10-11-2011

Demystifying the Data Interview: Developing a Foundation for Reference Librarians to Talk with Researchers about their Data

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

Demystifying the Data Interview: Developing a Foundation for Reference Librarians to Talk with Researchers about their Data

Jake Carlson
Data Services Specialist
Purdue University
jrcarlso@purdue.edu

Introduction

Interest in working with research data as an information resource is growing amongst academic libraries. However, research data sets and the issues surrounding making them accessible are much more complex than what librarians encounter with the materials that typically comprise library collections. These complexities can present a potential barrier for librarians seeking to engage in discussions with researchers about managing, sharing, and curating their data. Without possessing a significant depth of knowledge of the research process, or a strong understanding of data practices, norms and challenges within a particular field, librarians may be at a loss in how to begin.

In 2007, a team from the Purdue University Libraries and the Graduate School of Library and Information Science at the University of Illinois sought to gain a better understanding of the needs of researchers in sharing their data, and how librarians could potentially help address these needs. With support from the Institute of Museum and Library Services (IMLS), this team sought to identify "which researchers are willing to share their data, when, with whom, and under what conditions?" through interviewing science and engineering faculty. The findings of this research were shared as Data Curation Profiles (Witt *et al.* 2009). Each Data Curation Profile contains a description of a particular data set and its lifecycle, an account of how the researcher administers, shares or curates the data, and what the researcher would like to do with the data set but is not currently. In other words, his or her specific needs for the data set.

To assist librarians and other information professionals seeking to identify the needs of researchers in managing, sharing or curating their data, the Purdue Libraries have developed the Data Curation Profile Toolkit (DCP Toolkit). The DCP Toolkit provides the means for librarians to conduct data interviews with an individual researcher or small lab group and to construct Data Curation Profiles of their own. Information about the project, the DCP Toolkit, as well as completed profiles, can be accessed from the project's website: http://www.datacurationprofiles.org.

The DCP Toolkit was developed with the intention that any librarian or information professional would be able to make use of it. However, recognizing that conducting interviews with researchers about their data would be unfamiliar terrain for many librarians, the Purdue University Libraries, with additional support from the IMLS, developed a workshop to introduce librarians to the DCP Toolkit, explain how it could be used, and prepare them for conducting data interviews of their own. This workshop is now being offered at multiple locations in the United States. Although the workshop is open to any type of information professional, reference librarians were identified as likely

attendees early on. Reference librarians already have some relevant training and experience in conducting interviews, and many reference librarians already have developed relationships with researchers at their institutions through subject liaison responsibilities. As a result, the curriculum and the content of the workshop were developed with direct consideration of the needs of the front-line reference librarian.

A particular challenge in developing the workshop was the need to determine what base level of knowledge about data issues would be needed to enable librarians to conduct an effective data interview. In other words, what would a "typical" librarian need to know before conducting an interview with a faculty member regarding his/her research data and associated needs? To answer this question, the workshop development team analyzed the components of the DCP Toolkit to determine what specific concepts and definitions would need to be covered, sought out resources and examples that could be used to provide this level of knowledge, and finally determined how to incorporate this knowledge into the lesson plan of the workshop.

It should be noted that the goal of this investigation was to support the learning objectives of the workshop specifically, and not to provide librarians with a foundation in data curation work generally. The DCP Toolkit is meant to enable librarians to initiate discussions with faculty about their data and their related needs, and the primary purpose of the workshop is to prepare librarians to have these discussions. Therefore, the concepts, definitions and examples that were adopted had to be those that could easily be recognized and understood by both librarians and faculty. Furthermore, they would have to be relevant to a wide audience as the DCP Toolkit is meant to be an all-purpose tool that could be used to interview researchers from most any discipline.

Background

The starting point for determining what librarians would need to know to conduct a successful data interview were the findings from the research done by Purdue and the University of Illinois. The data interviews conducted in this project revealed a great deal of variation in the types of data researchers were willing to share and a wide range of potential concerns, requirements and desired services. For instance, the majority of the researchers in the study indicated a need to restrict access to their data for some period of time, or placed conditions on their willingness or ability to share their data with others. However, the length of time before a researcher would release the data and the exact conditions for release varied across participants (Witt, 2009). In addition, it was found that gaining an understanding of the nature of the specific data set under discussion and its lifecycle was a crucial aspect of determining researcher needs. The nature and form of the data at each stage in its lifecycle affect the researcher's perceptions of its likely value to others, and his or her willingness to share. In sum, the researcher's willingness to share data publicly with others hinged not only on disciplinary and sub-disciplinary cultures of the researcher, but on a range of individual considerations as well (Cragin, *et al.*, 2010).

The findings of this small-scale study echo the results of other research efforts to examine the behaviors and practices of researchers in handling and sharing their data. The Digital

Curation Center conducted case studies with multiple researchers from sixteen disciplines to examine the differences in sharing, reusing and preserving research data. One of the primary findings of this study was that disciplinary examinations of data practices were too broad in scope to be able to understand and explain researcher's actions and attitudes sufficiently. Observed variations in multiple areas, including the wide range of data types, research methods, data curation practices, and skill sets in managing data, led investigators to conclude that needs and requirements are best understood at the sub-disciplinary level or even finer levels (Key Perspectives, 2010). This conclusion is reinforced by the results of another series of case studies of information use and exchange between researchers in the life sciences. Investigators in this study observed that, although information exchange and use were taking place through a wide range of formal and informal channels, these cases of information exchange were best understood at a granular level of analysis. The cases of exchange were intricately structured and could not be fully understood through a simple linear or cyclical model (RIN, 2009). A recent survey conducted by DataONE found significant variation in data management and sharing practices based on multiple factors beyond the researcher's discipline. These factors include the researcher's primary funding source, age, work focus, and location (Tenopir, *et al.*, 2011).

