THE DATA INTERVIEW

Jennifer Walton

jwalton@mbl.edu
MBLWHOI Library
7 MBL Street
Woods Hole MA 02568

Abstract
Librarians are adapting their skills to the emerging new workflows in scholarly communications. The publication of data is an emerging topic for both librarians and scientific researchers. This paper demonstrates some ways that librarians may expand their existing skill in this area.

Keywords: Data Librarian, data management, libraries – data processing

The data management interview is an emerging topic for librarians; however, the number of librarians performing interviews remains low. Librarians have traditionally struggled with marketing their services in an effective way. Additionally, researchers have traditionally been the guardians of their data, so it can be difficult for librarians to establish a foothold in this area. However, librarians have skills and perspectives that researchers do not. This paper looks at the traditional library skills that can be adapted for use in the data interview and some methodologies that are emerging

The data interview can be viewed as a type of reference interview. Bopp & Smith (1995) define the reference interview as a "conversation between a member of the library reference staff and a library user for the purpose of clarifying the user's needs and aiding the user in meeting those needs."

This is often true with data. Making one's data accessible is a new area for many researchers. WHOI's data library collected data from the WHOI ships for years; however, technology has changed how those data can be shared and mandates have followed the technology. Scientists have always had spreadsheets of data and lab notebooks. But now researchers are being asked by funding agencies and publishers to come up with plans to make their data accessible to others. One of the truths of data management is that science did not make the transition from paper to electronic in a considered way. Scientists not necessarily know where to begin to met the mandates and make data openly accessible.

The first thing in any data interview is reference interview 101. Listen to what the researcher says and try to establish the real need. Use the reference interview tools:

- Paraphrase their needs back to them.
- Ask open questions.
- Clarify.
- Verify .

Just as when working at the reference desk, the librarian does not need to have all the answers, but does need to have the skills to help patrons find the answers.

Librarians have long been creating and discussing metadata. MARC and RDA are great examples of metadata. There are standards for creating the vocabularies or ontologies behind metadata with Authority files and classifications. Dublin Core came out of the library tradition and out of that grew Darwin Core, both of which are regularly used metadata schemes for data sets. Most data repositories have established their own metadata policies. Metadata schemas can range from extremely complex to very flat. Decisions on metadata may depend on how much interoperability or automation is needed. Although the schemas or ontologies may be unfamiliar to librarians, the basic tenants of metadata do not change

Librarians are already often responsible for Institutional Repositories IRs), which may be used for data in some cases. IRs can be useful for researchers submitting to journals that require them to supply the data behind the publication. The submission process has a few more steps and is better considered earlier in the process than a typical article submission. At MBLWHOI we are able to assign a doi for the datasets in our repository, which may be a condition of submission as well.

Institutional repositories are not the best choice for all data. However, a librarian familiar with the IR can assist researchers submitting data to another repository such as Figshare or Dryad. It is important for the librarian to help researchers find the best fit for their data, but it is not essential to memorize every subject repository, metadata schema, or submission process.

Librarians are very familiar with copyright laws and licensing in our countries. There are some nuances to dealing with data. For instance, data, at least in the U.S., is not copyrightable, but the arrangement of data is and can be licensed as well, as we are well aware of with library databases. There are several online tools including the open data commons that can be used as references for librarians to use when discussing Intellectual Property with researchers. In addition, librarians may remind researchers that the rules for citation apply in the case of data as well.

There are several ways that librarians in the U.S. are responding to this expansion of their responsibilities.

The E-Science movement in libraries includes liaison or embedded librarians who work with researchers through the scholarly communication cycle. This style of data interview involves a librarian who already knows quite a bit about the research and the researcher. The librarian may be working with researchers and students in the lab or classroom to ensure that good data management practices are being followed. NLM recently awarded grants to medical librarians to work as informationists on NIH funded projects.

Data Curation Profiles are specific sets of materials and instructions from Perdue University and UICU created through an IMLS grant that included a series of workshops describing their protocol, as well as making the materials available online. In 2007, Perdue and the GSLIS at University of Illinois began to research the needs of researchers in sharing data and how librarians could help them. In 2009, they published their findings and created the Data Curation Profiles Toolkit.

