The International Journal of Digital Curation
Issue 3, Volume 4 | December 2009

# Constructing Data Curation Profiles

Michael Witt, Jacob Carlson, D. Scott Brandt,
Distributed Data Curation Center, Purdue University
Melissa H. Cragin,
Graduate School of Library and Information Science,
University of Illinois at Urbana-Champaign

## Abstract

This paper presents a brief literature review and then introduces the methods, design, and construction of the *Data Curation Profile*, an instrument that can be used to provide detailed information on particular data forms that might be curated by an academic library. These data forms are presented in the context of the related sub-disciplinary research area, and they provide the flow of the research process from which these data are generated. The profiles also represent the needs for data curation from the perspective of the data producers, using their own language. As such, they support the exploration of data curation across different research domains in real and practical terms. With the sponsorship of the Institute of Museum and Library Services, investigators from Purdue University and the University of Illinois interviewed 19 faculty subjects to identify needs for discovery, access, preservation, and reuse of their research data. For each subject, a profile was constructed that includes information about his or her general research, data forms and stages, value of data, data ingest, intellectual property, organization and description of data, tools, interoperability, impact and prestige, data management, and preservation. Each profile also presents a specific dataset supplied by the subject to serve as a concrete example. The *Data Curation Profiles* are being published to a public wiki for questions and discussion, and a blank template will be disseminated with guidelines for others to create and share their own profiles. This study was conducted primarily from the viewpoint of librarians interacting with faculty researchers; however, it is expected that these findings will complement a wide variety of data curation research and practice outside of librarianship and the university environment.[1]

[1] This paper is based on the paper given by the authors at the 5th International Digital Curation Conference, December 2009; received October 2009, published December 2009.
The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. ISSN: 1746-8256 The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre.

# Data Curation Profiles

***Background***

Despite the recent increase of interest in data curation, defined in general terms by Lord, Macdonald, Lyon & Giaretta as "managing and promoting the use of data" (2004), very few tools exist regarding its learning and practice. This is especially the case for librarians, who are beginning to initiate or have been otherwise challenged to adapt to rapid changes in scholarly communication that include stewardship of research datasets (Association of Research Libraries [ARL], 2006). Responding to emergent changes in scholarly communication and recent reports on Cyberinfrastructure and e-Science, there has been a number of university-based initiatives to address both local and field-wide knowledge gaps on research practices and related data management problems. Several of these initiatives were led by university libraries and involved local environmental scans of the research activities, data being generated, practices and barriers, and other factors.

A team from the University of California-Santa Barbara Library published a report on their local informatics efforts that focused on data-intensive, interdisciplinary research (Pritchard, Anand & Carver, 2005). A significant contribution of this work was documentation of certain data-generation characteristics and the relationship to informants' sharing practices. The authors found that higher levels of automation in data generation or processing were often indicators of increased willingness to share data during the research cycle. Additional research is needed to identify similarities and distinctions across methods, research areas and sub-disciplines, but this study offers a view of some of the complexities that have a bearing on sharing and data publishing activities.

Librarians from the University of Minnesota published a report from a study on the research behaviour and related information service needs of their scientists and graduate students (Marcus et al, 2007) . The inclusion of graduate students in this study makes an important contribution to the knowledge base on research-related practices, as it identified some of the differences in social aspects and information needs between these two groups.

A multi-university study conducted by the Australian Partnership for Sustainable Repositories produced a report on a survey covering university data management practices (Henty, Weaver, Bradbury & Porter, 2008). Findings included great similarities in question responses across the three participating universities. Disciplinary difference was not an explicit goal of the survey, as respondents were not asked to identify their field or research area; some broad categorization was determined based on "extrapolation from departmental… or organizational affiliation", and fields included social science, medicine and health, business and economics, information technology, engineering and architecture, humanities and creative arts, science, and law. Interesting findings are included in the sections on "types of digital data" and the section on software applications used to generate digital data, as these provide a starting point for other groups undertaking local inventories.

