

NISO Thought Leader Meeting on Research Data

Report prepared by Maureen C. Kelly, consultant and facilitator of the October 1, 2008 meeting.

Introduction

On October 1, 2008 NISO held a Thought Leader meeting on the topic of research data. This was the final meeting in a series of NISO Thought Leader meetings sponsored by a Mellon Foundation grant. The goal of this meeting was to incubate new standards initiatives by discussing issues and areas where standards can help address pain points, push forward reuse of data, or drive application of systems in research and information exchange.

Meeting Preparation and Process

In preparation for the meeting, approximately 20 candidate participants were identified. This list was reviewed by the NISO Business Information Topic Committee and additional names were suggested. All names were prioritized and invitations were sent. The final list of attendees consisted of experts representing varying perspectives on the topic of research data, including working scientists, publishers, and others involved in information distribution and retrieval. There were 10 on-site participants and one remote participant. The list of participants appears in Appendix A.

In advance of the meeting, all invited experts were encouraged to identify key issues for discussion on October 1st. Contributions were submitted to a private blog set up for this purpose (<http://niso-researchdata.blogspot.com/>). Topics identified for consideration included: issues around provenance, metadata, citation and reference, version control and tracking, preservation, privacy, intellectual property, and other technical and cultural/social issues that impact data reuse.

The meeting began with a brainstorming discussion about barriers that exist to wider sharing of research data. Following the brainstorming discussion, the participants focused on more practical issues and considered how NISO might proceed with developing a sample standard (i.e., a standard for citing research data). This served to identify potential barriers to standards development in this area and allowed concerns to be shared and discussed. During final segment of the meeting, participants identified specific areas where standards would be useful; they then considered realistic ways in which NISO might proceed.

Issues & Challenges Affecting the Development of Standards for Sharing Research Data

Research data presents certain unique challenges in contrast to traditional scholarly literature. Citation practices for journal articles are well established. While differences exist across fields of scholarship with regard to citation formats, the metadata requirements are basically the same. They serve to enable the reader to direct the reader to the cited resource. Standards developed for the print environment have been adapted to serve this purpose in the electronic environment, and supporting systems, technologies and practices continue to be developed (e.g., DOI, link resolvers, most-appropriate-copy resolution, etc.).

Research data have only recently become a topic of interest for sharing and citation. As such, there are few widely used best practices or standards to support retrieval and reuse of research data. Whereas data may be represented in a graph or table in a journal article, the full data set typically does not accompany the article, and there are a number of factors that make this difficult to accomplish on a large scale. Scholarly publishers and societies have addressed these issues, but significant challenges remain. Good standards would make it easier for researchers to examine and reference data collections that are not collocated with the journal article.

The Thought Leader discussions explored both the challenges and opportunities for incorporating standards into data sharing practices. The group discussed several important issues that will need to be considered in conjunction with standards development:

1. Differing types of data collections: There are varying types of research data, each presenting different challenges for standards development. In particular, standardization efforts must recognize the differences for observational data, experimental data, and data from models and simulations. Observational data collections were considered most suited to standards development at this time.

2. Differences across fields of research: There are substantive differences in the nature of data collected across different disciplines in the sciences and social sciences. Differences also exist in the ways that data are maintained and used. These differences will have implications for standards development. Societies in some disciplines have begun to evaluate how data standards might be developed and used. Any new standards efforts should be undertaken collaboratively to take advantage of both existing work and special expertise.

3. Data Curation: In order to make data citation and reuse practical, there must be a measure of confidence about the persistence, accessibility and trustworthiness of the data. Good data curation practices are needed to create a reliable environment for data sharing. Some funding agencies, such as NIH, require grant recipients to deposit research data on completion of research projects. The experience of these agencies, including data submission guidelines and procedures for data access, could be instructive for future standards development. In addition, certain universities have recognized the importance of data curation and are developing programs in this area (e.g., UIUC offers a Master of Science Specialization in Data Curation: http://www.lis.uiuc.edu/programs/ms/data_curation.html). Such programs could provide insights useful for advancing standardization activities.

