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

Adrian Ogletree
Metadata Research Center, Drexel University
https://cci.drexel.edu/mrc/

INTRODUCTION

Metadata is necessary to find, use, and properly manage scientific data. Sharing metadata workflows across different communities is thus crucial for promoting data interoperability, reproducibility of results, and reuse. Improving the infrastructure that supports scientific data archiving contributes to digital stewardship efforts and ensures long-term access to scientific research data. This poster presents research examining the methods and processes used to acquire provenance, descriptive, and administrative metadata by domain scientists, and can contribute to the larger goal of achieving metadata interoperability within the DataNet Federation Consortium.

BACKGROUND & RATIONALE

Research Question

Where are people (and automated processes) creating metadata in the data life cycle, and what could be done to improve the quality?

Metadata Workflows

A **metadata workflow** is a workflow that generates metadata for a data collection. Data management and workflow tasks give metadata value and allow data to be reproducible.

DataNet Federation Consortium (DFC) Goals

- Implement a national data grid
- Enable collaborative research on shared data collections
- Enable reproducible data-driven research
- Integrate "live" research data into education initiatives

Literature on Research Data Management

"[The topic of metadata] provided further examples of serious inconsistency both within and across the disciplines. It seems that not only is there a body of researchers who have still to grasp the purpose and importance of metadata but, where the need for good metadata is understood, this does not necessarily translate into the sufficient use of standard structures." (Pryor, 2007, p. 143)

METHODOLOGY

Survey Questions

Demographics
- Time involved with DFC
- Affiliation within DFC
- Field of study
- Position
- Active research years
- Data types used

Work Practices
- When metadata is created by a person
- When metadata is created or captured by automation
- Standards in place for creating metadata
- Researchers adding metadata manually
- Metadata schemes utilized
- Metadata provided along with data to repositories
- Types of metadata included
- Tools that analyze their data

Open-Ended
- Information needed to reproduce their data
- Overarching research questions
- Additional comments about the survey and/or metadata workflows

SURVEY RESULTS

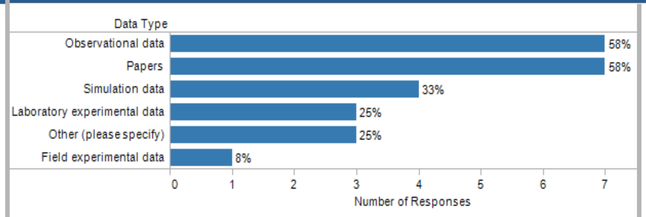

Figure 1: Depicts the data types created or used in research, including laboratory experimental data, field experimental data, observational data, simulation data, papers, or other. Most participants use observational data and papers as data sources.

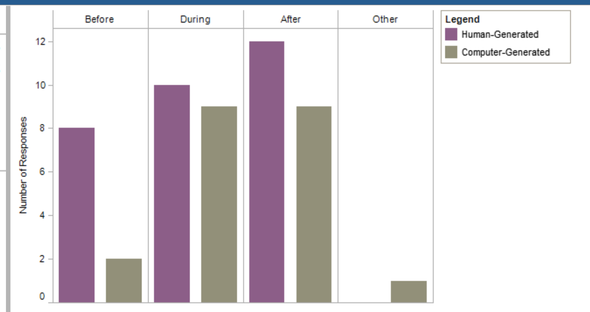

Figure 2: A major goal of this survey was to determine when metadata is most commonly created within the data life cycle. Figure 2 illustrates the similarities and differences between metadata creation by a person and metadata creation or capture by a computer. Both events are more likely to occur towards the end of the data collection process, but human-generated metadata is more common.

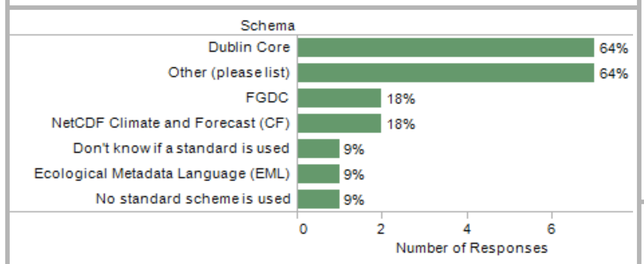

Figure 3: Depicts metadata schemes used. Dublin Core was the most commonly used metadata scheme, and 9% of respondents were unaware whether or not a metadata standard is used at their institution. See Table 1 for additional responses to this question. There are almost as many metadata schemes listed as there are respondents to this survey.

Our sensor data sets have specific metadata that is outside Dublin Core
free tag AVU in irods
MIxS
ddi
WaterML, GML
DDI

Table 1: Free-text responses of those who answered "other" in Figure 3.

Responses to Open-Ended Question
Need the workflows that generate the data set, the input files and parameters, and the output from the workflow to verify reproducibility.
Lots of specialized computing and virtual reality equipment as well as access to human subjects
Significant understanding of computing environment, executable programs, staging of data products in appropriate location
database version from which data was extracted, tool version and parameters with which analysis was done
Workflows, data provenance
Information about model set up including algorithm, programming language, initial, and boundary conditions.
I have not done intensive research, so I do not yet have any data that require reproduction. However, if I did, I would want metadata pertaining to the methodology, weights, and any specific stats software I used to analyze my data.

Table 2: Displays responses to the open-ended question, "What does another researcher need to reproduce your data set?" Most responses reference workflows, highly specialized knowledge, software, or equipment, and/or algorithms.

*Participants were asked to select all that apply

Responses to Open-Ended Question
An example is the TDLC. Each year they collectively examine which research questions should be explored. As they gain knowledge, the set of questions become more specific.
Full body tracking with no latency and no emcumbrances. Field of view filling head-worn displays with no latency, swimming, jitter, eyestrain, or nausea. Automated scenario and model (geometric and humanbehavior) generation for training VEs. THEN we could start evaluting and comparing VE-based training to live training.
Methods to better understand impact of environment on genotype/phenotype and vice versa for adaptation to change in environment.
What are the relationships between organisms' and species' distributions, their traits, and their environments.
+ science collaboration through data and software sharing + standards based data and metadata storage system design
Version tracking of data, provenance, and ranking of data based on their quality
My main interests align with data management education and getting researchers on board with data management. Making metadata creation/addition a more streamlined and easy process would definitely help with getting researchers interested in data management. So, I suppose the big questions for me would be: What steps are being made to make this easier/faster for researchers? How do we get the most metadata with the least amount of effort from our researchers without requiring them to learn about metadata standards, etc?

Table 3: Responses to the question, "What are the big questions in your field that you and/or other scientists would like to solve?" The purpose was to gain an understanding of the research projects represented, as well as potential applications of improved data management and workflows.

CONCLUSIONS

- More than half (58%) of participants 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

Acknowledgements | The DFC DataNet is supported by the National Science Foundation, OCI-0940841.