiming to examine overall quality in large data sets of library metadata in MARC 21 metadata scheme from the WorldCat global aggregation. The authors developed and tested these approaches in their recent study of WorldCat bibliographic metadata records quality and, more specifically, linked data application and quality of subject representation with the knowledge organization systems: classification schemes and controlled vocabularies. While these two papers focus on highly structured big data in the form of metadata, other articles in this issue explore various dimensions of less structured or unstructured big data. For example, an interdisciplinary team of data scientists and physics researchers, Zhao *et al.* comparatively analyse the performance of two alternative ways of extracting knowledge from unstructured scholarly data in the material science domain by computational semantic labeling. One of the two methods relies on neural networks, whereas another one relies on ontologies-based automatic indexing. Pengcheng Li and colleagues focus on machine learning applications for named entity recognition to support big data knowledge organization in somewhat structured data: a corpus of full-text research papers originally encoded in the extensible markup language format. While their study also focuses on entity recognition, Yu *et al.* report results of testing – in the cross-lingual corpus of English and Chinese text – the proposed framework for extracting entity relations. Their project contributes to big data knowledge organization solutions by extending the knowledge acquisition task in machine learning. Li and his collaborative team of information science and computer science researchers use graph-based MGraph semantic data modeling with the focus on "evolutionary knowledge" defined by the authors as knowledge schema and instances that are gradually derived from raw data, illustrated by knowledge extraction from video frames of a basketball match.

The novel and innovative approaches to research and practice of information and knowledge organization investigated and discussed in this issue have a potential of helping cultural heritage institutions in meeting the challenges and embracing the opportunities of the big data environment. It is not only the research and practical work of libraries, archives and museums in information organization and knowledge organization that is affected by the challenges and opportunities introduced by the big data environment but also, importantly, education for information professionals. While this issue does not include articles focusing specifically on education, we believe research papers published in it will also be useful for informing training and curriculum development.

Oksana Zavalina
Information Science, UNT, Denton, Texas, USA
Xiaoguang Wang
Wuhan University, Wuhan, China, and
Qikai Cheng
School of Information Management, Wuhan University, Wuhan, China

References

Dahlmeier, D. and Ng, H.T. (2010), "Domain adaptation for semantic role labeling in the biomedical domain", *Bioinformatics*, Vol. 26 No. 8, pp. 1098-1104.

Deng, Y., Li, Y., Du, N., Fan, W., Shen, Y. and Lei, K. (2019), "When truth discovery meets medical knowledge graph: Estimating trustworthiness degree for medical knowledge condition", paper

presented at the 28th Conference on Information and Knowledge Management (CIKM '19), November, Beijing, China, available at: http://arxiv.org/abs/1809.10404 (accessed 12 August 2020).

Greenberg, J. (2017), "Big metadata, smart metadata, and metadata capital: toward greater synergy between data science and metadata", *Journal of Data and Information Science*, Vol. 2 No. 3, doi: 10.1515/jdis-2017-0012 (accessed 12 August 2020).

Herring, J., Cavar, D. and Meyer, A. (2018), "Case law analysis using deep NLP and knowledge graphs", in Rehm, G., Rodríguez-Doncel, V. and Moreno-Schneider, J. (Eds), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC '18), 12 May, Miyazaki, Japan, European Language Resources Association (ELRA), Paris*, pp. 1-6, available at: http://lrec-conf.org/workshops/lrec2018/W22/pdf/7_W22.pdf (accessed 12 August 2020).

Paulheim, H. (2017), "Knowledge graph refinement: a survey of approaches and evaluation methods", *Semantic Web*, Vol. 8 No. 3, pp. 489-508.

Salo, D. (2010), "Retooling libraries for the data challenge", *Ariadne*, Vol. 64, available at: www.ariadne.ac.uk/issue/64/salo/ (accessed 12 August 2020).

Tomasi, F. (2018), "Modelling in the digital humanities: conceptual data models and knowledge organization in the cultural heritage domain", Historical Social Research/Historische Sozialforschung, Supplement 31, pp. 170-179, available at: www.jstor.org/stable/26533637 (accessed 12 August 2020).

Wang, R., Yan, Y., Wang, J., Jia, Y., Zhang, Y., Zhang, W. and Wang, X. (2018), "AceKG: a large-scale knowledge graph for academic data mining", paper presented at the 27th ACM International Conference on Information and Knowledge Management (CIKM '18), 22-26 October, Torino, Italy, pp. 1487-1490, 10.1145/3269206.3269252 (accessed 12 August 2020).

Xie, Z. and Fox, E. (2017), "Advancing library cyberinfrastructure for big data sharing and reuse", *Information Services and Use*, Vol. 37 No. 3, pp. 319-323, doi: 10.3233/ISU-170853 (accessed 12 August 2020).

Zhan, M. and Widén, G. (2018), "Public libraries: roles in big data", *The Electronic Library*, Vol. 36 No. 1, pp. 133-145, doi: 10.1108/EL-06-2016-0134 (accessed 12 August 2020).