D. Scott Brandt, Purdue University Libraries
Jacob R. Carlson, University of Michigan Libraries

Abstract: Enhancing the Data Curation Profiles to help Bridge the Gap between Researcher and Repository

This project will scope the outcomes for a roadmap leading to the redesign of the Data Curation Profiles Toolkit (DCPT). We will do this through partnerships with organizations and individuals who possess expertise in specific areas. The project is comprised of two parts. The first is to define four key areas of development for the next iteration of the DCPT, and get the input of a broad group to develop the roadmap. This will result in a proposal to design and build the next version of the DCPT. The second is to support a dialogue of experts in the library community on bridging the gap between the "active" stages of the data lifecycle to the curation stages with an end goal of making the transition from active use to dissemination and preservation a smoother and more seamless process. We will identify the key challenges, setting an agenda for community action and ensure that the next iteration of the DCP Toolkit is well situated to make an impact in further developing data services. This will be gathered and disseminated in a "bridging the gap" report.

Our approach in redesigning the DCPT has been informed from surveys of users and attendees of the DCP Workshop series, focus groups, the DCP Symposium and a usability study. Based on our preliminary work to this point, we have identified four areas of focus:

1. Incorporate the use of personas and scenarios
2. Creating a larger and more diverse question bank targeted to different stages across the data lifecycle.
3. Developing a reporting mechanism as a product of the DCPT interview that would contain recommendations for action by the researcher.
4. Generating better alignment with data repositories to incorporate best practices in data deposit.

These areas will contribute to developing more concrete outputs and results from using the enhanced DCPT, each building upon the other: a stronger question set will provide information that will help build personas and scenarios, and these in turn will identify metadata which is needed for deposit.

This planning grant will enable us to engage with key experts in each of these four areas. For our work in developing personas and scenarios, we will be collaborating with Dr. Suzie Allard, Associate Professor, School of Information Sciences, University of Tennessee Knoxville. Dr. Angus Whyte, Senior Institutional Support Officer, DCC at University of Edinburgh, and Sarah Jones, DCC Researcher at the Humanities Advanced Technology and Information Institute, University of Glasgow will provide input into creating a new question bank. Sherry Lake, Senior Data Consultant, at the University of Virginia and one of the developers of the DMVitals project will partner with us to develop reporting capabilities for the DCPT. Finally, in addition to interacting with personnel from Purdue's own data repository, PURR, we will work with Todd Vision, the Principal Investigator at Dryad Digital Repository and Chris Taylor, eScience Support Specialist. They will help us to understand the issues related to repository deposit across a variety of journals/use cases, and determine how to extrapolate these concerns into questions and scenarios.

The outcomes of this project will include a roadmap for the next iteration of the DCPT, a proposal for an implementation grant to the IMLS to put what we have learned into action and a report of our work in defining issues in transitioning data from being managed to being curated.

NLG Planning Grant 2014: Enhancing the Data Curation Profiles to help Bridge the Gap between Researcher and Repository

Summary Statement

This planning grant is requested to create a roadmap to scope the outcomes and work for redesigning the Data Curation Profiles Toolkit (DCPT) (PUL 2009). The ultimate goal is to help facilitate and increase the curation of research outputs—moving from active management of data and digital objects to dissemination and preservation of them. Consulting with experts in the field we will determine and prioritize outcomes that will enhance the next iteration of the DCPT. Funding will enable the PIs to work on-site with a group of experts, and provide the means to hold a workshop on bridging the gap between researchers and repositories. Forming solid partnerships with organizations and individuals who possess expertise in certain areas will be a critical component of developing the next iteration of the DCPT. Our primary outcome from this planning grant will be a roadmap for developing the next iteration of the DCPT as a more extensible and powerful tool for librarians and other information professionals to connect with stakeholders seeking to deposit their data into a repository. This roadmap will serve as the foundation for a proposal for an IMLS NLG Project Demonstration Grant in 2015 to build DCPT2.0. A second outcome will be a "bridging the gap" report that pulls together the thinking of experts on needs for moving from data management planning to active deposit of data for use/reuse and preservation.

