December 2007

Conducting a Data Interview

Michael Witt
Purdue University, mwitt@purdue.edu

Jake R. Carlson
Purdue University, jakecar@umich.edu

PURDUE UNIVERSITY **Libraries** *Access. Knowledge. Success.*

INTRODUCTION

Librarians at Purdue University are beginning to identify the scientific datasets that are being generated by our faculty and researchers as information assets to be collected, preserved, and made accessible as a function of the library's collection development. These librarians are subject-area specialists, and many have advanced degrees in their respective disciplines in addition to a degree in library science. They have all been trained in collection management; however, much of this training was related to traditional formats such as monographs and serials and not datasets. In our experience, one of the most effective tactics for eliciting datasets for the collection is a simple librarian-researcher interview. In this poster, we share a set of ten questions that a librarian can use as a starting point for such a "data interview". It is not a comprehensive strategy but instead a practical tool to draw out information that needs to be considered in order to evaluate the suitability of a dataset for the collection and the requirements for the infrastructure and services that will be needed for data curation.

#1 What is the story of the data?

Begin the interview with an open-ended question that allows the researcher to talk freely about his or her research, scientific workflow, and community of practice. This lends some insight into the value of the dataset and how it may fit into your collection and be used, and it provides the ***context*** for understanding how and why the dataset was created and how it was processed and analyzed.

#2 What form and format are the data in?

What computing environments (e.g., software) are required to use the data? If the data are in proprietary structures, you may consider reformatting them into agnostic formats or ones that can be more easily ***re-versioned***. Is there any existing ***metadata***, either external to the data or description that could be extracted from it? Ideally the data could be described to be discoverable by researchers from another discipline.

#3 What is the expected lifespan of the dataset?

In many cases, there are distinctions in the utility of a dataset as it begins in a raw state and then is analyzed and processed into new forms and versions as a result of different steps in the research workflow. Different entities may have custody of the data and use it for different purposes at different times, affecting its ***provenance***. Funding agencies may require that data be archived for a prescribed period of time or you may forecast its future value and the amount of time it should be retained. The data may be described and archived for effective ***preservation*** to ensure its accessibility and integrity over time.

#4 How could the data be used, reused, and repurposed?

This is a primary ***selection*** criterion that also impacts how the data are ***accessed*** and what ***policies*** may be needed to govern its use. As data are archived and shared, new and unintended uses for the data may increase its value. For example, a research dataset may be repurposed as a learning object.

#5 How large is the dataset, and what is its rate of growth?

It is important to quantify the size of the data for storage and network provisioning if you intend to ***ingest*** it into your repository. What is its physical (bits) and logical (records) ***scale***? Is the dataset static or dynamic? Ask for a sample of the data to examine.

#6 Who are the potential audiences for the data?

Information regarding potential users of the data and the users' needs is paramount. Along with potential uses for the data, this is another primary ***selection*** criterion. In some cases, the data may need to be embargoed or restricted to a limited group of users who are granted ***permission*** to access it.

#7 Who owns the data?

Establishing and maintaining the ***intellectual property*** represented by the data should be discussed at the earliest opportunity, and any conflicts should be resolved up-front. Many organizations have a submission policy that asks the contributor to verify that they own the data and have the right to submit it.

#8 Does the dataset include any sensitive information?

All data should be reviewed for information that violates ***confidentiality***, such as identification information on human subjects. Data curation activities should be informed by institutional review board requirements.

#9 What publications or discoveries have resulted from the data?

The researchers may have a bias regarding the importance of their data. The purpose of this question is to establish an objective metric for determining the value of the data for the collection. Different metrics may be more appropriate in determining the ***selection*** criteria for different kinds of data and data collections.

#10 How should the data be made accessible?

There is value in making data accessible using a conventional web-based user interface, but machine-to-machine interfaces should also be evaluated. These ***methods of access*** will be informed by the answers to the previous questions, and this question can be asked in an open-ended manner to fill in any gaps remaining at the conclusion of the interview.

SUMMARY

Although building robust collections of datasets present several complexities and challenges to resolve, the process of looking at scientific datasets as information assets and exploring what is needed to develop and manage data collections is similar to the traditional collection development practices that have been successfully employed by librarians for decades. We offer these ten "data interview" questions as a springboard for librarians to explore data curation in greater depth and specialization.

Michael Witt (mwitt@purdue.edu)
Assistant Professor of Library Science

Jake Carlson (jrcarlso@purdue.edu)
Data Research Scientist

Purdue University Libraries
Distributed Data Curation Center
http://d2c2.lib.purdue.edu

D2C2 Distributed Data Curation Center

"Conducting a Data Interview"

Michael Witt & Jake Carlson, Purdue University Libraries, West Lafayette, Indiana, USA