These findings pose significant challenges for agencies that are developing or maintaining repositories to enable the sharing, curation or preservation of research data. Traditionally individual researchers have functioned as the "gatekeepers" of their data, deciding when, with whom and under what conditions to share their data. Disciplinary communities and funding agencies are now pushing towards developing repository infrastructures to share research data more openly and at a larger scale. Depositing data into repositories requires that researchers relinquish their role as the "gatekeeper" of their data and transfer it to repository managers. If data repositories are to succeed in attracting submissions from researchers, repository developers and managers will need to be able to understand and respond to the needs and requirements of individual researchers. The demonstrated variations in researcher needs and requirements across disciplines, sub-disciplines, and amongst individuals insure that this process will not be a trivial undertaking. Data repositories will need assistance from people who are trained in conducting data interviews to understand the data and elicit requirements, and then negotiate and help prepare the submission of data into the repository.

Defining Roles for Librarians

With this in mind, the first task in developing the workshop was to articulate roles for librarians in helping to address issues in managing, sharing, and curating data. A vision was needed to describe how librarians could have an impact in addressing the challenges identified in the literature, and how the Data Curation Profiles Toolkit could be used to promote this vision.

The workshop is predicated on the belief that librarians, reference librarians in particular, are well-suited to raise awareness and identify researcher needs; skills that are essential given the diversity and variability of these needs. Libraries occupy a unique space in academia as they are charged with supporting the research and teaching activities of all

disciplines and departments at their home institution. As a result, libraries typically build services and collections to address a wide range of diverse information needs across a multitude of disciplines. In support of this work, reference librarians seek to engage with their constituencies, striving to develop relationships with individual faculty, administrators, students, and others at their institution. These individual relationships further inform the work of the libraries and enable further refinement of services and collections provided to address the specific information needs of clientele. The ability of reference librarians to work both within and across disciplines, to develop trusted relationships with faculty based on an understanding of their individual needs, and to cross administrative boundaries and bring different constituencies together are key elements in addressing the challenges described in working with data.

A foundation for this perspective is provided by research on content recruitment for institutional repositories. Around the turn of the century institutional repositories were introduced as a means to increase institutional prestige and as an alternative model of scholarly publishing, one in which faculty would gain more control over their work (Johnson, 2002). Despite the initial excitement and fanfare over institutional repositories many have languished due to a lack of contributed content (Davis & Connolly, 2007). The central problem being that institutional repositories services and software were developed without much consideration of the value propositions or direct needs of the faculty who were supposed to make use of them (Salo, 2008). A study done by the University of Rochester examined the disconnection between repository services and faculty needs through direct observations of faculty work practices. Their findings led to a reassessment and redesign of their institutional repository model and a new approach for recruiting content (Foster & Gibbons, 2005). A central component to their new approach is to train their liaison librarians on the features, benefits, mechanics, and context of their repository services, so they in turn can leverage their existing relationships with faculty to encourage and facilitate content submission to the repository (Bell, *et al.*, 2005). Other libraries are also turning to their reference librarians to assume significant roles in making connections between faculty and institutional repository services as a part of their liaison responsibilities (Chan, *et al.*, 2005; Palmer, *et al.*, 2008).

Perhaps informed by experiences with institutional repositories, the literature on possible roles for librarians in working with research data is recognizing the potential applicability of the skill sets possessed by reference librarians. For example, in November of 2008 attempts were made to identify a core set of skills for data librarians at the DCC sponsored Research Data Management Forum held in Manchester, England. The skill set for data librarians included several that are standard for conducting reference work: negotiation skills, coordination of practice across an institution, advocacy, promotion, marketing, raising awareness, and complaints and expectation management (Pryor, 2009). Anna Gold notes that some reference/subject librarians have incorporated data services into their work, particularly in the Social Sciences and geospatial data. Gold argues that what is needed now is an expanded scope of librarian's involvement with research data. Librarians have the opportunity to work both "downstream" in the data lifecycle, through providing discovery, selection, acquisition, and licensing for data sets, and "upstream" in supporting the use of documentation, best practices, or standards in the

production of data as collaborative partners with faculty (Gold, 2007). Tracey Gabridge at MIT looks to the work done by librarians in building and maintaining institutional repositories to inform roles in working with data. She believes that librarians must collaborate with others to build effective data curation systems and deliver appropriate data services through these systems. Gabridge believes that the subject liaison function of librarians can be reconfigured to extend library services to data curation (Gabridge, 2009). The Purdue University Libraries developed and carried out a pilot program to identify how the responsibilities of subject liaison librarians might translate into working with data sets in an institutional repository context. Although additional skills will be required of librarians seeking to develop and steward collections of data, the results of the pilot project place the relationship between liaison librarians and their faculty as an important foundation in working with data (Newton, *et al.*, 2010).

In addition to the literature, direct experiences in working with faculty at Purdue University have also informed this perception of the role of a data librarian. The focus of the Purdue Libraries has been on making connections with researchers working at the "small science" scale. In contrast to "big science" which takes place at a large scale and is well funded, "small science" is conducted on a limited budget by one lead researcher with the possible assistance of a collaborator, support staff, and a few graduate students. Small science constitutes the majority of the research done in the STEM fields at Purdue, which is likely to be the case at most research universities. "Big science" research may have resources and expert staff available, researchers engaged in "small science" self-report that they lack the means, time, and often the skills to address data curation by themselves (Heidorn, 2008).

Many of the "small science" researchers interviewed at Purdue are not used to giving much thought about what would be required to enable the dissemination, curation or preservation of their data. Furthermore, existing repository models often feel alien to researchers, as these models generally do not demonstrate how researcher needs and perspectives will be accounted for and represented in a repository in language that researchers will easily understand. Data management is typically performed by students and is likely to consist of local measures such as saving data to hard drives in the lab or backing up the data on to CDs. While students have received "research integrity" training, which includes on making data available upon request to the funder, publisher, or FOIA, etc., it is unlikely that they could produce a data set that would be usable by others easily or quickly (Brandt, 2010).

Equipped with knowledge about a data set and the needs of a researcher gained from a data interview, it is envisioned that reference librarians could assume the role of a trusted data consultant by working with researchers to help them navigate through what for them may be uncharted territory. The involvement of a reference librarian would extend across the lifecycle of the data, from the development of a data management plan that satisfies the needs of a funding agency, to following community based standards and practices in generating and managing the data, to the deposit of the data into a repository to ensure long-term access. Through building off on their existing roles, reference librarians are in a strong position to help researchers locate and understand relevant data tools, services,

standards and then to provide support to researchers in making appropriate use of these resources. Where solutions to the data needs identified in an interview do not yet exist, reference librarians could identify or form collaborations within or beyond their institution to help plan, design or create them. Echoing a call for reference librarians to engage in the recruitment of content for institutional repositories, reference librarians could take on the role of a data publishing associate. Through developing an understanding and knowledge of existing data repositories, the services they provide, and their requirements for submission, reference librarians could help prepare the researcher and their data to ensure a smooth transition of the data "gatekeeper" function from the researcher to an appropriate repository.

Defining "(Research) Data"

Success in this role depends on a librarian being able to talk with researchers about their data in ways that are understandable and meaningful to them. The term "data" is often defined very broadly and conceptualized differently depending on context. What constitutes data may be interpreted differently by different people at different times. Furthermore, data as a term is often associated with numerical, tabular data by default. Some disciplines, particularly in the Humanities, may not think in terms of working with "data". Therefore, establishing a clear understanding of what constitutes data is an essential precursor to any data interview. Without a shared definition of what data are the very premise of a data interview between librarian and researcher may be misunderstood.

The definition of data put forth in the workshop comes from the Office of Management and Budget Circular A-110, which reads: "Research data is defined as the recorded factual material commonly accepted in the scientific community as necessary to validate research findings." (OMB, 2011) This definition is broad enough that it could cover a wide variety of materials, yet compact enough to clearly delineate boundaries; it is also widely used by government funding agencies, and would likely be familiar to many researchers already. A potential problem exists with the use of the term "science community" in this definition, which may be seen as limiting. Therefore, this term is presented in the workshop in the broadest sense to include disciplines outside of the physical and natural sciences.

A data interview conducted using the DCP Toolkit is meant to capture the perspective of the researcher, and therefore must be driven by the researcher being interviewed, not the librarian. Although the definition of data provided by the OMB is sufficiently broad to serve as a backdrop to inform the data interview, ultimately each researcher will determine his or her own understanding of what is meant by data. The workshop includes a discussion of possible data types that were identified in data interviews conducted at Purdue to illustrate the broad variety of what a librarian may encounter. These data types included:

- "experimental & theoretical; raw numbers, algorithms, images; sometimes initial states that allow data reproduction"
- "notebooks (print & e-), data files, images; mostly "processed" data, some raw; Microsoft files and emails"

- "wide variety, from image to tapes to notes to bio-samples (not all on the computer)"
- "human records: surveys, videos, transcripts"

Defining "Data Set"
Researchers often work on multiple projects that generate multiple types of data for different purposes, uses and even audiences. A researcher may have different needs associated with the different types of data that he or she is generating. Therefore it is important to distinguish precisely which data are meant to be the subject of the interview before the discussion about the data begins. As different data may have different issues or challenges associated with them, limiting the focus in this way is meant to ensure that the needs expressed by the researcher are those that pertain unambiguously to a specific data set. However, discerning what constitutes a "data set" precisely can be a difficult process as the term does not have a universally accepted definition in scientific and technical literature (Renear, *et al.*, 2010).

In the workshop "data set" is defined as: the data collected and analyzed for a specific project or problem. The precise data set that will serve as the focus of the data interview should be negotiated with the researcher beforehand. A data set may still be comprised of multiple components, or data types. For example a series of text files, Excel spreadsheets and Matlab files may all be present within a particular data set. Not all of the data types that comprise a data set will have equal importance or value to the researcher for curation purposes. Therefore, it is important that the librarian be able to determine what constitutes the researcher's "primary" data versus the "ancillary" data as a part of the interview. In the workshop, primary data is defined as the data that are generated or analyzed specifically to achieve the project results. Ancillary data are defined as any additional data that are brought in or generated to assist in explaining or understanding the primary data, but are not used for research purposes directly. An example of primary data could be sensor data on the rate of traffic flow at a selected intersection. An example of ancillary data could be the weather conditions that are reviewed to potentially explain possible anomalies in the traffic flow data.

Defining "Data Lifecycle"
A data set is not typically "born" fully formed and complete. The published data that appear as a table or graph in a journal article will look very different from when the data was first generated. The data lifecycle identifies the stages that data will pass through and describes the transformations that occur at each stage.

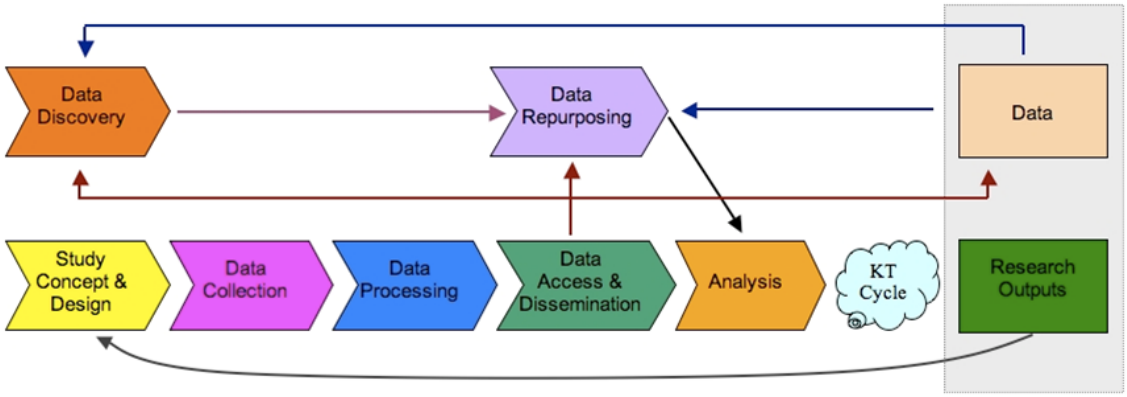
The data lifecycle is a useful approach for librarians to use as a framing device in a data interview for several reasons. First, researchers often identify their work with data as a series of stages which the data pass through. Second, conceptualizing the researcher's development and use of data as a series of stages within a lifecycle naturally supports discussions on the process, methods and tools used to work with the data at each stage. It is important to capture this information in order to ensure a more complete understanding of the data and the associated needs for its curation, and it would be more difficult to discern this from a general discussion. Finally, approaching the data interview from a

lifecycle perspective facilitates the identification of which elements in a data set the researcher may be willing or able to share with others and which may be targeted for curation.

Although useful, the idea of a data lifecycle can be difficult to employ as a part of an approach to a data interview. Every data lifecycle is different depending on the needs, aims and approaches of the researcher being interviewed. Therefore attempts to pre-define the lifecycle of a data set in an interview would likely result in a distorted view of the data and the researcher's needs. As the data interview is meant to be driven from the perspective of the researcher, the section of the DCP Toolkit that covers the data lifecycle prompts the researcher to define and describe the stages him or herself. However, this approach presents a challenge to the librarian conducting the interview as he or she must ensure that the discussion about the data lifecycle is as rich and complete as possible. Librarians who are just beginning to explore this area with researchers may not be familiar enough with data processes, workflows, and transformations to be able to articulate likely stages, or know if they are inadvertently overlooking some elements of the lifecycle.

Introducing librarians to the idea of a data lifecycle is an important element of the workshop. An example of a data lifecycle was needed to illustrate possible stages that may comprise a lifecycle; however the example would have to be relevant beyond a particular project or discipline. The example would also have to explicitly include data sharing and curation components. "The Life Cycle Model of Research Knowledge Creation" graphic developed by Charles Humphrey at the University of Alberta provides such an example (figure 1). Mr. Humphrey's graphic depicts the data lifecycle as a set of discrete stages and transition points where data loss may occur. His graphic also illustrates curation as a natural part of the data lifecycle and provides an example of how curation components may fit into it (Humphrey, 2006).

Figure – The Lifecycle Model of Research Data Creation



Although Mr. Humphrey's graphic serves as a solid foundation for introducing data lifecycle concepts, no example will be applicable to every researcher's real-world practice. Providing a means for librarians to be able to identify stages within a data lifecycle that is unique to the researcher being interviewed and likely unfamiliar to the

librarian presented a challenge. In the research conducted to develop the Data Curation Profile, a broad pattern amongst the researchers interviewed was observed. Most of these researchers included four types of stages to some degree in their description of their data lifecycle: a "raw" stage, a "processing" stage, a stage for "analyzing" the data in some way, and a "publishing" stage. The "raw" stage referred to the beginning of the process during which the data were generated or collected in some fashion. In the "processing" stage the data were subjected to some form of cleansing or preparatory actions in order to make them suitable for use by the researcher. In the "analyzing" stage, data were tested or transformed to provide information that would support or refute the researcher's hypothesis. In the "publishing" stage, a summarization of the data that best illustrated the researcher's work were created with the intent of dissemination in some way, shape, or form (generally as a component of an article or book).

These high-level commonalities in characterizing research stages were incorporated into the workshop. However, teaching the stages of the data lifecycle required further consideration. Just because these stages were observed broadly in many interviews conducted at Purdue does not mean that they would be present in every interview, especially in disciplines outside of the sciences and engineering. Furthermore, these stages may be exhibited to varying extents in real-world research practices. Some researchers may engage in several iterations of "processing" for example. Others may obtain their data from external sources in a state in which they are ready to be analyzed with minimal or no additional processing. Still others may perform stages that fall outside of these four loosely defined activities, such as reconciling different data types with each other. Therefore, the data lifecycle model and the four broad data stages are introduced at the workshop with careful explanation and caution about their use in real-world settings. These concepts are meant to serve more as guides than as rules. Examples and hands-on exercises for participants were developed to better convey these concepts in the workshop.

Defining "Data Curation"

Data curation is a term that seems to have acquired multiple meanings depending on the author and their particular perspective. These varying meanings can easily lead to confusion, especially in cross-disciplinary discussions. For example, in talking with some engineering faculty at Purdue, it was found that, for them, curation focused on quality control issues and review functions in selecting data to be added to a database. This is a more specialized definition of curation than would typically be employed by most librarians (Mullins, 2010).

Definitions of data curation employed by the library and information science field vary as well. It is not uncommon to see the terms "digital curation" and "data curation" used interchangeably. Some definitions of data curation incorporate archival and preservation functions (Lord, 2004), (UIUC, 2011), while other definitions explicitly separate these functions and define them independently from data curation (McGovern, 2009). The lack of a universally accepted definition both within and outside of the library and information science field is a hurdle that has to be recognized and addressed in the workshop.

Librarians need to be able to articulate clearly what is meant by their use of the word "curation" to the researchers they will be interviewing to ensure a productive discussion.

To address the idea of "data curation" in the workshop, the subject is approached from a broad vantage point by looking at the common elements of existing definitions. The definition provided by Phillip Lord is introduced as a broad framing device. His definition is well known in the field and is frequently cited, making it a useful starting point. The components of Lord's definition that are emphasized in the workshop are the management and promotion of data from the point of its creation, ensuring the fitness of data for contemporary purposes, and making data available for discovery and re-use (Lord, 2004). His inclusion of archival and preservation functions as components of curation are acknowledged but noted as being controversial. Another perspective on curation from the business world is then introduced. Steve Rosenbaum in the June 15, 2010 issue of Business Insider proclaims that "curation is king". By this he means that in the internet age content is no longer a specialized commodity, anyone can produce content. Instead, those who are adding value to content through enabling its discovery, aggregation, organization, and other curation functions, are today's movers and shakers (Rosenbaum, 2010). The focus on the core elements of data curation: planned management over time, availability for discovery and re-use, and adding value to enable or further usage, provides enough of an introduction to prepare librarians to hold discussions with researchers without getting overly bogged down in specific interpretations.

Defining "Data Sharing"

Investigating a researcher's willingness and ability to share their data set outside of their lab is at the heart of a data interview. However, given the potential number and diversity of possible researcher needs and concerns with sharing their data set, it can be difficult for a librarian to feel confident enough to discuss these issues in a data interview. Therefore, a significant portion of the workshop is spent providing a general introduction to some of the more common needs and issues mentioned by the researchers interviewed at Purdue and the University of Illinois. This type of instruction is designed to provide the librarian with enough background information to anticipate some of the issues that a researcher may raise and then to be able to navigate through the subsequent discussions.

In the workshop, data sharing is broadly defined as a researcher providing access to, making available, publishing, disseminating, or allowing others to view, access, or make use of their data. This definition is purposely loose as is it meant to include instances where the data set under discussion may already be shared with others on a small scale or through informal channels. For example, a researcher may share some of their data set through email to a colleague who attended a presentation of the researchers work. Discussing the nature and extent of sharing that has been done in the data interview may help to identify acceptable practices for the researcher and his or her peer groups, as well as to introduce discussion on potential needs.

The data interview questions in the DCP Toolkit are designed to ascertain when in the data lifecycle the researcher would be willing to share the data, with whom, and why.

Additional interview questions also seek to identify the potential audiences for the data set and the likely value of the data set to these audiences. The scope and complexity of these issues can make them difficult to convey to a novice audience. To illustrate these complexities to workshop participants, video clips of a data interview that took place between a librarian and a professor of Agronomy at Purdue are shown. In these clips, the Agronomy professor describes her willingness to share her data set after it has gone through a "processing" stage with anyone, provided that the data are described sufficiently for the potential audiences to be able to understand and make use of the data effectively. She also identifies researchers in agronomy, policy makers, and commercial enterprises as the likely audiences for her data set and explains how it might be useful for each group (see "Appendix A"). After viewing this clip attendees are asked to do an exercise in which they use an excerpt from the interview worksheet completed by the agronomy professor and part of the transcript of the interview to compose a section of a Data Curation Profile. Participants are then asked to share and discuss their work with each other. The approach of presenting a model data interview between a librarian and a faculty member as a component of the workshop helps connect librarians to high-level concepts from a real-world perspective.

In addition to addressing data sharing directly, the data interview in the DCP Toolkit contains several modules that address issues that indirectly relate to sharing a data set. Two of these modules are highlighted in the workshop: "organization and description", and "intellectual property". The approach used in presenting these modules is not to define these areas so much as to provide a brief description of some of the important issues and challenges associated with them. The objective is to provide workshop attendees with a sufficient level of information to enable them to understand issues that may arise during the data interview and to be able to pursue areas of interest with the researcher they are interviewing.

The purpose of the "organization and description" module of the data interview is to determine how the data set is currently organized and described, to identify any shortcomings in this area (from the perspective of the researcher), and to begin to determine whether there are community-based standards that could be applied to address these shortcomings. From Purdue's experience, it is fairly common for researchers to have organized and described their data set only to the extent that is needed for people closely associated with the research to be able to understand and make use of the data set. Researchers have varying degrees of understanding about metadata, but often do not have a sense of what metadata should be applied to their data set to enable it to be discovered, understood, administered or used by others. Similarly, librarians may have at least a base knowledge of what metadata is, but may not have an understanding of how it comes into play in supporting data sharing and curation functions. Librarians also need to have an understanding of the importance of standards generally, not just metadata standards, to enable effective curation. The workshop aims to provide attendees with enough of an understanding of these issues that they could feel comfortable discussing them when they are introduced in the data interview.

Intellectual property rights and protections is another subject that presents a set of potentially thorny issues that may affect a researcher's willingness or ability to share their data set. The data interview questions relating to intellectual property issues in the DCP Toolkit include ownership over the data, identifying the stakeholders and their possible influences over the data, the researcher's need for any particular terms of use, and attribution. Other issues related to intellectual property could arise during the interview, and so librarians should be prepared to discuss them if necessary. The workshop touches on some of these issues including copyright and its applicability to data, open access principles for data sharing, and privacy concerns for data involving human subjects, to introduce librarians to these subjects.

An important point to convey to librarians is that the purpose of the data interview is to investigate, not to advocate. Pushing a particular course of action too soon is likely to be counterproductive. Before any recommendations can be made on the sharing, management, or curation practices for a data set it is important that a librarian and others involved have as rich an understanding of the researcher's situation, issues, and needs as possible. Once this understanding is attained through analyzing the content of the data interview, then a librarian and others involved may craft a response with recommendations as needed.

Conclusion

As interest in improving data management, dissemination and curation practices continues to grow, academic libraries are seeing opportunities to develop resources and services aimed at supporting the needs of researchers in the 21st century. In considering how to respond to these opportunities, libraries would be well advised to learn from their experiences with institutional repositories. The literature on institutional repositories demonstrates that services that do not align with real-world needs of researchers will not be used. Reference librarians have been brought in to help address the deficiencies in the initial service model of institutional repositories through leveraging their existing relationships with faculty towards increasing awareness of repository services, content recruitment, and providing assistance in submitting or accessing content. In assuming these responsibilities reference librarians are moving towards a new type of relationship with faculty, one in which they are taking on more of a direct partnership role in the publishing process (Bell, *et al.*, 2005).

As with institutional repositories, designing effective strategies to develop capacity for libraries to work with research data will depend upon effective engagement with researchers and building a solid understanding of their real-world needs. Librarians will need to move beyond our focus on researchers' needs as information consumers, and work towards building awareness of their disciplinary and sub-disciplinary information cultures and norms, and of their individual data practices within their research lab environments. Acquiring this depth of knowledge needs to be made a pre-requisite before new infrastructures or services for research data are developed. Conducting data interviews with researchers is one approach towards achieving this foundational understanding.

Reference librarians are potentially well-suited to conduct effective data interviews, but will they feel confident enough in their ability to do so? Some librarians have had experience working with data as an information resource, but for most librarians talking to researchers about their data is unfamiliar, and perhaps uncomfortable, territory. In developing the curricula for the workshop, the primary goal was to provide librarians, reference librarians in particular, with enough familiarity with data terms and concepts to give them the ability and confidence to engage researchers in a data interview using the Data Curation Profiles Toolkit. A significant challenge in teaching librarians the art of the data interview is achieving the right balance of training them in the mechanics and use of the DCP toolkit itself with providing enough information about the underlying concepts and terminologies for them to understand and use the tool effectively. Achieving this balance is made even more challenging by the presence of multiple definitions of terms, the diversity of data cultures and practices across or even within fields of study, and the still emerging conceptualization of what roles and responsibilities librarians will be willing and able to assume in supporting data management, sharing and curation. Furthermore, the data interview is meant to capture and deliver the perspective of the researcher being interviewed, not that of the librarian. Therefore, explanations and examples of data concepts and terminology have to be presented with broad brush strokes to provide the ample footing needed to launch discussions between faculty and librarians without boxing either of them into a particular perspective.

Initial feedback from the librarians who have attended the workshop indicate that generally the workshop has helped prepare them to conduct a data interview. The real measure of success for the workshop however, will be the quantity and perceived quality of completed data interviews. In addition to surveying workshop attendees, Purdue University will host a symposium on the Data Curation Profile Toolkit and librarians' engagement in data curation issues in May of 2012. Workshop participants who have conducted data interviews and developed Data Curation Profiles of their own will be invited to present their experiences, findings and the results of their work. Presenters at the symposium will be asked to participate in a focus group to discuss the challenges they encountered, future directions for librarian-faculty engagement in data curation, and what additional educational programs or tools may be needed.

As roles and responsibilities for libraries in working with data become more apparent the nomenclature surrounding data may become better defined. For now, there is a real need to develop paths to engagement through enabling librarians to better understand researchers' needs with data management, sharing and curation. The Data Curation Profile toolkit and the workshop that was developed to teach its use are an attempt to provide one such path for librarians and other information professionals.

Acknowledgements:
The author would like to thank Ms. Eugenia Kim who served as a Graduate Student Intern on the Data Curation Profiles project for her many contributions to this research and to the development of the Data Curation Profiles workshop. The Data Curation Profile Toolkit and workshops were developed with generous support from the Institute of Museum and Library Services.

References:

Bell, S., Foster, N.F. & Gibbons, S. (2005), "Reference librarians and the success of institutional repositories". *Reference Services Review*, 33(3), p.283-290. Available at: http://www.emeraldinsight.com/10.1108/00907320510611311 (accessed 11 October 2011).

Brandt, D.S. (2010), "Provost Fellowship: Final Report" Purdue University (unpublished).

Chan, D.L.H., Kwok, C.S.Y. & Yip, S.K.F. (2005) "Changing Roles of Reference Librarians: the Case of the HKUST Institutional Repository" *Reference Services Review*, 33(3), p.268-282. Available at: http://www.emeraldinsight.com/10.1108/00907320510611302 (accessed 04 October 2011).

Cragin, M.H., Palmer, C.L., Carlson, J.R., & Witt, M. (2010), "Data Sharing, Small Science, and Institutional Repositories", *Philosophical Transactions of the Royal Society A*, Vol. 368, No. 1926, pp. 4023-4038.

Davis, P.M. & Connolly, M.J.L. (2007), "Institutional Repositories: Evaluating the Reasons for Non-use of Cornell University's Installation of DSpace", *D-Lib Magazine* Vol.13, No. 3/4, available at: http://www.dlib.org/dlib/march07/davis/03davis.html (accessed 04 October 2011).

Foster, N.F. & Gibbons, S. (2005), "Understanding Faculty to Improve Content Recruitment for Institutional Repositories", *D-Lib Magazine* Vol.11, No. 1, available at: http://www.dlib.org/dlib/january05/foster/01foster.html (accessed 04 October 2011).

Gabridge, T. (2009), "The Last Mile: Liaison Roles in Curating Science and Engineering Research Data", *Research Library Issues: A Bimonthly Report from ARL, CNI, and SPARC*. No. 265, available at: http://www.arl.org/bm~doc/rli-265-gabridge.pdf (accessed 04 October 2011).

Gold, A. (2007), "Cyberinfrastructure, Data and Libraries, Part 2: Libraries and the Data Challenge: Roles and Actions for Libraries." *D-Lib Magazine* Vol.13, No. 9/10, available at: http://www.dlib.org/dlib/september07/gold/09gold-pt2.html (accessed 04 October 2011).

Heidorn, P.B. (2008), "Shedding Light on the Dark Data in the Long Tail of Science" *Library Trends* Vol. 57, No.2, pp.280-299. DOI: 10.1353/lib.0.0036, available at: http://muse.jhu.edu/journals/library_trends/v057/57.2.heidorn.pdf (accessed 11 October 2011).

Humphrey, C. (2006), "E-Science and the Life Cycle of Research", available at: http://datalib.library.ualberta.ca/~humphrey/lifecycle-science060308.doc (accessed 11 October 2011).

Johnson, R.K. (2002), "Institutional Repositories: Partnering with Faculty to Enhance Scholarly Communication", *D-Lib Magazine* Vol.8, No. 11, available at: http://www.dlib.org/dlib/november02/johnson/11johnson.html (accessed 04 October 2011).

Key Perspectives (2010), "Data dimensions: disciplinary differences in research data sharing, reuse and long term viability. SCARP Synthesis Study", Digital Curation Centre. Available at: http://www.dcc.ac.uk/scarp (accessed 11 October 2011).

Lord, P., Macdonald, A., Lyon, L. and Giaretta, D. (2004), "From Data Deluge to Data Curation" *Proc 3th UK eScience All Hands Meeting*, p.371–375, available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.111.7425&rep=rep1&type=pdf (accessed 11 October 2011).

McGovern, N. (2009), "Digital Preservation Management: Implementing Short-term Strategies for Long-term Problems" [workshop] ICPSR, May 2009, Ann Arbor, MI.

Mullins, J.L. (2010), "The Challenges of E-Science Data-set Management and Scholarly Communication for Domain Sciences and Engineering: a Role for Academic Libraries and Librarians", in Marcum, D.B. and George, G. (Ed.), *The Data Deluge: Can Libraries Cope with E-Science?* ABC-CLIO, Santa Barbara, CA, pp. 33-42.

Newton, M.P., Miller, C.C. & Bracke, M.B. (2010) "Librarian Roles in Institutional Repository Data Set Collecting: Outcomes of a Research Library Task Force" *Collection Management*, Vol. 36, No.1, pp.53-67, available at: http://dx.doi.org/10.1080/01462679.2011.530546 (accessed 4 Oct 2011).

OMB Circular A-110, available at: http://www.whitehouse.gov/omb/circulars_a110 (accessed 11 October 2011).

Palmer, C.L., Teffeau, L.C. & Newton, M.P. (2008) "Strategies for Institutional Repository Development: A Case Study of Three Evolving Initiatives" *Library Trends*, Vol. 57, No 2, pp. 142-167, DOI: 10.1353/lib.0.0033, available at: http://muse.jhu.edu/journals/library_trends/v057/57.2.palmer.html (accessed 04 October 2011).

Pryor, G. & Donnelly, M. (2009), "Skilling Up to Do Data: Whose Role, Whose Responsibility, Whose Career?" *International Journal of Digital Curation* Vol. 4, No 2, pp. 158-170, available at: http://www.ijdc.net/index.php/ijdc/article/view/126 (accessed 11 October 2011).

Renear, A.H., Sacchi, S. & Wicket, K.M., (2010), "Definitions of Dataset in the Scientific and Technical Literature", In *Proceedings of the 73rd ASIST Annual Meeting*. Available at: http://www.asis.org/asist2010/proceedings/proceedings/ASIST_AM10/submissions/240_Final_Submission.pdf (accessed 11 October 2011).

RIN (2009), "Patterns of information use and exchange: case studies of researchers in the life sciences", *Research Information Network*, available at: http://www.rin.ac.uk/our-work/using-and-accessing-information-resources/patterns-information-use-and-exchange-case-studie (accessed 11 October 2011).

Rosenbaum, S. (2010), "Content Is No Longer King: Curation Is King" *Business Insider*, June 15, 2010, available at: http://articles.businessinsider.com/2010-06-15/tech/30097151_1_content-creation-king-curation (accessed 11 October 2011).

