Purdue University
Purdue e-Pubs

Data Curation Profiles Toolkit

11-29-2010

# The Data Curation Profiles Toolkit: Interview Worksheet

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

Follow this and additional works at: http://docs.lib.purdue.edu/dcptoolkit

Part of the Library and Information Science Commons

Recommended Citation
Carlson, Jake, "The Data Curation Profiles Toolkit: Interview Worksheet" (2010). *Data Curation Profiles Toolkit.* Paper 3.
http://dx.doi.org/10.5703/1288284315652

This work is licensed under a Creative Commons Attribution-Noncommercial-Share Alike 4.0 License.
This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.

# The Data Curation Profiles Toolkit

# Interview Worksheet

| | |
|---|---|
| **Author** | Jake Carlson |
| **Publisher** | Purdue University Libraries / Distributed Data Curation Center |
| **Contact** | jrcarlso@purdue.edu |
| **Date of Creation** | November 29, 2010 |
| **Date of Last Update** | November 29, 2010 |
| **Version** | V 1.0 |
| **Acknowledgement** | Based on research funded by the IMLS (LG-06-07-0032-07) "Investigating Data Curation Profiles across Research Domains" by D.S. Brandt, J. Carlson, M. Witt (Purdue University Libraries), M. Cragin, C. Palmer (GSLIS - University of Illinois Urbana-Champaign). |
| **URL** | http://www.datacurationprofiles.org |
| **License** | Creative Commons Attribution-NonCommercial-ShareAlike 3.0 License. |

Name of Interviewee: _____

Department or Center: _____

Institution: _____

This worksheet is designed to elicit the information necessary to develop a data curation profile about data from a particular research project. In your responses to this worksheet and to the interview questions, please limit your focus to the data associated with the particular research project you have selected.

This worksheet is meant to be filled out as a part of the interview and your responses to the questions in each module will serve to guide the conversation. The interviewer will then ask you some follow up questions about your responses to gather additional details and to better understand your priorities and needs.

If you have any questions, need more information, or would like clarification about any item listed in this worksheet, please do not hesitate to ask your interviewer.

**Please do not turn the page until you are asked to do so.**

## Module 1 – The Data Set

Please provide a brief description of the data:

_____

_____

_____

_____

_____

_____

## Module 2 – The Lifecycle of the Data Set

**1. Initial Data Stage: _____**

- Approximately how many data files exist at this stage?

  _____

- What is the approximate average size of each file at this stage?

  _____

  (Please include the unit of measurement – kb, MB, GB, etc.)

- What format(s) are the data in?

  _____

  (For example: MS Excel 2007, MySQL database, JPEG 2000 images, a raw data file from a Campbell CR10 data logger, etc.)

**2. Second Data Stage: _____**

- Approximately how many data files exist at this stage?

  _____

- What is the approximate average size of each file at this stage?

  _____

  (Please include the unit of measurement – kb, MB, GB, etc.)

- What format(s) are the data in?

  _____

  (For example: MS Excel 2007, MySQL database, JPEG 2000 images, a raw data file from a Campbell CR10 data logger, etc.)

**Please continue on to the next page**

3. **Third Data Stage: _____**

- Approximately how many data files exist at this stage?

  _____

- What is the approximate average size of each file at this stage?

  _____

  (Please include the unit of measurement – kb, MB, GB, etc.)

- What format(s) are the data in?

  _____

  (For example: MS Excel 2007, MySQL database, JPEG 2000 images, a raw data file from a Campbell CR10 data logger, etc.)

4. **Fourth Data Stage: _____**

- Approximately how many data files exist at this stage?

  _____

- What is the approximate average size of each file at this stage?

  _____

  (Please include the unit of measurement – kb, MB, GB, etc.)

- What format(s) are the data in?

  _____

  (For example: MS Excel 2007, MySQL database, JPEG 2000 images, a raw data file from a Campbell CR10 data logger, etc.)

**Please continue on to the next page**

5. **Fifth Data Stage: _____**

- Approximately how many data files exist at this stage?

  _____

- What is the approximate average size of each file at this stage?

  _____

  (Please include the unit of measurement – kb, MB, GB, etc.)

- What format(s) are the data in?

  _____

  (For example: MS Excel 2007, MySQL database, JPEG 2000 images, a raw data file from a Campbell CR10 data logger, etc.)

