# Metadata Workflows Across Research Domains: Challenges and Opportunities for Supporting the DFC Cyberinfrastructure

Adrian Ogletree
Drexel University
adrianogletree@gmail.com

**Keywords:** metadata workflows; metadata generation; DataNet Federation Consortium (DFC); research data; cyberinfrastructure.

## 1. Introduction

This poster presents research results from a survey studying metadata workflows. In the context of this study, a 'metadata workflow' is defined as a workflow that generates metadata for a data collection. The following research question guided this investigation: Where are people (and automated processes) creating metadata in the data life cycle, and what could be done to improve the quality?

## 2. Background

Metadata is necessary to find, use, and properly manage scientific data. Sharing metadata workflows across different communities is thus crucial for promoting data interoperability and reuse. The DataNet Federation Consortium (DFC) is a project within the NSF Office of Cyber-Infrastructure DataNet initiative. One widespread problem that the DFC seeks to address is the unfortunate reality that "many scientific fields lack a common integrated data infrastructure, which often results in non-standardized, local data management practices" (Akmon, 2011, p. 330-331). Carole Goble, Robert Stevens, Dave De Roure, and others have made significant contributions to the study of e-science workflows and reproducibility. In addition, Taverna and Kepler are two open-source, community-driven, scientific workflow management systems with large user bases in the eScience community (Taverna; Kepler). However, data management needs vary substantially across disciplines. Willis, Green, and White (2012) call for future research to examine in greater detail the "community-specific practices and workflows as well as constraints caused by the technological environment and trends at the time of scheme creation" (p. 1517).

## 3. Methodology

A survey was distributed via e-mail to the DFC listserv in order to better understand how scientific metadata is created. DFC scientists, researchers, and data curators involved in any aspect of creation or use of scientific metadata were invited to participate in this study.

## 4. Results and Discussion

Fourteen (14) participants responded to the survey, representing a 34% response rate (the DFC listserv contains 41 members). They were affiliated with eight different DFC project partners: the Ocean Observatories Initiative (OOI),[1] the iPlant Collaborative,[2] the Odum Institute for Research in Social Science,[3] the National Oceanic and Atmospheric Administration (NOAA),[4] the Renaissance Computing Institute (RENCI),[5] the University of Virginia, the Data Intensive Cyber

[1] http://oceanobservatories.org/
[2] http://www.iplantcollaborative.org/
[3] http://www.odum.unc.edu/odum/home2.jsp
[4] http://www.noaa.gov/
[5] http://www.renci.org/

Environments (DICE) Center,[6] and the School of Information and Library Science at the University of North Carolina at Chapel Hill. The participants' fields of study included hydrology, biology, climatology, ecology, library sciences, computer science, engineering, social sciences, and information science. The composition of the participants' positions were as follows: 2 professors, 1 associate professor, 1 assistant professor, 1 postdoc researcher, 1 doctoral student, 2 master's students, 2 administrators, 1 software engineer, 1 scientific analyst, and 1 IT project team lead (one participant did not respond to this question). Five (5) of the participants had 5 to 10 years of research experience.

The following types of data were created or used in the participants' research: observational data (7), papers (7), simulation data (4), laboratory experimental data (3), "other" (3), and field experimental data (1). Participants were asked to select all that apply. Observational data has the most long-term value for researchers because it is often unique, irreplaceable, or costly to collect (Anderson, 2004).

Figure 1 below shows metadata creation by a person and metadata creation or capture by a computer. Participants were asked to select all that apply; for instance, some researchers add metadata at every point within the data collection process. Eight (8) of the participants who responded to this question manually create metadata before data is collected, 10 manually create metadata during data collection, and all 12 manually create metadata afterward. Only 2 of the participants report that computer-generated metadata is created before data is collected; 9 report that automated metadata creation occurs during or after data collection, with one respondent selecting "other," who had no automated metadata collection. Data management best practices recommend that data documentation happen at the very beginning of the research project, before data collection. However, these results indicate that more scientific metadata is created during or after the data collection process than before, and that few researchers take advantage of automated metadata generation workflows.

