
EDITORIAL CONTENT

Introduction to Data Models Special Collection

Nicholas M. Weber¹ and Karen M. Wickett²¹ University of Washington, US² University of Texas, USCorresponding author: Nicholas M. Weber (nmweber@uw.edu)

Keywords: Data Models; Knowledge Representation; Ontology development; Metadata; Data Curation

Innovations such as data harmonization (Niemi, Näppilä & Järvelin, 2009), interoperability frameworks (Hughes et al., 2016), and feature extraction tools (Bhattacharyya, Organisciak & Downie, 2015) are greatly improving the capabilities of research communities to access and manipulate data in computing systems. Underpinning these new systems-level features and functionalities are a number of robust conceptual, logical, and physical data models. These include data- and curation-oriented models such as the Open Provenance Model (Kwasnikowska, Moreau & Bussche, 2015) and the Research Object Model (Belhajjame et al., 2015), as well as ontologies such as the the Semantic Web for Earth and Environmental Terminology (SWEET)¹ and the Gene Ontology². However, the formal literature of data science often glosses over or excludes the design work that goes into developing and implementing data models (Simsion, 2007). As a result it is often unclear how or why design decisions were made, or what advances and new techniques have been developed for data modeling and knowledge representation as they are applied to research data.

The Data Science Journal will curate a new special collection of papers that are thematically aligned with these issues in order to fill a needed gap in the data modeling literature. Our intention is to publish research and practices papers that shed light on new methods of design and development, as well as reviews of recently released data models. In the following sections we give a brief overview of four initial contributions to this the data modeling special collection and highlight their unique contribution to the field. We then describe some potential topics of interest for future submissions to the special collection, as well as future directions in data modeling research more generally.

Special Collection Publications

The special collection on data models has published four initial papers that cover a range of important topics concerning the formal representation of research data. Mayernik et al. describe the development of EarthCollab, a harmonized ontology that uses the VIVO software suite for linking and making discoverable geodesy and polar science resources – including data, publications, and instrumentation (2016). Their contribution focuses specifically on the process of ontology selection and consolidation, with extended discussion of their method for reusing concepts from existing geoscience ontologies. The highlight of this work is a set of recommendations for good ontological modeling practices in the polar science community, which draws attention to problems in ontology versioning and maintenance, as well as lessons learned in consolidating classes of different ontologies with similar named concepts (e.g. dataset vs project).

Almas and Schroeder describe their work in applying an existing model of identification for scholarly texts in Classics (the CTS model) to a new corpus of digitized coptic material (2016). This contribution sheds considerable light on requirements engineering processes in digital humanities research – as they given an in-depth account of the domain-specific scholarly research and citation practices, the structure of the

¹ SWEET Ontologies: <http://sweet.jpl.nasa.gov/>² Gene Ontology Consortium: <http://geneontology.org/>

textual data, and the data management workflow that was necessary to incorporate into their modeling decisions. Similarly, Blackwell and Smith offer a unique perspective on developing abstract, conceptual models for identifying, citing, and reliably retrieving textual fragments in their work in the Homer Multitext project (2016). In doing so, they describe a model of representation, Ordered Hierarchy of Citation Objects (ohco2), which assigns sub-properties to data-objects (text fragments) for identification and reference. This work is based on a series of use-cases developed from scholars working on the Homer Multitext project – providing a helpful empirical application of their model.

Finally, Hester provides background the design of a programming interface to evaluate data transfer standards (2016). This contribution gives a thorough overview of background research that has been conducted in software development aimed at assisting scientists writing input or output modules for a data standard. Hester then builds upon an existing framework, the Olog approach, and provides a set of use cases that demonstrate the utility of these concepts he refines for achieving semiautomatic data file translation and transfer.

The contributions in this special collection draw from data modeling expertise in a variety of domains, and highlight data modeling and representation challenges from within those domains. Throughout these papers, there is attention to balancing overall data modeling strategies against the perspectives driving research within the originating areas. This sensitivity to domain knowledge and to the unique aspects of data arising from the domain is essential for effective data modeling. By bringing these perspectives together in the special collection, we can surface some of the fundamental aspects of data modeling and representation for current data management and sharing. Across the four initial contributions, we see emphasis on enabling collaboration through data handling and management techniques from a number of disciplinary applications, including those making up the digital humanities, earth sciences and physical sciences. These contributions also demonstrate the connections between the structures that are applied to data in collection, use, and sharing – and how those structures shape the potential reuse of those data.