4. Scale of Data Collection: The size of the data collection will impact how it is curated and how it is reused. Standards for reuse of large data collections should include provisions for reuse and citation of portions of the data set as well as entire data sets.

5. Versioning: Versioning represents a significant challenge to data reuse. Unlike journal articles, data collections are typically dynamic and change as new data are collected and/or

modified through subsequent research. It is important to ensure that a citation will direct researchers to the same version of the data that was cited. Guidelines and best practices are needed for managing and citing data versions.

6. Role of Data in Scholarly Research: Academe regards publication of scholarly journal articles as important for career advancement. The same value is not currently assigned to the publication of research data. This reduces the incentives for making research data available. In addition, since data are an important part of the research process that underlies the findings published in journals, researchers often guard their data until all their results have been published lest another researcher use their data to 'scoop' them. This has implications for the availability of research data. While not directly related to standards, the lack of incentives for sharing data does have implications for standards development.

Key Recommendations

The group was in agreement that there are challenges to be met before effective standards can be developed and adopted, but the group also agreed that NISO can make useful contributions toward moving this forward. Following are the key recommendations.

1. Survey and summarize successful data management and citation conventions for existing data repositories.

By documenting existing practices, NISO can establish a foundation for subsequent standards development. This survey could be undertaken as a large project or as a series of short guidelines on specific topics. The survey could address best practices for data deposit, data citation, and data curation along with rights and restrictions for data reuse. In addition to using this effort to identify successful best practices, it could also promote these best practices and encourage wider adoption.

2. Develop a thesaurus of terms relevant to data sharing.

The lack of a common vocabulary can deter the collaboration needed for effective standards development. By developing a thesaurus, NISO could facilitate standards development. This thesaurus could include entries for data types, relevant metadata, experimental methodology, etc. and could indicate variations across research domains. There would also be value in developing an ontology for access, privacy and preservation policies related to citable data.

3. Develop guidelines for data citation.

The ability to accurately cite data collections is a key requirement for encouraging data deposit and reuse. This is important for both technical and cultural reasons. If data sets cannot be accurately cited, then researchers cannot locate the data for reuse and contributors cannot receive recognition for their work. Should NISO undertake development of a standard for data citation, it is recommended that this work be limited to a specific type of data such as

collections of observational data. It is further recommended that NISO identify and build on existing best practices in this area.

4. Work collaboratively with other organizations that are addressing these issues.

Some concern was expressed that this is a new area for NISO standards development and that NISO may not be well known to other organizations working in this area. It was also recognized that NISO's experience with information and publishing standards can provide valuable context for the development of standards for sharing research data.

In order to overcome this constraint and make best use of NISO's capabilities, it was recommended that NISO work collaboratively with other groups already exploring standards for data sharing. Given that data sharing practices will vary across disciplines and data types, community collaboration will be help to ensure relevance and wider adoption of standards that are developed.

Further, because the sharing of research data is still in its early stages, it was recommended that NISO use collaborative technologies (e.g., WIKI) in documenting best practices and/or developing a thesaurus. This will allow the community to cooperate in the development and maintenance these documents so that they retain their relevance going forward.

Appendix A: Attendees

Clifford Lynch, Executive Director, Coalition for Networked Information
cliff@cni.org

Ellen Kraffmiller (Substituting for Merce Crosas, Director of the Dataverse Network)
ekraffmiller@hmdc.harvard.edu

Paul Uhlir, National Academy of Sciences
puhlir@nas.edu

Lars Bromley, AAAS Project Director
lbromley@aaas.org

Robert Tansley, Google Scholar Software Engineer
roberttansley@google.com

Jean Claude Bradley, Chemist Drexel University
bradlejc@drexel.edu

Camelia Csora, product manager for 2collab
C.Csora@elsevier.com

MacKenzie Smith - Associate Director for Technology, MIT Libraries - DSpace
kenzie@mit.edu

Stuart Weibel, Senior Research Scientist at OCLC
Weibel@oclc.org

Matthew J. Dovey, JISC
m.dovey@jisc.ac.uk

John Sack, HighWire Press (participating remotely)
sack@stanford.edu