Editorial

Enayat Rajabi*

Shannon School of Business,
Cape Breton University,
Sydney, NS, Canada
Email: enayat_rajabi@cbu.ca
*Corresponding author

Amin Beheshti

Department of Computing,
Macquarie University,
Sydney, NSW, Australia
Email: amin.beheshti@mq.edu.au

Biographical notes: Enayat Rajabi is an Assistant Professor of Business Analytics with the Shannon School of Business at Cape Breton University in Canada, as well as an Adjunct Professor of Information Management with Dalhousie University's School of Information Management. He received his PhD in Information and Knowledge Engineering from the University of Alcala, Spain. His work has focused on semantic web and linked data since 2010, and he has published his related research in several JCR journals.

Amin Beheshti is the Director of AI-enabled Processes (AIP) Research Centre and the head of the Data Analytics Research Lab, Department of Computing, Macquarie University. He is also the Senior Lecturer in Data Science (Macquarie University) and Adjunct Academic in Computer Science (UNSW Sydney). He has been recognised as a high-quality researcher in Big-Data/Data/ Process Analytics and has been invited to serve and served as Keynote Speaker, General-Chair, PC-Chair, Organisation-Chair and program committee member of top international conferences. He is the leading author of the book entitled *Process Analytics*, published by Springer.

1 Introduction

The continuous improvement in connectivity, storage, and data processing capabilities allow access to a data deluge from the big data generated on various islands of data, from social (e.g., on Twitter and Facebook) and Internet of Things (IoT, e.g., CCTVs and smart cars) to private (e.g., personal and business data) and open (e.g., news and websites) data islands. In computing, data is information that has been translated into a form that is efficient for storage and processing. Metadata is the data that provides information about other data (Beheshti, 2020). The goal of metadata is to help users properly understand the raw data. Metadata can be used to express the semantics of information and facilitates information seeking, retrieval, understanding, and use (Sicilia, 2006). Per W3C RDF document,[1] metadata can be used in many applications such as "cataloging for describing the content and content relationship" or "by intelligent software agents to facilitate knowledge haring and exchange". Metadata (e.g., descriptive, structural, and administrative metadata) can be generated automatically and can be used for information management and retrieval (Haynes, 2018). Tracing is a mechanism to generate metadata and will facilitate transforming the raw data into actionable insights. To achieve this goal, it will be vital to link and use the data and metadata in an intelligent way. In this editorial, we introduce several challenges that exist in different phases of metadata and ontology lifecycle management process.

2 Review of special issue papers

Leveraging metadata in different systems facilitates effective data management, discovery, and use. Metadata can be created manually to be more accurate, or it can be automatically extracted using extraction tools. One of the applications of metadata extraction is using it in identifying scanned texts. Accessing scanned texts in the web domain, for searching and extracting information about semantic topics and concepts, is a challenging task. The handling and processing of these documents require the adoption of alternative solutions, such as word spotting techniques. Word spotting methods aim to recognise complete words and can be applied mainly in areas in which OCR systems fail owing to unknown printed fonts or degraded printed texts. In the special issue, Varthis et al. provide a simple and effective pipeline word spotting framework for the automatic generation of metadata or search indexes for machine-printed texts such as Patrologia Graeca (PG). PG,

as an example, presents difficulties for OCR owing to its polytonic script (letters with diacritics) and the degraded and noisy scanned pages. The proposed framework in this study has direct applicability for digital libraries, such as automatically creating metadata and search indexes for the PG collection, by cropping regions of interest.

Metadata can be further structured and fostered to be used as a form of ontologies in different systems. Adding semantics relations brings meaningful information and new insights. Many studies in the domain of metadata and ontologies highlighted that combination of metadata, ontology and semantic-based approaches, thus significantly improving the quality of the results achieved in different applications. Process mining is one of the domains that can leverage the advantages of ontologies and semantics. Process mining aims at monitoring and improving real-world processes by extracting knowledge from event logs readily available in different information systems.

In this special issue, Ghalibafan et al. employ a semantic-based approach based on ontologies to address the main challenges arising in process mining. The authors propose a novel technique for extracting an ontology from an event log and enriching the ontology extracted from a process event log using databases. Using a bin log where all the operations, including insert, delete, and update in the databases are stored, they resolve the inadequacies of the event log including incorrect and missing data by extracting ontologies. The ontology is created based on the event log and the bin log in the analysis of a process and match the extracted ontologies for the process of data cleaning. The authors develop an algorithm for the purpose of enhancing the accuracy of ontology matching.

Ontologies can be also used to organise concepts, manage metadata, formulate some data searching strategies, and construct knowledge graphs. A knowledge graph is created based on a subset of a multilingual semantic network, WordNet, which is a large lexical database for semantic relations between words. The main purpose of building WordNet was to allow dictionaries to understand the conceptual relationships between words and phrases in the English language. WordNet combines nouns, verbs, adjectives, and adverbs in terms of cognitive synonyms (i.e., Synsets), each conveying a special concept. Synsets are interlinked by various semantic and lexical relationships. Albesher and Rabie performed a comprehensive review of Arabic WordNet (yielded 101 research papers) and explored how Arabic WordNet has been used so far. They discussed the challenges in regard to the Arabic language that makes applying this knowledge graph difficult. Several research gaps in the field of Arabic WordNet are discussed in this article.

The last phase of the metadata and ontology lifecycle is using them efficiently in data analytics platforms for knowledge discovery. Knowledge discovery refers to a process to extract useful knowledge from data. This process requires a lot of data that needs to be properly validated, cleaned, and integrated. Polychronou et al. propose leveraging a knowledge discovery system in a big data platform and performed experimentation over the currently provided datasets by focusing on the needs of the food safety sector. They describe a big data platform architecture that includes different components including data integration, data indexing, and a semantic API, and test the scalability and performance of the platform by different metrics.

References

Beheshti, A., Benatallah, B., Sheng, Q.Z. and Schiliro, F. (2020) 'Intelligent knowledge lakes: the age of artificial intelligence and big data', *International Conference on Web Information Systems Engineering*, Springer, Singapore, pp.24–34.

Haynes, D. (2018) *Metadata for Information Management and Retrieval: Understanding Metadata and Its Use*, Facet Publishing.

Sicilia, M.A. (2006) 'Metadata, semantics, and ontology: providing meaning to information resources', *International Journal of Metadata, Semantics and Ontologies*, Vol. 1, No. 1, pp.83–86.

Note

1 https://www.w3.org/RDF/FAQ