Outside of libraries, several recent projects have been conducted investigating research domains and data practices in relation to repositories, including Project StORe (Pryor, 2007) and DataShare (Rice, 2007). Project StORe was developed to "increase the value of research output by implementing bi-directional links between published papers and reports and the datasets behind them". The investigation was conducted across seven research fields, including astronomy, biochemistry, biosciences, chemistry, physics, social policy and political science. Pryor (2007) reported findings from the large survey and follow-up interviews conducted to clarify the use and non-use of source (research data) repositories, as well as researchers' needs for usable and useful systems. In addition to identifying some of the differences between research groups (i.e., faculty and graduate students) and the disciplines, the study found that even for fields where data submission was expected with publication, only a very small portion of those researchers' data was ever deposited. The UK DataShare Project was initiated to explore novel approaches to support academic researchers who want to share data over the Internet (Rice, 2007). The project was conducted at three universities, which implemented the study to match current, local repository efforts. The lead group on this project also worked on the development and piloting of the Data Audit Framework (Jones, Ball & Ekmekcioglu, 2008) which was developed to assist "organisations with the means to identify, locate, describe and assess how they are managing their research data assets".

The most comprehensive study to date about researchers' views and data sharing activities was undertaken by the Research Information Network (RIN) report (2008), which conducted over 100 interviews with researchers from eight fields, including astronomy, chemical crystallography, classics, climate science, genomics, rural economy and land use, social and public health, and systems biology. The investigation addressed three areas: how data are shared or made available to others, current roles of primary research data in scientific production and communication, and quality assurance practices. Significant to this investigation, the RIN (2008) report identifies several gaps in the curation knowledge base, recommending the need to take "full account(s) of the different kinds of data that researchers create and collect… and make clear the categories of data that they wish to see…shared with others".

## Introduction

The investigation described in this paper begins to address the gap indicated by the RIN report and builds on the efforts identified above by profiling researchers and their data in order to inform data curation activities in academic libraries. Investigators from Purdue University and the University of Illinois conducted a series of in-depth interviews with a convenience sample of 19 faculty members at their respective institutions to identify needs related to data curation in each of their domains and then profile them in concise, structured documents - Data Curation Profiles [2] - that are suitable for sharing and annotation. For each faculty subject, a profile was constructed that includes information about his or her general research, data forms and stages, value of data, data ingest, intellectual property, organization and description of data, tools, interoperability, impact and prestige, data management, and preservation. Each profile also contains detailed information about a specific dataset supplied by the subject as a real-world exemplar. In total, 12 research domains are being explored (the

[2] The completed Data Curation Profiles and more project information can be found at http://datacurationprofiles.org/

number of profiles is in parentheses): Agronomy & Soil Science (5); Anthropology (3); Biochemistry (1); Biology (1); Civil Engineering (1); Earth and Atmospheric Sciences (2); Electrical and Computer Engineering (1); Food Science (1); Geology (3); Horticulture and Plant Science (2); Kinesiology (1); Speech and Hearing (1). In addition to the profiles, investigators are examining two of the domains as case studies[3], conducting focus groups with participating subject-specialist librarians, and investigating practical applications for institutional repositories.

Developing and completing the Data Curation Profiles served as a vehicle for the investigators to interact directly with data producers, understand their perceptions and scientific workflows, and determine what information to collect about their data needs that are pertinent for curation. As an output of the study, the Data Curation Profiles can be used by librarians and others to inform decisions such as the selection and deselection of datasets, the presentation of data for human and machine consumption, and the provision of metadata (see Figure 2). The profiles can also facilitate the determination of new roles in archival and systems librarianship as the needs expressed by the faculty subjects can be associated with systems and services that can be provided by libraries and librarians. The Data Curation Profiles are being published to a public wiki for questions and comments, and a blank template will be disseminated with instructions for others to create and share their own Data Curation Profiles. In this way, the profiles can be referenced and enhanced by practicing librarians, and the wiki can become a on-going resource for the applied learning and professional development of librarians who will play a role in data curation.