Statement of Need:

Much has been achieved in data curation by libraries in the past eight years. Since IMLS awarded grants to develop digital curation education programs in 2006, libraries have seen the rise of digital curator as a profession, an increase in digital management tools, and the advent of the institutional data repository. (Ray 2012) The NSF's data management plan "mandate" in 2010, motivated many academic libraries to rise to the challenge of consulting with researchers on creating DMPs. The ARL's e-Science Institute, initiated in 2011, helped many university libraries plan for developing data services. Academic research libraries are not alone in this, as archives and museums are seen as collaborators who are addressing similar problems in making their resources discoverable. (Duff et al, 2013) And yet there is still much to do, especially moving beyond data management and planning to the deposit and dissemination of research outputs.

The 2013 White House OSTP memorandum on "Increasing Access to the Results of Federally Funded Scientific Research," made it clear that researchers need to go much further than making plans for data, emphasizing the need for deposit, access and usability. And while many feel the growing "mandate" is aimed at scientific researchers who received NSF and NIH funding, many other players—from publishers holding millions of articles to repositories and museums sitting on a vast wealth of undiscovered digital (or soon to become digitized) objects—are driven by an understanding of the need to make contents more openly available. Planning tools such as the DMPTool and DMPOnline are crucial for thinking about data at the beginning of the research lifecycle, however more help is needed

at points further along the data lifecycle. In particular, research has identified the process of transferring data from its creators to its curators is a bottleneck in the data lifecycle (Witt 2008). The act of transferring a data set or digital object to a third party is often a new and unfamiliar experience to many researchers. As such deposit is not a part of the typical scholarly workflow, researchers often do not give much consideration towards how the transfer will take place. Critical information that needs to accompany the outputs, such as documentation and description, are often lacking which in turn reduces the utility of the data set for dissemination and reuse. As a result, the transfer, if it happens at all, is a cumbersome process that requires a heavy investment of time and resources on the part of the curating agency. As deposit mandates become more commonplace repositories will likely face an increasing demand for their services which will stretch their already limited resources even further.

The Data Curation Profiles Toolkit was created in 2010 with support from the IMLS as a resource for librarians to engage researchers in discussion about their data. Specifically, the DCPT is an interview protocol designed to capture information about a particular data set being developed or managed by a researcher across its data lifecycle, how the researcher and her lab are managing and working with the data set currently, and what the researcher would like to do with the data set but is not for whatever reason. In other words, what are the unmet needs of the researcher for her data set? The output, a Data Curation Profile (DCP), is a document that can be shared amongst the researcher, service providers and other stakeholders as a means of informing a plan of action. The DCPT has been widely adopted and used by research librarians all over the world to help them connect with student and faculty researchers and as a means to inform library initiatives and projects to assist these researchers in addressing their needs. Notable uses of DCPs include Cornell's project to reimagine the services offered through their DataSTaR repository (Wright et al. 2013) and Purdue's work to understand the needs of graduate students developing data in a field station (Carlson and Stowell-Bracke 2013). The value of completed Data Curation Profiles transcends the individual interaction between librarian and researcher. Completed DCPs serve as a community resource that can help to inform the development of data services in libraries. The Data Curation Profiles Directory, which provides access to completed DCPs, went online in November of 2013 (previously, completed DCPs were posted on the project's website) (PUL 2013). In its first year, more than 4,000 copies of DCPs have been downloaded from the collection.

Since the DCPT was introduced in 2010, we have been collecting and analyzing data about how it is being used by librarians and about how librarians are engaging with researchers more generally. Librarians and informational professionals who participated in workshops on using the DCPT professed an increase in confidence in discussing data sharing, description and intellectual property, but also noted the time and effort it took to develop a DCP as a barrier to its use (Carlson 2013). Experts, leading figures, and practitioners who participated in a symposium discussing the DCPT workshop, recognized the utility and impact of the DCPT and strongly encouraged that it be enhanced to further facilitate data curation. (Brandt & Carlson 2013) We aver that an extended and easier to use version of the DCP will do that.

We seek to apply what we have learned about how librarians have (or have not) engaged with data producers towards developing the next iteration of the DCPT. The goal of the next iteration of the DCPT will be to enable librarians to assist data producers in any setting in responding to the increasing pressure from federal, publishing and scholarly communities to make research data and digital objects more widely available through deposit into repositories. Our intent is to develop a tool that will bridge the gap between the "active" stages of the data lifecycle management to curation stages of discovery, access and preservation based on a solid understanding of current practices with data in research labs. The tool would then respond with recommendations based on established best practices. While much of our work has focused on academic research environments, we know that colleagues in state archives and museums are also interested in a "DCPT2.0" that could work in their settings as well.

