Demystifying the Data Interview

Eugenia Kim - Data Curation Intern, Purdue University Libraries - eugenia.kim24@gmail.com

Jake Carlson - Data Research Scientist - Purdue University Libraries / Distributed Data Curation Center - jrcarlso@purdue.edu

QUESTION

What would a typical librarian would need to know ahead of time in conducting an interview with a faculty researcher regarding his/her research data and associated needs in managing or curating this data?

METHODS

• The authors analyzed the DCP Toolkit to identify specific definitions, concepts, resources, and examples to provide a basic understanding of data curation.

• By compiling this information with additional resources for reference, the authors intend to help librarians take full advantage of the DCP Toolkit.

ABOUT THE DATA CURATION PROFILES

• From 2007 - 2010, the Purdue University Libraries and the Library School at the University of Illinois conducted research on "Which researchers are willing to share data, when, with whom and under what conditions?"

• Research was supported by the Institute of Museum and Library Services

• The results led to the creation of the Data Curation Profiles (DCP) Toolkit

• The DCP Toolkit is designed to assist librarians in identifying the data management needs of researchers.

Data

Purpose: A clear definition for "data" is essential for setting the context of a data interview.

Challenges: The term "data" is often defined very broadly. It may mean different things to different people. Some disciplines, such as the Humanities, may not think in terms of "data".

Examples of Data: notebooks, data files, images, e-mails, surveys, videos, transcripts, drafts of publications.

Selected Definition: "Research data is defined as the recorded factual material commonly accepted in the scientific community as necessary to validate research findings" -OMB Circular A-110, Source: http://www.whitehouse.gov/omb/circulars_a110.

Data Lifecycle

Purpose: Librarians need to have a reference guide to the stages that data go through as part of the lifecycle. Understanding the lifecycle helps librarians identify and understand researcher data curation needs.

Challenges: Each researcher may have their own definition of "data lifecycle" that may not apply to other research. Additionally, not all data may require curation.

Selected Approach: The stages of a data lifecycle can be broadly categorized into Raw, Processed, Analyzed, and Published. For an example of a data lifecycle, see the graphic to the right.

Humphrey, Charles. "e-Science and the Life Cycle of Research" (2006) Retrieved 4/20/10: http://datalib.library.ualberta.ca/~humphrey/lifecycle-science060308.doc

Data Librarian

Purpose: Librarians need to be able to explore how their knowledge, skills and perspective could be applied towards supporting researchers' needs without feeling overwhelmed.

Challenges: Data librarians may come from all kinds of backgrounds and may be called upon to fulfill a wide range of roles requiring a variety of skillsets. Furthermore, the position of "data librarian" has not yet been established at many institutions or may be tied to traditional library services rather than curation.
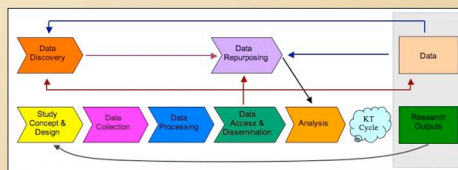
Base Identity: Broadly defined, a librarian with responsibilities relating to the curation and maintenance of data.

Data Curation

Purpose: To understand or be familiar with the tasks and responsibilities that are generally associated with the term 'data curation'. Specific activities associated with curation are "archiving" and "preservation".

Challenges: Not every dataset follows the same lifecycle or understands curation in the same way. A librarian may also find that a researcher asks for a task to be performed *after* the interview or has an unusual type of need.

Selected Definition: "The activity of managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purpose and available for discovery and re-use." - Lord, Macdonald, Lyon & Giaretta (2004) "From data deluge to data curation." Proceedings of the UK e-Science All Hands Meeting 2004, 31st August - 3rd September, Nottingham UK.

Data Sharing

Purpose: Allows for transparency of how funding is spent, accountability for outcomes and their validity, and advancing research innovations.

Challenges: Researchers may be resistant to sharing data for multiple reasons, including the lack of preparation of the data for viewing by others, issues of ownership, control over its use, IP rights, and the lack of incentives to publish one's data.

Selected Approach: Rather than define "data sharing" itself we seek to provide enough background information on the challenges listed above for librarians to ask questions of researcher.

RESULTS

•Data concepts and terminology and meaning are often vary between various fields.

•Concepts do not necessarily have examples that can be easily referenced by librarians.

•These results informed the development of the Data Curation Toolkit Workshops.

CONCLUSION

• The results of this project will be further analyzed through:

o Assessing the impact of the training delivered in the DCP workshops.

o The effect of the workshop and DCP toolkit in enabling librarians to conduct data interviews.

o Other sources such as the community forums on the DCP Toolkit website.

RECOMMENDED RESOURCES

• Data Curation Profiles Community http://www4.lib.purdue.edu/dcp/
• Resources for Digital Curators | Digital Curation Centre http://www.dcc.ac.uk/resources/
•Digital Curation | ICPSR (Inter-University Consortium for Political and Social Research) http://www.icpsr.umich.edu/icpsrweb/ICPSR/curation/
•Digital Curation: A How-To-Do-It Manual | Ross Harvey http://www.neal-schuman.net/curation/
•Digital Curation Exchange | A Space for all things 'digital curation' http://www.digitalcurationexchange.org/

Demystifying the Data Interview

Eugenia Kim, Data Curation Intern
Purdue University Libraries,
eugenia.kim24@gmail.com

Jake Carlson, Data Research Scientist
Purdue University Libraries/Distributed Data Curation Center
jrcarlso@purdue.edu

Data

Purpose: A clear definition for "data" is essential for setting the context of a data interview.

Challenges: The term "data" is often defined very broadly. It may mean different things to different people. Some disciplines, such as the Humanities, may not think in terms of "data".

Examples of Data: notebooks, data files, images, e-mails, surveys, videos, transcripts, drafts of publications.

Selected Definition: "Research data is defined as the recorded factual material commonly accepted in the scientific community as necessary to validate research findings" -OMB Circular A-110, Source: http://www.whitehouse.gov/omb/circulars_a110.

Data Librarian

Purpose: Librarians need to be able to explore how their knowledge, skills and perspective could be applied towards supporting researchers' needs without feeling overwhelmed.

Challenges: Data librarians may come from all kinds of backgrounds and may be called upon to fulfill a wide range of roles requiring a variety of skillsets. Furthermore, the position of "data librarian" has not yet been established at many institutions or may be tied to traditional library services rather than curation.

Base Identity: Broadly defined, a librarian with responsibilities relating to the curation and maintenance of data.

Data Curation

Purpose: To understand or be familiar with the tasks and responsibilities that are generally associated with the term 'data curation'. Specific activities associated with curation are "archiving" and "preservation".

Challenges: Not every dataset follows the same lifecycle or understands curation in the same way. A librarian may also find that a researcher asks for a task to be performed *after* the interview or has an unusual type of need.
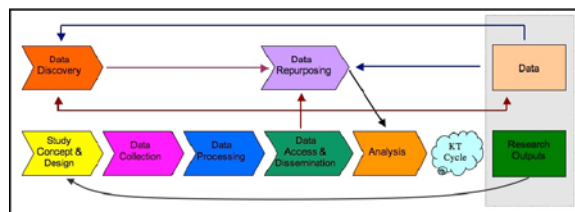
Selected Definition: "The activity of managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purpose and available for discovery and re-use."
- Lord, Macdonald, Lyon & Giaretta (2004) "From data deluge to data curation." Proceedings of the UK e-Science All Hands Meeting 2004, 31st August - 3rd September, Nottingham UK.

Data Lifecycle

Purpose: Librarians need to have a reference guide to the stages that data go through as part of the lifecycle. Understanding the lifecycle helps librarians identify and understand researcher data curation needs.

Challenges: Each researcher may have their own definition of "data lifecycle" that may not apply to other research. Additionally, not all data may require curation.

Selected Approach: The stages of a data lifecycle can be broadly categorized into Raw, Processed, Analyzed, and Published. For an example of a data lifecycle, see the following graphic.

Humphrey, Charles. "e-Science and the Life Cycle of Research" (2006) Retrieved 4/20/10: http://datalib.library.ualberta.ca/~humphrey/lifecycle-science060308.doc

Data Sharing

Purpose: Allows for transparency of how funding is spent, accountability for outcomes and their validity, and advancing research innovations.

Challenges: Researchers may be resistant to sharing data for multiple reasons, including the lack of preparation of the data for viewing by others, issues of ownership, control over its use,

IP rights, and the lack of incentives to publish one's data.

Selected Approach: Rather than define "data sharing" itself we seek to provide enough background information on the challenges listed above for librarians to ask questions of researcher.

Recommended Resources
Data Curation Profiles Community
http://www4.lib.purdue.edu/dcp/

Resources for Digital Curators | Digital Curation Centre
http://www.dcc.ac.uk/resources/

Digital Curation | ICPSR (Inter-University Consortium for Political and Social Research)
http://www.icpsr.umich.edu/icpsrweb/ICPSR/ curation/

Digital Curation: A How-To-Do-It Manual | Ross Harvey
http://www.neal-schuman.net/curation/

Digital Curation Exchange | A Space for all things 'digital curation'
http://www.digitalcurationexchange.org/