## Developing the Data Curation Profile

One purpose for the creation of the Data Curation Profile is to address a perceived shortage of robust models for the systematic description of datasets for sharing and curation. Creating the profile required two elements: 1) the conceptual development of the function and content and 2) the generation of a template. Three initial prototypes were generated using literature-based cases of data-handling and curation efforts in three exemplary fields. Researchers in astronomy, ecology and crystallography have made significant advances in developing standards and infrastructure for managing, sharing and curating data. The methodology for this process was essentially a review and distillation of sets of published literature and project-based documentation pertaining to research data and their dissemination, management, description, use, or other curation-related issues.

A review of astronomy focused primarily on the work that had been done by the US National Virtual Observatory and the Sloan Digital Sky Survey. In ecology and environmental sciences, materials from the Long Term Ecological Repository (LTER) and the Center for Embedded Sensing (CENS) projects were reviewed. In the biological sub-discipline of crystallography, work done by the eBank UK project and the eCrystals repository was examined.

Passages from this literature that described the lifecycle of the data, as well as the passages that discussed or addressed community needs or functionalities of the project relating to data, were identified and excerpted into a separate document. The lifecycle for datasets in each field was then reconstructed and annotated. Categories were

[3] Similar case studies are being conducted by the Digital Curation Centre in its SCARP Project http://www.dcc.ac.uk/scarp/

deduced by analyzing the identified needs and functionalities within each field. Once this categorization had been done in each of the fields, a "card sort" exercise was performed to sift out the common categories of need across all fields as well as "depth" of the categories (how many of the needs identified fell into a particular category). This literature-mining process was effective in identifying issues common across these fields and therefore informing the development and structure of the Data Curation Profile template.

While working with the literature-based profiles, a qualitative methods protocol was developed and approved by the Institutional Review Boards of both institutions authorizing research with human subjects. Qualitative data (along with a few quantitative variables) were collected through two stages of interviews, interview "worksheets", sample datasets and documents. This blended data approach helped to bring into focus each scientist's specific data types and their related curation needs. Preliminary analysis revealed that the initial interviews about the participants' research and data forms were not sufficient to elicit the granularity of requirements details needed to consider related curation policies. The complementary nature of interviews with the integrated structured worksheets were of particular value in situating the participants' data management needs in the context of their research cycles.

Interviews were conducted using an Interview Guide to help focus attention on data issues. In the first stage, a Pre-interview Worksheet was distributed prior to the interview asking the subject to identify their research area and to describe two recent or on-going projects "from the perspective of the data." Questions in the Interview Guide tended to be general (e.g., "How long do you usually keep your data?"), which allowed the subjects the freedom to speak freely from their perspective and understanding of data curation. Interviews ranged from 60-120 minutes. A second stage of follow-up interviews included a Requirements Worksheet, designed to gather more granular information (e.g., "How many years should this specific dataset be preserved?") about curation needs and requirements for the specific forms of data subjects had stated they were willing to share. This was supplemented with customized follow-up questions to fill gaps from the initial interviews. While the Purdue project investigators worked with subjects who were already known to them, the investigators at Illinois enlisted the help of their subject-specialist librarians to identify and recruit subjects. Preparation for the interviews required learning about the subjects from public material available on the Web and information provided by the subject-specialist librarians.

All interviews were recorded and fully transcribed. The initial code list was developed through independent manual coding of selected interviews by multiple investigators. The investigators worked together to ensure the selection of all broadly relevant and useful terms, a shared understanding of the terms and their codification, and the optimization of intercoder reliability. Transcripts were then coded using qualitative analysis software (NVivo), applying the initial coding terms followed by iterative micro-analysis of data related to strong emergent themes. Results from the data generated with the Requirements Worksheet were analyzed to identify patterns and contrasts regarding the data forms that the subjects were willing to share, when they were willing to share them (e.g., before or after the publication of a paper), followed by further analysis of the interview data to draw out associated motivations and rationales (Witt, 2009).