Without this Planning Grant, we feel we are in jeopardy of working on the next version of the DCPT without much needed additional perspectives and context, or not being able to work on the project at all. We have been successful in identifying ideas and future directions for development of the DCPT, but recognize we need the input of other experts to ensure that our roadmap is solid and complete. Furthermore without a formal commitment of our time on a grant, we have had to take this on in small spurts of effort here and there. In the Purdue Libraries, a grant affords the formalization of "buying out" our time in other work (e.g., not co-teaching for a semester, having a colleague temporarily take liaison duties or committee work, etc.). Thus, this planning grant is needed for us to be able to dedicate a significant amount of our time to move forward in advancing the DCPT, and in forging the needed connections with relevant experts.

This planning project will enable us to scope the outcomes for a roadmap leading to the redesign of the DCPT. The project addresses two overlapping components, and provides two tangible outcomes, a roadmap and a white paper.

The first component is to dig deeper into four key areas that our work has shown will be critical to address. This will guide development for the next iteration of the DCPT so these areas are addressed and a solution for them implemented. The second component is to facilitate a dialogue of experts in the library community on issues and needs in bridging the gap between the "active" stages of the data lifecycle to those of the curation lifecycle. The end goal is to make the transition from active use to dissemination and preservation a smoother and more seamless process. The results of addressing these components will provide a roadmap in the form of a proposal to move forward. By analyzing these components we will identify the key challenges, setting an agenda to ensure that the next iteration of the DCPT makes an impact in further developing data services. And because key problems relating to research data will be identified, reviewed and analyzed by experts as a part of this process, we want to share that with the broader research community irrespective of the DCPT.

Impact:

The initial research, feedback and input noted above helped us identify many of the "on-the-ground" problems that practitioners face engaging in data curation. Many of the issues are not new: researchers feel they don't have the time or knowledge to address curation, and librarians feel they need more techniques and tool to help researchers prepare to share or deposit data. Projects like the SHared Access Research Ecosystem (SHARE), of which Purdue is a contributing member, will provide frameworks and mechanisms for sharing data, but does not explicitly address what researchers need to do (ARL 2013). While the current version of the DCPT does not provide specific recommendations for curating data, it does walk through the complete research lifecycle and helps identify needs to be addressed.

This aforementioned work has provided many ideas on what the next version of DCPT could be, and how it could address moving data from a management setting to a repository and preservation environment. This planning grant will allow us to collect and organize these and other ideas, prioritize them, and refine them into a plan a roadmap to proceed with developing "DCPT2.0." We will bring together the collective knowledge of leading figures, experts and practitioners through site visits, focus groups and interviews.

In particular, we have identified, and seek to explore, four critical areas for the further development of the DCPT. These four areas are:

1. Incorporate the use of both personas and scenarios that would provide insight into how researchers work and enable the librarian to ask questions that would better reflect the specific research being performed and data being generated.
2. Create a larger, more specific and more diverse question bank that would address curation issues and needs at various stages along the data or research lifecycle.
3. Develop a means to rank or prioritize researcher needs within the structure of the DCPT and then, through applying existing best practices and standards, to craft a report that includes actionable recommendations to the participating researcher.
4. Generate a greater continuity and alignment with the needs of repositories seeking data deposits through identifying what information and materials are needed, as well as desired, from researchers seeking to make deposits. Defining what would constitute an "ideal deposit" will inform best practice and shape the content and context of DCPT 2.0.

This planning grant will enable us to engage with these key experts in each of the four areas. This engagement will form the initial part of our work in charting the direction for the next DCPT. Once we have addressed these key areas, we will work with others in workshop environment to ensure we have the broad perspective to make the DCPT useful for many environments.