The profiles are designed as a structured framework for librarians. There is a set of step-by-step instructions to follow, including a script for the librarian and supplemental materials for the researcher. This can be beneficial for librarians who are not familiar with the material, but it can be time intensive. They also recommend taping the interviews with the researcher's permission and then transcribing the interview. This can be beneficial for institutions that may need forms of assessment tied to data management in the same way as with other library services.

Institutions without many resources and dedicated programs may choose to create portals to help assist researchers creating data management plans and identifying resources. Librarians may still be available to help researchers; however, it is only part of their job. Many libraries have created web pages to do this, often as a way to become involved with data management. This can help demonstrate to researchers the ways that the library is a resource for them.

A checklist and an outline for the interview process are helpful for the data interview process. The checklist will help you if the conversation does not follow your outline. The list should be as detailed as necessary to gather the needed information. The following topics are suggested as important parts of the data interview.

> *WHY?* The conversation should begin with this question: why does the researcher want to create a data management plan or make their data accessible? If the need has already been articulated, paraphrasing the answer is still useful: "The Proceedings of the Royal Society Proceedings B is requiring you to submit the data supporting your article, so to make those data available you want to add them to the Institutional Repository." This will inform the rest of the reference interview as well as define the scope of the project at hand. It can be

useful to narrow the scope to a manageable size as opposed to creating and implementing a data plan for all research. However, this can also be an opportunity to learn generally about the research and to gain specific information about the task at hand.

Once the basics are known, the interview may proceed in an order that is comfortable or intuitive; however, it should likely include the following topics:

- Description of Data. This will include: types of files created;, essential metadata; what metadata are recorded automatically; do researchers consistently note the equipment used; their methodologies; file naming conventions; etc. This is also a good time to ask what issues they have in their data collection and management.
- Lifecycle. This can be as complex or simple as is applicable. It is important to understand the whole cycle of the researcher's data and where the data under discussion fit in. This can be fairly complex. The Purdue module for "The Lifecycle of the Data Set" is a good reference for this portion of the interview. Some questions to consider here would be:
  - Where the data at hand fit with the data for the entire project; how are the raw data used or stored; how many stages of data do they store; is this a static, closed dataset; is versioning a concern; how often are the data updated; and how do the different stages of their data need to be preserved? Researchers may envision having access to their data forever. In some disciplines such as long-term ecological data, the value may lie in preserving the raw data. However, with the vast amounts of data being produced, the cost of storing and maintaining may not be feasible. It is important for researchers to critically look at their data and make a guess at their lifespans.
  - Sharing and Access. When describing sharing and access, many scientists voice fears concerning re-use of data out of context or using them to refute findings. As a result, researchers may want to maintain their role as gatekeepers of their data. In many cases, this may mean allowing others access via FTP to data on institutional servers. However, due to mandates from a funder, a publisher or in some cases their community, they must make their data openly available. Although librarians cannot change this mandate, they can help by providing resources on this topic, such as opendatacommons.org and Dryad's use of the Creative Commons Zero Waiver. Librarians can also remind researchers of data citation, long-term access and DOIs.
- It is also important to discuss metadata and standards. If the data are going into a repository, there will be standards set by the repository that need to be investigated.
- Measuring Impact or Assessing Use of the Data. Assessment requirements or preferences will influence decisions on where and how data is deposited and should be discussed in the initial interview.

The interview is the beginning of the process. After the initial interview, the librarian should expect to continue the conversation with the researcher and move forward on the decisions based on the information garnered in the interview.

The scholarly process is in a period of change. Data accessibility, reuse and value are topics confronting many researchers.

References

Bopp, Richard E. and Smith, Linda C. *Reference and Information Services: An Introduction*, Second Edition. Englewood, CO: Libraries Unlimited, 1995.

E-Science Portal for New England Librarians. http://esciencelibrary.umassmed.edu/index.

NLM Administrative Supplements for Informationist Services in NIH-funded Research Projects. http://www.nlm.nih.gov/ep/AdminSupp.html.

Data Curation Profiles Toolkit. http://datacurationprofiles.org/.