At the start of the interview process, most of the subjects who participated in the study expressed interest in sharing at least some of their data with others beyond their own research teams at some point in the data's lifecycle. Several of the subjects had already shared data with other researchers informally, through e-mail or by mailing a CD or hard drive of their data. However, very few of the subjects had invested a great deal of time, effort or resources on curating their data or ensuring its fitness for dissemination or use by others. During the interviews, many subjects confessed ignorance on how they could or should document and manage their data to enable its dissemination and curation. Lacking experience with and knowledge of curation practices, it was clear from the transcripts of the initial interviews that many subjects were not able to provide the level of detail that would be needed to develop policies in a language that could be expressed for machine implementation.

A key aspect of the Data Curation Profile was to represent the curation needs of the subject for his or her data as articulated by the subjects themselves. Therefore, the Data Curation Profile had to be flexible enough to accommodate subjects' different needs, yet structured to enable cross-discipline analysis and consistency. Once the first interviews of the faculty subjects were transcribed, sample profiles were generated and the template was amended. Each of the four investigators took slightly different approaches to constructing their profiles, but each draft profile was centered on a ground-up approach: reviewing the content of the transcripts and using this analysis to inform the design of the structure and content of the Data Curation Profile. Once completed, the four drafts were compared and reviewed in an iterative fashion by the project team to develop a uniform set of categories and a structure that could accommodate the divergent nature of the data and capture the needs of subjects across multiple domains. This review included subject-specialist librarians from Purdue and Illinois who offered their perspectives on the utility of the profile and the potential usefulness of its content to their practice of librarianship.

The last phase of the development of the Data Curation Profile was to seek feedback and validation from a panel of external reviewers on the usefulness of the profiles. The members of this external review panel were recruited from practicing science librarians, librarians actively involved in digital preservation, a computer scientist, and a technologist from the CIO's office of an American, research-extensive university. Reviewers were provided with two draft Data Curation Profiles and were asked to evaluate the utility of the profiles for their work. Specifically, reviewers were asked if the information contained in the profiles would be sufficient for their institutions to be able to take on the responsibility of curating the dataset described in each of the profiles. The general response was affirmative. Several of the reviewers desired more detailed information than was presented in the draft profiles. In some cases the investigators revisited the transcripts to "backfill" this information; for some profiles this information was not available. This and other feedback from the external reviewers was examined, and their feedback was incorporated into the final version of the Data Curation Profile template.

## Structure and Content of the Data Curation Profile

This section provides an illustration of the Data Curation Profile. Each begins with a brief summary of the research area of interest, and a general statement on the subject's needs relating to his or her data. This summary is designed to highlight the key aspects of the data and the elements or aspects of data curation that are of primary

importance to the subject. After the summary, the Data Curation Profile is comprised of two broad sections, each of which is made up of more specific content categories.

### *Details of the Example Dataset*

The first section is designed to capture details about the example dataset, including its data forms, lifecycle stages, and other contextual information that will be needed by the data curator to understand the data and handle it effectively.

> The [subject] studies real-time traffic signal performance measures project in which he measures the movement of traffic, specifically the number of vehicles passing through an intersection and the amount of time they spend at an intersection on a movement-by-movement basis over a 24 hour period. The result is a profile of traffic movement for an intersection...

Figure 1. Excerpt "Research Area Focus" from Data Curation Profile (Civil Engineering).

#### *Overview of the research.*

This overview provides a high-level summary of the research to give the curator contextual background about the data and its use by the subject.

