A critical introduction to metadata for e-science and e-research

Emmanouel Garoufallou*

Department of Library Science and Information Systems,
Alexander Technological Educational Institute of Thessaloniki,
PO Box 141, GR-574 00, Thessaloniki, Greece
Email: mgarou@libd.teithe.gr
*Corresponding author

Christos Papatheodorou

Database & Information Systems Group (DBIS),
Laboratory on Digital Libraries and Electronic Publishing,
Department of Archives, Library Science and Museology,
Ionian University, GR-49 100, Corfu, Greece
and
Digital Curation Unit,
Institute for the Management of Information Systems,
'Athena' Research Centre, Athens 15125, Greece
Email: papatheodor@ionio.gr

Abstract: Scientific research is moving towards multi-disciplinary, multi-institutional collaboration and therefore powerful tools and infrastructures based on interoperability principles are needed to support this trend. This paper introduces the special issue on the metadata for e-science and e-research of the *International Journal on Metadata, Semantics and Ontologies*. This special issue seeks to draw attention to the on-going challenges that scientists and systems developers face in the area of metadata and data management for e-science and e-research. In particular, the objectives of this special issue are (a) to present some of the latest research in this field, especially in relation to the use of metadata for addressing challenges associated with the management of scientific and research data across a broad range of applications; and (b) to highlight some of the challenges associated with the use of metadata, and encourage further research in this area. The special issue includes four papers reporting innovative approaches to key issues in the area of metadata for e-science and e-research, such as metadata modelling and standardisation, data quality and data re-use.

Keywords: e-science; e-research; metadata standards; metadata interoperability; data quality.

Reference to this paper should be made as follows: Garoufallou, E. and Papatheodorou, C. (2014) 'A critical introduction to metadata for e-science and e-research', *Int. J. Metadata, Semantics and Ontologies*, Vol. 9, No. 1, pp.1–4.

Biographical notes: Emmanouel Garoufallou is a Lecturer at the Department of Library Science and Information Systems and project manager of the Programme 'Open Source Digital Library Services of Alexander TEI of Thessaloniki' at the Alexander Technological Educational Institute of Thessaloniki (ATEITh), Greece. He is also coordinator of ATEITh libraries. He was the Chair of the 7th MTSR2013 Conference and is a member of the MTSR steering committee. He served as programme committee member of various international conferences, while he serves as an editorial board member of various international journals and as associate editor of the *Program: electronic library and information systems* journal.

Christos Papatheodorou holds a BSc and a PhD in Computer Science from the Department of Informatics, Athens University of Economics and Business, Greece. He is Associate Professor and the Chairman of the Department of Archives, Library Science and Museology, Ionian University, Greece. He is also a research fellow at the Digital Curation Unit, "Athena" Research Centre, Greece. He is the Chairman of the Steering Committee of the International Conference on the Theory and Practice of Digital Libraries (TPDL). His research interests include digital libraries evaluation, metadata interoperability, and digital curation/preservation.

1 Introduction

The parallel growth in scientific data (Big Data) and cloud computing has revolutionised the way scientific content is communicated to and used by researchers. E-science and e-research applications have extended the traditional forms of scholarly cyber-infrastructure, such as institutional repositories and digital libraries, to include new tools intended to satisfy new requirements in academic communication. Thus research groups focusing on a knowledge domain or interdisciplinary research communities need to collaborate and communicate both their workflows and the processes they followed to generate results and new knowledge. Indicative workflows include storing, manipulating, enriching and annotating, linking, disseminating and publishing their results (Jeffery, 2007), as well as the data generated during the various steps of the scientific inquiry, such as raw and processed datasets, data about methodologies, research instruments and models, information about individual researchers and research groups as well as data regarding funding bodies and research proposals (Castelli et al., 2013). These 'information objects' or 'research objects' (Bechhofer et al., 2010; Balatsoukas et al., 2012) might be not only textual but also multimedia, e.g. graphical representation of molecular structures, 3D building engineering structures, clinical guidelines, source codes, climate models, epidemiological data, economic models, social policy simulations, or human genome structures.