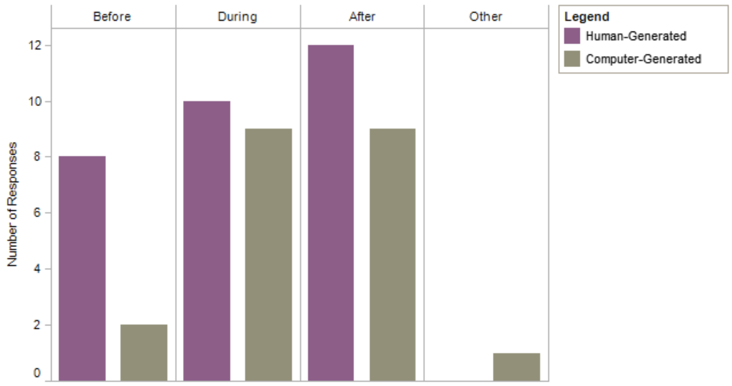

FIG. 1. Metadata creation by humans and automated processes.

Six (6) of the participants reported that their organization has a specified standard in place for creating metadata. The following metadata schemes were used: Dublin Core (7), "Other" (7), FGDC (2), NetCDF Climate and Forecast (CF) (2), "Don't know" (1), EML (1), and "No standard scheme is used" (1). Participants were asked to select all that apply. Six (6) of the participants who selected "other" named the following metadata schemes: free tag AVU in irods, MIxS, DDI (2), WaterML, and GML. Based on the survey results, many different metadata schemes were used, consistent with Greenberg's (2005) study of digital repositories that "hundreds of metadata schemes [are] being used, many of which are in their second, third, or *n*th iteration" (p. 18).

[6] http://dice.unc.edu/

When asked what information another researcher would need to reproduce their research, responses include: information about workflows, highly specialized knowledge, software, or equipment, and/or algorithms and parameters used. Similarly, Borgman (2012) observes that research reproducibility requires "the precise duplication of observations or experiments, exact replication of a software workflow, degree of effort necessary, and whether proprietary tools are required" (p. 17). Without contextual information and high-quality metadata, even "open" data is unusable.

## 5. Conclusions

Overall, the results met expectations based on other similar studies of scientists' data management practices and perceptions (Akers, 2013; Anderson, 2004; Borgman, 2012; Chavan & Penev, 2011; Greenberg, 2005). The following list represents the key findings of this survey:

- More than half (58%) of participants create or use observational data
- Metadata is more likely to be created after data collection
- Scientists and researchers suffer from a lack of awareness of metadata standards
- Data sharing is complicated by the need for highly specialized knowledge, software, and/or equipment in order to reproduce research

This study makes a contribution towards methods of survey design for the purposes of studying metadata workflows. Although the responses to this survey represent multiple scientific disciplines, positions, and institutions, this study was limited by the small sample size. Future research should include larger populations, and different research domains can be categorized in order to study the similarities and differences of data management needs between communities. Another area of interest for the DFC is the ability of the iRODS data grid to capture the provenance information associated with execution of a workflow. This research could be useful for creating a definition of a sufficient context to enable re-use of data.

## Acknowledgements

I would like to thank Jane Greenberg and Reagan Moore for their guidance and support. The DFC DataNet is supported by the National Science Foundation, OCI-0940841.

## References

Akers, Katherine G. and Jennifer Doty. (2013). Disciplinary differences in faculty research data management practices and perspectives. The International Journal of Digital Curation, 8(2), 5-26. doi:10.2218/ijdc.v8i2.263

Akmon, Dharma, Ann Zimmerman, Morgan Daniels, and Margaret Hedstrom. (2011). The application of archival concepts to a data-intensive environment: Working with scientists to understand data management and preservation needs. Archival Science, 11(3-4), 329-348.

Anderson, William L. (2004). Some challenges and issues in managing, and preserving access to, long-lived collections of digital scientific and technical data. Data Science Journal, 3, 191-201.

Borgman, Christine L. (2012). The conundrum of sharing research data. Journal of the American Society for Information Science and Technology, 63(6), 1059-1078. doi:10.1002/asi.22634

Greenberg, Jane. (2005). Understanding metadata and metadata schemes. Cataloging & Classification Quarterly, 40(3-4), 17-36. doi:10.1300/J104v40n03_02

Kepler. Retrieved from https://kepler-project.org/

Taverna. (2014). Retrieved from http://www.taverna.org.uk/

Willis, Craig, Jane Greenberg, and Hollie White. (2012). Analysis and synthesis of metadata goals for scientific data. Journal of the American Society for Information Science and Technology, 63(8), 1505-1520.