#### *Data forms and stages.*

Data Forms and Stages include a narrative as well as a table that describes the data at each stage of its lifecycle. Different datasets involve different lifecycles, but many data can be generally mapped to four, base stages: raw, processed, analyzed and published. The information captured about the data at each stage consists of the output, typical file size, format, and any additional notes that would assist the curator.

| Data Stage | Output | Typical File Size | Format | Other / Notes |
|---|---|---|---|---|
| "Raw" | Sensor data | 100k in 1 file per day | proprietary to the sensor | ftp downloads are mostly automated. |
| "Processing Stage 1" | Sensor data – normalized, screened for outliers/errors, and moved to an open/accessible format | Roughly 6kb | .csv / .xls | Data are formatted into .csv before bring reformatted into a mySQL database. |
| "Processed" | Data vectors | 800 records per intersection per day. Each record has about 38 fields (floating point) | SQL | MySQL database typically holds 3-4 months worth of vehicle signature data, traffic signal data and the corresponding images. |
| "Analyzed" | Pivot charts/graphs | unknown | .xls | Data needs to be placed into charts and graphs for interpretation. Visualization is necessary to give it meaning and for presentation. |

| | | | | |
|---|---|---|---|---|
| "Published" | Pivot charts/graphs | unknown | .ppt | Data are presented to others (incl. funders) via power point. |
| **Ancillary and Augmented Data** | | | | |
| Video | unknown | unknown | Several formats – primarily "Real Video" but also .wmv, .mpeg | Video taken are correlated with the data for verification purposes. |
| Image | Stills taken from the video | unknown | .gif /.jpg / .ppt | Images are generated as still shots from the video. |
| 3rd Party Data - Weather information from "Weather Underground" website | unknown | unknown | .csv files | Collected via screen scrape. Correlated with collected data for explanatory/ descriptive purposes. |
| 3rd party data – road conditions from INDOT's databases | unknown | unknown | unknown | Collected on an ad hoc basis as needed for explanatory/descriptive purposes. |

Table 1. Excerpt "Data Forms and Stages" from Data Curation Profile (Civil Engineering).

***Value of the data.***

Value of the Data captures the subject's thoughts and opinions on the value of his or her datasets outside of its immediate purpose and how it might be used or re-purposed by different audiences.

***Data for Ingest.***

Data for Ingest identifies the particular slice or stage in the lifecycle of the data that the subject has identified as having the most value for scholarship that should be curated.

***Other Descriptions***

The second section of the Data Curation Profile identifies and describes both the subject's current practices in managing, disseminating and archiving their data as well as the subject's articulated needs and functional requirements in working with and curating their datasets.

> …[Locally developed] metadata is stored alongside of the data in the mySQL database. Metadata tables within the database are: sensors, sets, devices, lanes, assets, and state codes…

Figure 2. Excerpt "Locally Developed Metadata" from Data Curation Profile (Civil Engineering

These practices and needs are organized according to several over-arching categories, defined in the profile as:

- ***Intellectual Property*** – details who the owner(s) of the data are, who the stakeholders might be, what terms of use might be needed, and if any privacy or confidentiality issues exist with the data.
- ***Organization and Description of the Data for Ingest*** – describes how the data are currently organized and described, including any associated metadata formats and standards as well as how the data may need to be described for sharing and use.
- ***Ingest*** – provides information on how the data may be ingested into a repository, including the process issues and scale.
- ***Access*** – covers the subject's overall willingness, motivations and conditions to share data as well as any stated needs or requirements for limiting user access (e.g., embargo).
- ***Discovery*** – describes what metadata and points of access may be needed for searching and browsing within a data repository as well as helping users and user agents find the data from outside of the repository (e.g., search engines, other service providers).
- ***Tools*** – any software or other tools that may be needed to use the data or enhance its utility such as visualization, data mining, or analysis tools.
- Interoperability – from the perspective of the subject, how these data may need to interact and integrate with other, external data or tools.
- ***Measuring Impact*** – attribution and prestige; getting adequate credit for contributing data to scholarship and tracking the provenance and future applications of the data by others.
- ***Data Management*** – identifies and addresses a broad range of issues relating to the maintenance of the data while under the care of the curator (e.g., audits, backups, redundancy).
- ***Preservation*** – describes the archival practices and issues related to preserving the data (e.g., policy, format migration, persistence).