Metadata and ontologies are integral tools for the curation of e-science and e-research infrastructures ensuring open, comprehensive and persistent access to scientific material. Yet, challenges still exist regarding the role of metadata in the process of storing, preserving, managing, modelling, retrieving, representing and disseminating this type of information (Greenberg and Garoufallou, 2013). Challenges arise for several reasons, including: the heterogeneous, highly granular and unstructured nature of data (Balatsoukas et al., 2012); data provenance and the iterative scientific workflows involved in the process of data generation and dissemination with severe implications for data quality and re-use (Deelman et al., 2009; Simmhan et al., 2005); and, the complexity of the information governance issues surrounding the generation and dissemination of data (e.g. numerous collaborators with different access rights, security of sensitive and confidential data) (David and Spence, 2003).

This special issue seeks to draw attention to the on-going challenges scientists and systems developers face in the area of metadata and data management for e-science and e-research. In particular, the objectives of this special issue are:

- To present some of the latest research in this field, especially in relation to the use of metadata for addressing challenges associated with the management of scientific and research data across a broad range of applications.
- To highlight some of the challenges associated with the use of metadata, and encourage further research in this area.

Although special issues of several journals have focused on e-science and e-research – for example a special issue of the *Philosophical Transactions of the Royal Society* (Walker et al., 2011), the very recent issue of the *Future Generation Computer Systems* (Katz and Ambrason, 2013) and the proceedings of the annual *Metadata & Semantics Research Conference* (*MTSR*) – the uniqueness of the present special issue is the focus on metadata, and its role in the process of developing and implementing e-science and e-research infrastructures. Therefore, the selection of papers in this issue should benefit researchers and developers of e-science/ e-research infrastructures as well as metadata professionals and researchers in the area of metadata standardisation or the semantic web.

This paper is structured as follows. In the next section we present some definitions of the concept of e-science and e-research in the context of this special issue. The following section discusses the main challenges in the area of metadata for e-science and e-research. The next section introduces the papers featured in this special issue, while the editorial concludes by envisioning the future research trends in the domain.

2 Challenges and the future of metadata for e-science and e-research

The distinction between e-science and e-research is still unclear. Anecdotally, some researchers have attempted to draw a line between the two terms. For example, Beaulieu and Wouters (2009) approached e-research in a broader manner than e-science. According to their interpretation, e-science consists of three fundamental elements. These are: sharing of computational resources, access to big volumes of data and the use of platforms that promote collaboration and communication between stakeholders (Beaulieu and Wouters, 2009). In their definition, e-research was defined in a broader way to include also the presence of specific research methods and disciplinary research practices and workflows.

In the context of this editorial the terms e-science and e-research are used interchangeably to denote the types of application developed to support the harvesting, analysis, sharing and re-use of scientific and research data (or Big Data). In this manner, both terms are characterised by technologies that bring together three fundamental characteristics of e-science (as defined by Beaulieu and Wouters, 2009), but also the contextual research practices and workflows of e-research. Moreover, both terms are used in a broad manner to include solutions developed across all scientific fields, such as pure sciences, medical and life sciences, engineering, social sciences and humanities. Therefore, in the context of this editorial, terms such as e-social sciences, or e-humanities (digital humanities) also

form part of the concept of an e-science or e-research. This decision was made because the focus of this special issue was on applications and technologies developed across different academic fields. Also, this decision was motivated by the interdisciplinary nature of modern science, which, in many cases, makes it difficult to make distinctions between the different disciplines.

In particular, the special issue welcomed submissions focused on the application of metadata and related semantic technologies (such as ontologies and vocabularies) for solving problems related to a range of data management issues across different disciplines and research communities. Examples of this type of issue were: the modelling of scientific content; infrastructures, systems and services for knowledge organisation; ontology approaches, models, theories and languages; semantic representation of scientific content and remote collaboration; auto-generated vs. human generated e-science metadata; visualisation techniques for metadata, content, repositories; interoperability in e-research environments; workflow management models; open data and linked open data for e-science; cloud facilities and supercomputing for e-science; archiving and preservation metadata and conceptual models.

Given the complexity of e-science and e-research infrastructures, there are several challenges that need to be addressed in relation to the use of metadata. For the purpose of illustration, this section presents three broad challenges that researchers and systems developers should investigate. These are: metadata modelling and standardisation; data quality; and data re-use.