Salo, D. (2008), "Innkeeper at the Roach Motel" *Library Trends*, Vol. 57, No 2, pp. 98-123, DOI: 10.1353/lib.0.0031, available at: http://muse.jhu.edu/journals/library_trends/v057/57.2.salo.pdf (accessed 04 Oct 2011).

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A.U., Wu, L., Read, E., Manoff, M. and Frame, M. (2011), "Data Sharing by Scientists: Practices and Perceptions" *PLoS ONE* Vol. 6, No. 6, available at: http://www.plosone.org/article/info:doi/10.1371/journal.pone.0021101 (accessed 11 October 2011).

The University of Illinois at Urbana-Champaign (UIUC), Graduate School of Library and Information Science (2011), "Master of Science: Specialization in Data Curation", available at: http://www.lis.illinois.edu/academics/programs/ms/data_curation (accessed 11 October 2011).

Witt, M. (2009), "Eliciting Faculty Requirements for Research Data Repositories" Conference Proceedings of the *4th International Conference on Open Repositories* Georgia Institute of Technology. http://hdl.handle.net/1853/28509 (accessed 11 October 2011).

Witt, M., Carlson, J., Brandt, D.S., & Cragin, M.H. (2009), "Constructing Data Curation Profiles", *International Journal of Digital Curation* Vol. 4, No. 3, pp. 93-103, available at: http://www.ijdc.net/index.php/ijdc/article/view/137, (accessed 11 October 2011).

Appendix A – Excerpts from a Data Interview

I = Interviewer
R = Researcher

I: Could you provide me with a brief overview of the data set?

R: So there's a lot of interest in understanding how certain nutrients in fertilizer interact with the environment. And so this experiment was set up to demonstrate just that, to look at some of the environmental questions. And so basically we have a field that has different micro-environments in it, and we have partitioned the field so that we can sample the field at multiple times in different environments. And it was an experiment that we conducted over two years, we have two years of data. So we have three tillage treatments, and then within that we put in some fertilizer treatments and there are three different fertilizer treatments, and then we just put these experimental treatments out across the field and then, because the field is highly variable, we can go back to specific environments and take samples. We sampled for various soil attributes as a function of depth within the rooting zone. So you go down from the surface, down to, as far down as a root will grow in a season and you partition that up into layers. And then we also collect throughout the growing season some plant samples which are indexes of how the plant is doing; the leaf tissue samples. And then we have yield. And so the data would be things like weights, or concentrations of nutrients in a sample, either soil or a plant tissue.

I: So you indicated earlier that you would only share your raw data will your immediate collaborators, but then once the data had been cleaned and processed, you would be willing to share it with other researchers at [name of institution] as well as others within and outside of your field.

R: Yes. For this type of data I feel that there's as much potential to misunderstand and as much need for description amongst any of these groups, and quite honestly, with the way technology is anybody can access me with equal ease and it might be equally annoying to me to spend time annotating data for someone who's halfway around the world and going through a translator as it is to someone who's down the hall who wants me to sit with them and say "okay this is this and this is this". So, once you get beyond the group that might be in the meeting with you, you know, quarterly, to discuss what's going on and how you're doing things, there's pretty much an equal need to have it carefully described and then you're done for all of these outside groups.

I: Which groups in particular do you think might find your data set to be particularly useful?

R: So, I would break it down into the people who actually want to use it to do research, through aggregation or something like that, or people who don't necessarily think your results go far enough and want to understand what's beneath your synthesis of data that

you wrote in a report. And that might be more industry or maybe it is policy type people who have research assistants who are delegated to aggregate and synthesize, for example EPA has its own board that will gather information and prepare it for Congress. So for this type of data, where you have kind of a linkage between yields and a huge environmental issue, this is kind of a hot topic data because it talks about potential value-added traits with high yield corn so industry might be interested in looking at, okay what did you actually find, what are potential research areas for us beyond what I've put in the paper. And they would want to look at the data set for their own purposes and, by golly I don't know why we shouldn't let them. And then likewise I might have colleagues who are doing the exact same study in different regions of the country. And better knowledge is gained form aggregating the data. Or, maybe they just want to re-analyze it completely because there might be a different result if you pull together the same type of information, yield data, you know the data on the K, the data on soils, and you dump it in and re-mine the data, you might come up with a different conclusion and since I'm not going to do those experiments elsewhere in the country.

I: So, if you get asked by EPA, or a colleague, what would you have to do in order for someone else to understand and use your data? Would you have to re-package it…

R: Oh my goodness yes.

I: You would have to pull it together and annotate it…

R: Yes, and I think key barriers are one, the annotation, and two, "you". You know, you made the statement "you would have to do it" and that's often, even if I have to delegate, if the student who did the particular analysis, [student name] is gone! You know, he's the student, he's now got a job someplace else, he's not here to do that, and that means one of the professors is going to have to do it.

I: So the annotation is something that probably could be handled by a graduate student but there's a time lag between the time you're doing the study and the time you're ready and willing to share…

R: Yeah, and the graduate student themselves may no longer be available and so then if you just ask some other graduate students, it's not their project. They don't have the same corporate knowledge of the project that is owned by these co-authors.

I: And so that resides with you…

R: [laughs] And departs with me, or whatever. Yeah, so theoretically the graduate student could do this as they develop the data set; they already document their work in lab notebooks. But my students don't really follow set procedures in writing up their lab notebooks. And it's not easy to connect the information in their paper notebooks with the Excel spreadsheets they generate. I can get at the information if I need to, but it's not really accessible to anyone else.