## Conclusion

The first profiles were published to the project wiki in October 2009. More profiles will be published as they are completed, along with a blank template and instructions to help librarians create and share their own profiles. Feedback has been enabled (but is moderated) to allow librarians to add information to the profiles as well as ask and answer questions about them. The creation of new profiles and the dialog surrounding existing profiles will increase the breadth and depth of domain-specific knowledge in terms that are practical for librarians.

The value of the Data Curation Profile is dependent on its uses, of which several have been suggested. The initial investigation focused on building a prototype profile that was based on user perspective and perceptions that could contribute to curation; issues such as scalability and resource allocation for their production and use have not been addressed. Based on preliminary feedback from the project's subject-specialist librarians and external reviewers, it is evident that the profiles will be useful for those engaged in both upstream and downstream data management and curation services. The profiles can be useful guides for exploring, learning about and interacting with data producers and collecting information about datasets and collections. It is believed

this supports new roles for academic and research librarians, especially for liaison activities such as exploring researcher interests related to sharing data further "upstream" in the research cycle. As efforts around the development of data collections grow, tools like the Data Curation Profile can be used to help gather information to make local data development policies and selection and deselection decisions. It is proposed that profiles can be used to support professional development or applied learning for librarians who can view and share profiles to learn more about curation in a particular domain.

While this work is presented from the perspective of academic librarians, it is hoped that Data Curation Profile will complement a wide variety of data curation research and practice outside of librarianship in the university environment and that the wiki will serve as an on-going resource for the broader research community.

## Acknowledgements

This research was supported by a National Leadership Grant from the Institute of Museum and Library Services, LG-06-07-0032-07. The investigators express their gratitude to the six external advisors who reviewed the draft Data Curation Profiles: Leslie Delserone (University of Minnesota Libraries), Michael Grady (Office of the CIO, University of Illinois), Ron Jantz (Rutgers University Libraries), Ardys Kozbial (University of California-San Diego Libraries), Reagan Moore (School of Information and Library Science, University of North Carolina) and Brian Westra (University of Oregon Libraries).

## References

Association of Research Libraries. (2006). *To stand the test of time: Long-term stewardship of digital data sets in science and engineering*. Retrieved June 30, 2009, from http://www.arl.org/bm~doc/digdatarpt.pdf

Henty, M., Weaver, B., Bradbury, S., & Porter, S. (2008). *Investigating data management practices in Australian universities*. Retrieved July 7, 2009, from http://hdl.handle.net/1885/47627

Jones, S., Ball, A., & Ekmekcioglu, Ç. (2008). The Data Audit Framework: A first step in the data management challenge. *International Journal of Digital Curation, 3*(2), 112-120.

Lord, P., Macdonald, A., Lyon, L., & Giaretta, D. (2004). From data deluge to data curation. *Proceedings of UK e-Science All Hands Meeting*, August 31-September 3, 2004, Nottingham.

Marcus, C., et al. (2007). *Understanding research behaviors, information resources, and service needs of scientists and graduate students: A study by the University of Minnesota Libraries.* Retrieved July 7, 2009, from http://lib.umn.edu/about/scieval/Sci%20Report%20Final.pdf

Pritchard, S.M., Anand, S., & Carver, L. (2005). *Informatics and knowledge management for faculty research data.* Retrieved July 7, 2009, from http://net.educause.edu/ir/library/pdf/ERB0502.pdf

Pryor, G. (2007). Project StORe: Making the connections for research. *OCLC Systems & Services, 23*(1), 70-78.

Research Information Network. (2008). *To share or not to share: Publication and quality assurance of research data outputs*. Retrieved June 30, 2009, from http://www.rin.ac.uk/data-publication

Rice, R. (2007). DISC-UK DataShare Project: Building exemplars for institutional data repositories in the UK. *IASSIST Quarterly, 31*(3/4). Retrieved July 7, 2009, from http://www.iassistdata.org/publications/iq/iq31/iqvol313rice.pdf

Witt, M. (2009). Eliciting faculty requirements for research data repositories. *4th International Conference on Open Repositories*.