There will be two direct outcomes from this project. The first outcome is a roadmap for the next iteration of the DCPT. This roadmap will serve as the foundation for a demonstration grant proposal to the IMLS in 2015. The second outcome will be a report of our work to

define the challenges of bridging the gap between data producers and data curators and to identify action areas for addressing these challenges. The report will be independent from the DCPT roadmap and will be released as a separate document on Purdue University's institutional repository. The impact of this planning grant will depend on two criteria. First, by our success in obtain funding based on the roadmap created in this project to pursue the implementation of the next iteration of the DCPT. Second, by the impact our "bridging the gap" report has on various data curation and scholarly communities. Metrics will include the number of downloads, the strength and duration of the social media activity, and the number of citations or mentions in the literature.

Project Design:

The goals of the project are to lay the groundwork for constructing the next iteration of the DCPT, to form solid partnerships with organizations and individuals who possess the expertise that we will need to be successful in rethinking the DCPT, and to articulate issues clearly about depositing data into a repository and to chart courses of action to help address these issues.

The project plan is to gather and use expert input to develop a roadmap for redesigning the DCPT. Currently we have several sources of input and feedback on the current tool. We have evaluations from 119 workshop participants from immediately after the workshop as well as three months later. We have feedback from 12 practitioners who created a Profile and participated in a focus group as a part of the DCPT Symposium. We also have input from a dozen experts and leaders in data curation who participated in the symposium. And we have recent data collected from a survey of 150 people who downloaded the Toolkit. While we believe this information and our experience working in this area, especially developing and teaching others to use the DCPT, gives us deep insight into issues and problems in data curation, we know that we need to get more input to build a more useful tool. The roadmap will help define the goals, outcomes and objectives needed for building "DCPT2.0." This includes accounting for system requirements for technical development, implementation and testing of the new tool.

In order to accomplish this we first will build collaborations with partners with expertise in each of our four areas of focus. Then these partners to help lead a workshop which will include a larger group of experts to ensure we have affirmation for the direction in which we'll be moving, and are inclusive of their viewpoints (e.g., from DMPTool, JHU's Conversancy, The Smithsonian, Cornell's DataSTaR, etc.).

For our work in developing personas and scenarios, we will be collaborating with Dr. Suzie Allard, Associate Professor, School of Information Sciences, University of Tennessee Knoxville. In addition to her position at UT, Dr. Allard is a co-investigator and co-leader of the Sociocultural Issues Working Group of DataONE. She has led work of that WG to develop personas that describes different kinds of people who would interact with a particular system, and scenarios that describe different use cases of how people interact with systems. We will work with Allard to design a conceptual structure for using responses to DCPT questions, current and future, to develop personas and scenarios.

Through our collaboration with Dr. Allard, we will reach out to important communities including DataONE and LIS students to provide additional perspectives. Because Dr. Allard sits at the nexus of so many overlapping areas and serves in numerous roles, we are asking that she be a formal partner in this planning grant.

In considering the development of a question bank across the data lifecycle, the Digital Curation Centre (DCC) has crafted instruments for interviewing data producers, as well as organizations and institutions: the Digital Asset Framework (DAF), Collaborative Assessment of Research Data Infrastructure and Objectives (CARDIO) and DMPOnline are well known examples. Dr. Angus Whyte, Senior Institutional Support Officer, DCC at University of Edinburgh, and Sarah Jones, DCC Researcher at the Humanities Advanced Technology and Information Institute, University of Glasgow are currently working on expanding and enhancing a knowledge bank of questions for DCC instruments. We will work with Whyte and Jones to develop a framework for building questions based on themes related to data management and curation.

Providing actionable information to researchers based on their environment and their needs will be an important element of the next DCPT. DMVitals is a tool used to assess data management, which provides feedback and recommendations for practices from similar scenarios, best practices and standards. Sherry Lake is a Senior Data Consultant in the Data Management Consulting Group at the University of Virginia. She oversees the development and use of DMVitals. We will work with Lake to understand how responses to questions can be weighted or ranked to provide customized feedback and best case recommendations, and linked to personas and scenarios for reinforcement.

We see open dialogues with data repositories on their needs and expectations for data deposit as a critical step in addressing the gap between researchers and repositories. Dryad is a curated disciplinary repository for data underlying the international scientific and medical literature. Dryad's approach to integrating data submission for a growing list of journals allows for a variety of data curation use cases—from direct submission by authors to facilitated submission by journals. Todd Vision, Principal Investigator at Dryad Digital Repository and Associate Director of Informatics at NESCent, along with Chris Taylor, eScience Support Specialist will be working with publishers and editors to create instructions and coaching to enhance data deposit. In addition to working with personnel from the Purdue University Research Repository, we will work with Taylor to understand the issues related to repository deposit across a variety of journals/use cases, and determine how to extrapolate these concerns into questions and scenarios.

All of these will contribute to developing more concrete outputs and results from using the enhanced DCPT, each building upon the other: a stronger question set will provide information that will help build personas and scenarios, and these in turn will identify metadata which is needed for deposit. And these will help drive the workshop to validate our work to date and include broader perspective going forward.

The project will consist of the following seven activities:

- **Activity 1** – Preparation: (1 month, October) We will collate, analyze and categorize the information we have collected thus far about the DCP, its use by librarians and librarian's engagement with data producers to build a framework for interacting with our expert collaborators. This framework will aid us in determining the specific objectives and outcomes for each of our four areas of inquiry. We will supplement this activity with a literature search to include most recent perspectives in library and information science. Based on this work we will form strategies to develop each of our four areas of inquiry.

- **Activity 2 –** Gathering expert input: (2 months, November – December) We will gather expert opinion and knowledge on the key issue (moving from data planning and management to curation and preservation) as relates to vision and goals of the DCPT. We will travel to our collaborators' institutions to work with them on reviewing our initial preparation and research, and gathering their feedback and input. The primary experts with whom we will collaborate work at institutions which are well known for innovations in many areas related to data curation—research, application, and education. Thus, we will identify others at these locations who could provide further input, guidance and perspective on this project. We will meet with these individuals to conduct interviews or focus groups with them.

- **Activity 3 –** Workshop: (1 month, January) We will hold a workshop at Purdue University on issues and strategies in bridging the gap between data producers and data curators focusing on the deposit of data into repositories. The workshop will be comprised of a limited number of selected invitees, including our expert collaborators, with sufficient expertise to be able to help address the topic. The workshop will primarily be a working session and focused on addressing issues and how to resolve them.

- **Activity 4 –** Submit Implementation Proposal: (2 months, January - February) Concurrent with Activities 2 and 3, we will prepare and submit a proposal for a demonstration grant to the IMLS. The purpose of the demonstration grant will be to provide funding to design and implement the roadmap we develop for the next iteration of the DCPT.

- **Activity 5 –** Draft Report: (3 months, February – April) Applying what was learned from the workshop held in Activity 3, we will draft a report that defines the issues in transitioning data from being managed to being curated, and plots courses of action for the data curation community to address these issues. We will continue our discussions with our expert collaborators and incorporate their guidance and feedback into this report.

- **Activity 6 –** Peer Review and Revision of Report: (3 months, May – July) Together with our collaborators, we will identify additional experts and leaders in data curation, as well as LIS educators, digital archivists and researchers, who can provide further input, guidance and direction on the subject of connecting the data

management and data curation aspects of the data lifecycle. We will share a draft of the report with them and solicit their reviews and suggestions.

The report will be published by the end of July 2015 in Purdue's institutional repository.

- **Activity 7** – Disseminate and Publicize Report: (2 months, August – September) We will publicize and promote the report and lead discussions on the issues it raises. This will include holding a webinar and engagement through social media.

Personnel, Time, Budget:

As investigators of two previous Data Curation Profiles grants, Scott Brandt and Jake Carlson of Purdue will lead this NLG Planning Grant. Dr. Suzie Allard of Tennessee will be a formal partner and co-PI. As a leading figure in research data planning and management, a member of multiple DataONE working groups, and an LIS educator, she posses a great body of knowledge, experience and expertise directly related to this area. She has interacted with the Purdue team previously, having consulted at the DCP Symposium held in 2012. Both Scott Brandt and Jake Carlson will devote 10% of their time to the project which will be cost shared by Purdue. Dr. Allard will devote 5% of her time to the project.

As noted, we seek specific expertise and knowledge to address four key areas in building a roadmap. We have identified experts with whom we will collaborate in these areas. The first is Suzie Allard, for her expertise in understanding and developing personas and scenarios that assist users in understanding context for engaging in data curation, and provide examples that may help guide them. These will be extracted from responses to questions, and we recognize that we need to build a larger and more extensive question bank to guide users through data curation decision points.

Second, we will work with Dr. Angus Whyte of the DCC and Sarah Jones of HATII to develop a question bank appropriate to the goals and structure of DCPT 2.0. Dr. Whyte and Ms. Jones have been working on a similar approach to building question banks for their work on the DAF, CARDIO and other projects.

Third, to create mechanisms for feedback and recommendations, we recognize the need to build into the system some kind of response ranking or weighting to provide context relevant feedback. The DMVitals system is a prototype for doing this, and we will work with Sherry Lake, one of the architects of DMVitals, to determine how we can best incorporate user generated response in the new tool. Ms. Lake is the Senior Data Consultant at the University of Virginia and has a wealth of experience in developing solutions to researcher's data management and curation needs.

Fourth, as we want to ensure that curation matches the needs of repositories, we will work with Dr. Todd Vision and his colleagues at the Dryad data repository. Dryad works closely with many researchers, authors, editors and publishers, and is the prototype for proactive data curation. Dr. Vision is the Associate Director of Informatics at the National

Evolutionary Synthesis Center at University of North Carolina at Chapel Hill and has a wealth of knowledge in community focused informatics.

Dr. Whyte, Ms. Jones, Ms. Lake and Dr. Vision will each contribute a total of five days of their time to the project. This includes time that will be spent with Scott Brandt and Jake Carlson during visits, travel to Purdue to participate in the workshop and communications throughout the life of the project. We will also take advantage of our close proximity to work with Michael Witt, project director for the Purdue University Research Repository, one of the first university sponsored institutional data repositories in the US.

References:

Association of Research Libraries. (2013). SHared Access Research Ecosystem. http://www.arl.org/focus-areas/public-access-policies/shared-access-research-ecosystem-share

Brandt, D. S. & Carlson, J. (2013) "Final Performance Report: Understanding Curation through the use of Data Curation Profiles" (IMLS RG-06-10-0101-10). Unpublished.

Carlson, J. R. (2013). Opportunities and Barriers for Librarians in Exploring Data: Observations from the Data Curation Profile Workshops. *Journal of eScience Librarianship*, 2(2), 2.

Carlson, J., & Stowell-Bracke, M. (2013). Data management and sharing from the perspective of graduate students: An examination of culture and practice at the Water Quality Field Station. portal: Libraries and the Academy, 13(4). p.343-361. doi: 10.1353/pla.2013.0034

Data Curation Profiles Template (2010). Data Curation Profiles Toolkit. http://datacurationprofiles.org/sites/all/docs/DCP_Template_v1.doc

Duff, W. M., Carter, J., Cherry, J. M., MacNeil, H., & Howarth, L. C. (2013). From coexistence to convergence: studying partnerships and collaboration among, libraries, archives and museums. *Information Research*, 18(3).

Purdue University Libraries (2009). Data Curation Profiles Toolkit. http://datacurationprofiles.org

Purdue University Libraries (2013). Data Curation Profiles Directory. http://docs.lib.purdue.edu/dcp/

Ray, J. (2012). The rise of digital curation and cyberinfrastructure: From experimentation to implementation and maybe integration. *Library Hi Tech*, 30(4), 604-622.

Witt, M. (2008) Institutional repositories and research data curation in a distributed environment. *Library Trends*, 57(2), 191-201.

Wright, S. J., Kozlowski, W. A., Dietrich, D., Khan, H. J., Steinhart, G. S., & McIntosh, L. (2013). Using Data Curation Profiles to Design the DataSTaR Dataset Registry. *D-Lib Magazine*, 19(7), 2.

Schedule of Completion: Enhancing the Data Curation Profiles to help Bridge the Gap between Researcher and Repository

Activity	Oct 13	Nov 13	Dec 13	Jan 14	Feb 14	Mar 14	Apr 14	May 14	Jun 14	Jul 14	Aug 14	Sep 14
1. Preparation	■											
2. Gathering Input		■	■									
3. Workshop				■								
4. Submit Proposal				■	■							
5. Draft Report					■	■	■					
6. Peer Review & Revision								■	■	■		
7. Publish and Publicize											■	■