MeS Workshop Report by Jane Greenberg and Sarah Carrier/SILS/MRC UNC
Submitted Dec. 11, 2008

Metadata for Scientific Datasets (MeS) Workshop

The MeS workshop was held on Thursday, Sept. 25, at the Dublin Core 2008 (DC-2008) Conference in Berlin, Germany. The workshop was open to all DC-2008 participants. The workshop provided an opportunity for people to converse about metadata for scientific data and associated challenges and solutions. The overriding goal was to explore interest in forming a Dublin Core community around the topic of metadata for digital scientific data.

The workshop description is posted both on the **DC-2008 website** (*see* workshop16) at: http://dc2008.de/programme/workshops; and the **SILS/Metadata Research Center's <MRC> website** at: http://ils.unc.edu/mrc/sci_metadata/. The MRC website also presents the workshop agenda and provides access to the presenters' slides.

The workshop was motivated by the fact that research funding agencies in the US, Europe, Asia, and Australia are increasingly attentive to the curation of scientific data sets, and recognize that the full value of research investments rely not just on adequate preservation, but also *metadata* supporting sufficient description and structure such that their value can retained more readily for scientific pursuits. The workshop CfP announcement outlines several of key challenge in this area, where metadata can play a fundamental role.

- *Canonical identification of datasets*. [Metadata is] Critical for establishing provenance, auditing value and use, and attracting social-networking attention that will enhance their value.
- *Unfound data is unused data*. Datasets that cannot be discovered will not be reused, and the value of the data is thus unavailable for further exploitation.
- *Unstructured data is difficult to use*. Datasets should be designed and structured for reuse at the time of experimental design.

The workshop had five parts: 1. Introduction, 2. Presentations, 3. Commentary, 4. Open discussion, and 5. Charting the next steps.

1. Introduction

The workshop started off with a welcome by Jane Greenberg (SILS/MRC UNC); she introduced the workshop, reviewed the agenda, and stated the overriding goal of exploring the possibility of establishing a Dublin Core community specific to the theme of metadata for scientific data. Jane noted that Stu Weibel (OCLC) was instrumental in conceptualizing the workshop, and helping her to plan the meeting; but, unfortunately, he was not able to attend the Berlin conference.

There were 35 participants, and individual introductions where held off until after the presentations, during the *round-robin session.*

2. Presentations

Four brief presentations covered a range of metadata issues central to scientific datasets and to help inform the latter discussion.

- **John Kunze** (California Digital Library) presented "Data Curation: an NSF DataNet Perspective." He outlined key issues of a major distributed data curation approach, mapping how various partners might address interoperability and data preservation challenges via metadata. The ideas presented are mapped out in an NSF DataNetONE proposal (*pending*), and seeks to develop scalable, flexible, sustainable access to data, using new technologies and approaches.

- **Jörg Holetschek** (Botanischer Garten und Botanisches Museum Berlin-Dahlem) presented "Metadata on Primary Biodiversity Information," and emphasized biodiversity informatics. He began by defining biodiversity and biodiversity data, and underscored a chief goal of metadata research in this area is to unify data into one system. He reviewed the GBIF (http://www.gbif.org), and BioCASe: Biological Collection Access Service (http://www.biocase.org/) efforts. In his conclusion, he noted intellectual challenges of distinguishing metadata and data, and the impact these differences have on system operations and functionality.

- **Jane Greenberg** (SILS/MRC UNC) presented, "The Dryad Repository (http://www.datadryad.org/repo/) for Data Underlying Scientific Publications." She highlighted Dryad's application profile for scientific datasets; an instantiation study examining the application of bibliographic relationships for managing datasets; research on scientists' PIM (personal information management) metadata behaviors (see White's DC-2008 poster: http://dc2008.de), and the HIVE (Helping Interdisciplinary Vocabulary Engineering) project (https://www.nescent.org/sites/hive/Main_Page) for dynamically integrating vocabularies encoded in SKOS.