Future topics for the special collection

The special collection will continue to grow with submissions made to the Data Science Journal. The goal is to build a body of discourse around data modeling and representation to support data management and sharing. As new techniques are developed and new models are proposed, we hope to bring together contributions in order to foster communication to aid practice and to illuminate the underlying goals and techniques that guide data modeling.

Future topics might include any or all of the following:

Design choices: A designer of a data model often faces choices between expressiveness, ease of use, and computational complexity – How are these tradeoffs accounted for in doing requirements engineering at the beginning stages of developing a curation system?

Harmonization: What are complications in, or best practices for harmonizing conceptual models? (e.g. FRBR + CIDOC CRM = FRBRoo)

Interoperability: How have data models been developed to facilitate cross or interdisciplinary data interoperability?

Requirements Engineering: Research data systems are often developed by working closely with data producers and potential systems users. How are requirements for a data model generated from these types of interactions?

Ontology Development: Ontologies capture a conceptualization of a domain. How are the essential aspects of research domain or a research data system to be analyzed for representation? How can an existing ontologies be evaluated for potential implementation or refinement?

Sustainability: Knowledge organization and representation activities contribute greatly to the sustainability and long-term success of a research data curation systems – How do these activities co-evolve with the discipline or domain that they serve? How have data models and metadata schemas been edited and revised to accommodate changes in scale, complexity, or heterogeneity of research data?

Education: What are the competencies necessary for doing knowledge representation work and research data systems development? How are these skills taught in classrooms, workshops, and continuing education programs?

Competing Interests

The authors have no competing interests to declare.


References

- Almas, B** and **Schroeder, C T** 2016 Applying the Canonical Text Services Model to the Coptic SCRIPTORIUM. *Data Science Journal*, 15: 13. DOI: <http://doi.org/10.5334/dsj-2016-013>
- Belhajjame, K, Zhao, J, Garijo, D, Gamble, M, Hettne, K, Palma, R, . . ., Klyne, G** 2015 Using a suite of ontologies for preserving workflow-centric research objects. *Web Semantics: Science, Services and Agents on the World Wide Web*, 32: 16–42. DOI: <https://doi.org/10.1179/0308018814Z.000000000105>
- Bhattacharyya, S, Organisciak, P** and **Downie, J S** 2015 A Fragmentizing Interface to a Large Corpus of Digitized Text:(Post) humanism and Non-consumptive Reading via Features. *Interdisciplinary Science Reviews*, 40(1): 61–77.
- Blackwell, C** and **Smith, N** 2016 Modeling Citable Textual Analyses for the Homer Multitext. *Data Science Journal*, 15: 17.
- Hester, J** 2016 A Robust, Format-Agnostic Scientific Data Transfer Framework. *Data Science Journal*, 15: 12. DOI: <https://doi.org/10.5334/dsj-2016-012>
- Hughes, J S, Hardman, S, Crichton, D J, Martinez, S, Law, E** and **Gordon, M K** 2016 A Working Framework for Enabling International Science Data System Interoperability. *41st COSPAR Scientific Assembly, Abstract S. 2-5-16*. (Vol. 41).
- Kwasnikowska, N, Moreau, L** and **Bussche, J V D** 2015 A formal account of the open provenance model. *ACM Transactions on the Web (TWEB)*, 9(2): 10. DOI: <https://doi.org/10.1145/2734116>
- Mayernik, M S**, et al 2016 Building Geoscience Semantic Web Applications Using Established Ontologies. *Data Science Journal*, 15: 11. DOI: <https://doi.org/10.5334/dsj-2016-011>
- Niemi, T, Näppilä, T** and **Järvelin, K** 2009 A relational data harmonization approach to XML. *Journal of Information Science*. DOI: <https://doi.org/10.1177/0165551509104231>
- Simsion, G** 2007 *Data modeling: theory and practice*. Technics Publications.

How to cite this article: Weber, N M and Wickett, K M 2016 Introduction to Data Models Special Collection. *Data Science Journal*, 15: 14, pp.1–3, DOI: <http://dx.doi.org/10.5334/dsj-2016-014>

Published: 23 November 2016 **Published:** 23 November 2016 **Published:** 06 December 2016

Copyright: © 2016 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

 *Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS 