Apr 6th, 12:30 PM

Developing a Data Management Plan (DMP) in the Cognitive Sciences

Jessie Cass
Simmons College, Graduate School of Library & Information Science, cassj@simmons.edu

Follow this and additional works at: http://escholarship.umassmed.edu/escience_symposium

Part of the Cognitive Psychology Commons, Communication Sciences and Disorders Commons, and the Scholarly Communication Commons

This work is licensed under a Creative Commons Attribution-Noncommercial-Share Alike 4.0 License.
OBJECTIVE: To experience the process of using principles of scientific research data management (SRDM) to work with a researcher to create a data management plan (DMP). SRDM is an area where research in the traditional sciences intersects with information science. SRDM guides researchers through all stages of the data life cycle. A DMP is a document explaining how a study will progress through the data life cycle that is increasingly required by research funders. This project was undertaken as part of a class on SRDM through the Simmons College School of Library and Information Science.

METHODS: After corresponding via email with a researcher studying the cognitive and linguistic skills of deaf children with autism, a set of questions was created based on an interview instrument developed by the Digital Curation Centre and a Skype interview was conducted. Using the information gathered during the interview and in follow-up emails, as well as knowledge of SRDM principles learned in class and through independent research, a DMP (following National Science Foundation guidelines) was created. Additionally, aspects of the researcher’s study which proved challenging when creating a DMP were identified. RESULTS: A seven-part DMP was created. Challenging aspects were identified as a set of teaching points. These included: data being collected via video camera; children as subjects; subject IDs; repository requirements. CONCLUSIONS: This project was successful in teaching both this author and the interviewed researcher about SRDM and DMPs. This will improve the cognitive science community’s understanding of the principles and importance of SRDM.

Challenges Throughout the Data Cycle

A major part of the study in question required videotaping the subjects. This brought up two interesting issues. The first is that the video camera used was recording onto obsolete hardware. A related hardware malfunction caused data to be lost and the lack of support due to the obsolescence of the hardware meant that the data could not be recovered. Additionally, videotaped data are often inherently personally identifiable information as they often show subject’s faces and/or voices.

One of the funders for this study was interested in the researcher submitting the data to a repository. This could be a way for the data to be useful to additional people, for both analysis and re-use. Unfortunately, for several reasons, including the videotaped nature of much of the data, and difficulty obtaining appropriate consent, these data were not ultimately deposited. If the researcher had known about the repository’s strict requirements from the beginning and planned for that throughout the data cycle, it is possible that things could have been done differently, resulting in successful re-use of data.

All of the subjects in the study were children. This complicated the researcher’s ability to share gathered data. The researcher had the parents of subjects sign consent forms, however these did not give the researcher permission to submit the data to a repository. If the researcher had been looking at the whole life cycle of the data from the beginning, the wording of the consent form could have been different.

Preserving Data

One issue the researcher had early in the data life cycle was with the naming of files and folders. To maintain confidentiality, the researcher needed to use IDs, rather than names, when referencing subjects and naming files. In this case the researcher used the sex of the subject, the relevant diagnosis of the subject, and the chronological number of the subject to make an ID. This was problematic because it would be possible for someone to figure out which subject went with which ID, and therefore which data, as long as they knew the order in which data were taken and the sex and diagnosis of the subject. A common alternative solution is to use a number randomizer to create IDs that are unrelated to any aspect of the subject. These IDs are written in a protected spreadsheet in order to keep track of which subject has which ID.