Metadata modelling and standards: To date, the development of traditional scholarly repositories and information retrieval systems has been followed by several attempts to define metadata standards for the description of information objects. Yet, the establishment of e-science infrastructures has not been supported by any organised attempts to promote metadata standardisation and modelling. Although the Common European Research Information Format (CERIF; Jeffery, 2007) data model provides an endeavour towards the use of a common format for information management across research projects and their publications, the heterogeneity of metadata schemas and application profiles, both across and within disciplines, is high with severe implications for data re-use.

Metadata for data quality: Data quality has been a problem monopolised by many data-intensive environments, such as business processes, stock markets or health and social policy making (Pipino et al., 2002). The advent of Big Data and modern cloud computing has revolutionised the way scientists interact with data. Data can be accessed simultaneously by many researchers from distributed laboratories and research groups around the globe, and new versions of this data can be generated collaboratively, stored, disambiguated and disseminated for further re-use (Simmhan et al., 2005). Therefore, the development of new data quality metrics becomes a challenge for a series of e-science and e-research workflow processes that can generate large volumes of data. This challenge is influenced by the nature of the data produced, which is characterised by malleability, volume-scalability, granularity and complex transformations. Although novel methods for addressing the problem of data quality in e-science have been proposed, such as provenance and curation metadata, the increasing volume and the nature of the data produced require further investigation.

Metadata and data re-use: Data re-use can be influenced by the level of data and metadata heterogeneity and semantic interoperability. Although several methods for improving interoperability exist, such as semantic mapping methods, metadata cross-walks and linked data, there is still variability in the effectiveness of these methods across the different scientific disciplines. Also, despite the popularity of linked open data for addressing issues of semantic interoperability between pieces of information and applications on the web, there are still problems associated with its implementation in e-science and e-research. In many cases, these problems can be associated with the process of creating links between multiple linked open datasets as well as the integration between linked open datasets, ontologies and RDFs (Jain et al., 2010; Bechhofer et al., 2013).

3 Overview of the papers

This special issue brings together some of the latest research in the field of metadata for e-science and e-research. It includes four papers, which propose novel approaches to solve problems associated with the challenges mentioned in the previous section (i.e. metadata modelling and standardisation, data quality and data re-use), and identify areas for further research.

In particular, the paper ‘Research information management: the CERIF approach’, by Keith Jeffery, Nikos Houssos, Brigitte Jörg and Anne Asserson, deals with the first category of challenges, metadata modelling and standards. The authors provide a detailed presentation of the CERIF data model and discuss some of the implications of this approach for data management, data re-use and metadata generation. CERIF is a three-layer data model, maintained by the euroCRIS community, which aims to describe uniformly heterogeneous research datasets, their creators, providers and administrators.

The paper ‘A provenance-based approach to evaluate data quality in e-science’, by Joana E. Gonzales Malaverri, André Santanchè and Claudia Bauzer Medeiros, deals with the second category of challenges, the data quality. The authors specify a framework for data quality measurement based on provenance metadata. Specifically, they provide a methodology to evaluate the quality of digital artefacts based on their provenance. The proposed methodology is validated experimentally by a prototype workflow system.

The paper ‘A linked open data approach for geolinguistics applications’, by Emanuele Di Buccio, Giorgio Maria Di Nunzio and Gianmaria Silvello, focuses on the third challenge, data re-use and interoperability. It

provides an innovative approach, based on Linked Open Data, to increase the level of data re-use in geolinguistic systems. Such systems explore the relationship between language and cultural adaptation and change and a functional requirement for them is reusability of linguistic tools and semantic integration of data collections. For this purpose, the authors define an ontology for geolinguistic resources, and provide a linked open dataset and an application to generate dynamically linguistic maps.

In the same category of challenges, the last paper of the special issue is entitled 'Metadata based management and sharing of distributed biomedical data', by Fusheng Wang, Cristobal Vergara-Niedermayr and Peiya Liu. The extremely rapid evolution of biomedical disciplines demands collaboration among researchers, and therefore flexible and powerful infrastructures are needed to facilitate them to re-use experiments and validate approaches. The authors present a novel metadata-based framework for managing and sharing distributed biomedical data. They present the conceptual and technical characteristics of this framework, and introduce SciPort, a web-based collaborative biomedical data management platform that makes use of metadata to facilitate sharing and re-use of distributed data.