- **Robin Rice,** (EDINA and Data Library, University of Edinburgh) presented, "DISC-UK DataShare and Institutional Data Repositories" and spoke about challenges describing datasets housed in institutional data repositories. Her presentation underscored the need for data librarian expertise, and to combine their skills with that of repository managers. She also shared some lessons learned in the process, such as the need to keep things simple, demystify data for librarians, decipher various definitions for data, and encourage communication about the development of metadata for scientific datasets to those creating data sets. (Robin presented a poster at DC-2008 on aspects of this work: http://dc2008.de/wp-content/uploads/2008/10/12_rice_poster.pdf).

3. Commentary

After the presentations, P. Bryan Heidorn (NSF Program Officer, Division of Biological Infrastructure), briefly commented on NSF's goal to fund the type of work presented in the workshop, and several underlying intended outcomes of NSF's DataNet program. He also noted the need to educate professionals about data curation, highlighting work at UIUC (Univ. of Illinois at Urbana-Champaign) and relating this work to several points emphasized in Robin Rice's presentation. He then summarized three needs that ran across all of the presentations; these included: 1. the need to use metadata standards, including a shared standard for datasets; 2. the need for metadata describing datasets not to be too cumbersome to produce or sustain; and 3. the need to have metadata be interoperable with a range of standards—technical, operational, and so forth. The commentary concluded with an outline of the following questions that were meant to stimulate discussion.

Questions for discussion:

1. Which data collections need Dublin Core?
2. Who are the users of those datasets?
3. How does Dublin Core help users in their tasks?
4. Are there particular sciences that need Dublin Core more than others?
5. Considering institutional repositories and discipline repositories, where does Dublin Core fit in for each?

4. Open discussion, determining next steps

Following the presentation of questions, there was a *round-robin/3 minute madness session* where each person gave their name and institutional affiliation; and had the opportunity to state a metadata issue or concern relating the questions posed above (by P. Bryan Heidorn) or other issues.

Among discussion issues, it was noted by several people that that institutional repositories are proliferating; federal money is being directed more toward domain repositories; and interoperability between these two types of is lacking. One participant noted that institutional repositories are not as populous as one might anticipate.

One participant urged that in further discussions in this area take note of the individual's role, emphasizing that "users" (individuals wanting to find and use domain resources/data) differ from "depositors" (individuals wanting to share resources/data and potentially collaborate).

A comment was made about limitation of institutional federated resources for sharing data, and noted a *Dublin Core extension or application profile for scientific data* may address some of the current problems. It was noted that UKOLN at the University of Bath in the UK is carrying out a JISC commissioned study addressing the feasibility of an application profile for scientific data.

One commenter raised a question about "international issues"—noting that challenges don't only stem from the difference between institutional and domain repositories, but also global participation and language differences.

A final discussion involving several participants focused on issues of granularity, noting the value of rich detailed schemes supported by geospatial reference metadata, compared to general problems, including lost data, when information is often buried in natural text passages associated with data, but not marked-up as metadata.

5. Charting the next steps

The last segment of the workshop included an *all-participant* discussion about interest in forming a Dublin Core community that would serves as a venue for sharing ideas. There was also discussion about potentially developing a *Dublin Core extension or application profile for scientific data*, an undertaking that would require forming a task force. The **Dryad metadata application profile** was noted as a model here. The discussion highlighted the need to conduct a feasibility study as a first step for developing an application profile for scientific data. The question was called about interest in forming a community focused on science and metadata, and a vote was taken. There was unanimous support for establishing a Dublin Core community for metadata for scientific datasets and several people volunteered to help conduct a feasibility study.

Action Items

- To summarize meeting notes, and post the notes and slides from all of the presentation on the SILS/MRC (metadata research center) meeting homepage @: http://ils.unc.edu/mrc/sci_metadata/.
- Propose a Dublin Core community specific to the topic of metadata for scientific datasets, possibly taking on the name of (MeS).
- To identify community leaders for this proposed group.
- To set up a listserv, and send email to all workshop participants to inform them of the group's establishment, electronic mailing list, and homepage if it is approved by the DC Advisory Board. (If the community is not approved, it would also be good to inform all workshop participants of the reasons given, and suggest appropriate next steps for people to further engage in discussion of the topic of metadata for scientific datasets.