4 Outlook

Scientific research is moving towards multi-disciplinary, multi-institutional collaboration and hence research data (raw data, secondary data and experimental workflows) should be inter-linked, discoverable and re-usable. Therefore there exist significant efforts for the creation of virtual research environments enabling communication, storage and preservation of scientific data.

Moreover, the scholarly communication paradigm is changing steadily and requires workflows that integrate data and publications. Data add value to a publication and facilitate its understanding. Therefore, several technologies have been proposed either to cite data or to incorporate data descriptions in publications. The most known are Linked Data as well as OAI-ORE (http://www.openarchives.org/ore/1.0/primer.html), which permit data providers and aggregators to publish source metadata and share authority files and vocabularies.

The current trend in scientific communication and collaboration is the development of research infrastructures that (i) ensure data exchange and interoperability between content resources, (ii) provide content storage and preservation, processing and provision functionalities, and (iii) provide services and workflows to the communities to exploit the content such as authorised access to resources, data curation workflows and communication services. Up to now, significant steps have been made to develop research infrastructures aiming to establish common data models, standards and best practices (e.g. DARIAH for humanities (http://www.dariah.eu/), and Research Data Alliance (https://rd-alliance.org/) with a more general vision for 'facilitating research data sharing and exchange'). In any case, metadata management and interoperability constitutes a vital parameter for the development of such infrastructures, and hence the future research on metadata for e-science and e-research should be directed to investigating the integration and seamless access to scientific data.

References

Balatsoukas, P., Williams, R., Carruthers, E., Ainsworth, J. and Buchan, I. (2012) 'The use of metadata objects in the analysis and representation of clinical knowledge', *Metadata & Semantics Research* (*MTSR 2012*), pp.107–112.

Beaulieu, A. and Wouters, P. (2009) 'e-Research as intervention', in Jankowski, N. (Ed.): *E-Research: Transformation in Scholarly Practice*, Routledge, pp.54–72.

Bechhofer, S. et al. (2013) 'Why linked data is not enough for scientists', *Future Generation Computer Systems*, Vol. 29, No. 2, pp.599–611.

Bechhofer, S., De Roure, D., Gamble, M., Goble, C. and Buchan, I. (2010) 'Research objects: towards exchange and reuse of digital knowledge', *The Future of the Web for Collaborative Science* (*FWCS 2010*).

Castelli, D., Manghi, P. and Thanos, C. (2013) 'A vision towards scientific communication infrastructures: on bridging the realms of research digital libraries and scientific data centre', *International Journal on Digital Libraries*, Vol. 13, Nos. 3/4, pp.155–169.

David, P. and Spence, M. (2003) *Towards institutional infrastructures for e-Science: the scope of the challenge*, Oxford Internet Institute, Research Report no.2.

Deelman, E., Gannon, D., Shields, M. and Taylor, I. (2009) 'Workflows and e-Science: an overview of workflow system features and capabilities', *Future Generation Computer Systems*, Vol. 25, No. 5, pp.528–540.

Greenberg, J. and Garoufallou, E. (2013) 'Change and a future for metadata', *Metadata & Semantics Research* (*MTSR 2013*), pp.1–5, DOI: 10.1007/978-3-319-03437-9_1.

Jain, P., Hitzler, P., Sheth, A., Verma, K. and Yeh, P. (2010) 'Ontology alignment for linked open data', *ISWC'10*, pp.402–417.

Jeffery, K.G. (2007) 'Technical infrastructure and policy framework for maximising the benefits from research output', *ELPUB 2007*, pp.1–12.

Katz, D. and Ambrason, A. (2013) 'Recent advances in e-Science', *Future Generation Computer Systems*, Vol. 29, No. 2.

Pipino, L., Lee, Y.W. and Wang, R.Y. (2002) 'Data quality assessment', *Communications of the ACM*, Vol. 45, No. 4, pp.211–218.

Simmhan, Y., Plale, B. and Gannon, D. (2005) 'A survey of data provenance', *ACM SIGMOD*, Vol. 34, No. 3, pp.31–36.

Walker, D., Atkinson, M., Brooke, J. and Watson, P. (2011) *Selected papers from the 2010 e-Science all hands meeting. Philosophical transactions of the Royal Society*, Vol. 369 (1949), pp.3251–3253.