**Please list any additional stages below or on the back of this sheet as needed**

## Module 3 – Sharing

1. In the previous module you identified the stages that your data goes through during its lifecycle.

   In the table below, please indicate what data you would be willing to share and with whom? (Please place a checkmark in as many boxes as apply)

| | Would not share with anyone | Would share with my immediate collaborators | Would share with others in my research center or at my institution | Would share with others in my field | Would share with others outside of my field | Would share with anyone |
|---|---|---|---|---|---|---|
| Initial Data Stage | | | | | | |
| Second Data Stage | | | | | | |
| Third Data Stage | | | | | | |
| Fourth Data Stage | | | | | | |
| Fifth Data Stage | | | | | | |
| Additional Data Stage(s) – (if needed) | | | | | | |

2. In the table above, if you indicated that you would be willing to share any of your data with:

   - others in your field,
   - others outside of your field, or
   - with anyone

   then please indicate **when** you would be willing to share the data with each group:

   (ex. immediately after the results are published, 6 months after the project is completed, etc.)

   _____

   _____

3. Who would you imagine would be interested in this data?
   (For example, other researchers in my field, researchers outside of my field, practicing professionals, policy makers, etc.)

4. How would you imagine this data being used by the groups / people you listed in the previous question? What value would the data have for these groups / people?

5. Would you place any conditions on sharing your data with the groups or people you have identified (such as requiring some form of acknowledgement, etc.)?

   If so, what would those conditions be?

## Module 4 – Access

1. Have you ever deposited any of this data into a data repository?

   Yes No I don't know

   a. If you answered "yes" to this question, which data and to what particular repository?

      _____

2. Would you be willing to submit your data to a data repository?

   Yes No I don't know

3. If you answered "yes" to question #2, please answer the following questions: (otherwise please leave these questions blank)

   a. At what stage in the data's lifecycle would you submit your data to the repository?

      _____

   b. Some data repositories offer to embargo data sets before making them accessible to others. An embargo can be defined as restricting access to the data for a designated period of time before releasing the data for public access.

      If you would be willing to submit this dataset to a repository but require an embargo, how long will the embargo need to last?

      ___ No embargo - People should be able to access my dataset immediately
      ___ 1 - 3 month embargo
      ___ 4 - 6 month embargo
      ___ 7 month to 1 year embargo
      ___ More than 1 year but less than 2 years
      ___ 2 or more years but less than 5 years
      ___ 5 years or more
      ___ After my retirement
      ___ After my death

      ___ Other (please explain): _____

      _____

**4.** Please prioritize your need for the following types of services for your data.

| | Not a priority | Low Priority | Medium Priority | High Priority | I Don't Know or N/A |
|---|---|---|---|---|---|
| The ability to cite this dataset in my publications. | | | | | |
| A requirement that others cite this data set if they were to use it in their research. | | | | | |
| The ability to access the data set at a secondary (mirror) site if the repository is "off-line". | | | | | |
| The ability to restrict access to the data set to authorized individuals. | | | | | |

## Module 5 – Transfer of Data / Ingest into a Repository

1. What preparations or actions (if any) would need to take place before your data could be ingested into a repository or otherwise transferred out of your direct control for curation purposes?

_____

_____

_____

_____

2. Please prioritize your need for the following types of services for your data.

| | Not a priority | Low Priority | Medium Priority | High Priority | I Don't Know or N/A |
|---|---|---|---|---|---|
| The ability for me to be able to submit this data to a repository myself. (I initiate and perform the submission process) | | | | | |
| The process of submitting this data to a repository is automated. (The submission process is initiated through a trigger – a date, an event, etc., and done without my intervention) | | | | | |
| The ability to batch upload this data into a repository. | | | | | |

## Module 6 – Organization and Description of Data

1. Please explain briefly how this dataset is organized, and how it has been described (e.g. "detailed annotations", "a code book", "a data dictionary", “column headings in a spreadsheet”, etc.).

   If you have used any standardized forms of description or metadata please identify the standard(s) in your response.

_____

_____

_____

_____

2. Is this amount of organization and description sufficient for another person with similar expertise to be able to understand and properly use the data?

   Yes No I don’t know

3. Please prioritize your need for the following types of services for your data.

| | Not a priority | Low Priority | Medium Priority | High Priority | I Don't Know or N/A |
|---|---|---|---|---|---|
| The ability to make the data accessible in multiple formats. | | | | | |
| The ability to apply standardized metadata from your field or discipline to the dataset. | | | | | |

## Module 7 – Discovery

1. Please prioritize your need for the following types of services for your data.

| | Not a priority | Low Priority | Medium Priority | High Priority | I Don't Know or N/A |
|---|---|---|---|---|---|
| The ability for researchers within my discipline to easily find this dataset. | | | | | |
| The ability for researchers from outside of my discipline to easily find this dataset. | | | | | |
| The ability of the general public to easily find this data set. | | | | | |
| The ability for people to easily discover this dataset using Internet search engines (e.g., Google). | | | | | |

2. How do you imagine that people would find your data set?

_____

_____

_____

## Module 8 – Intellectual property

1. Who is the owner of the data?

_____

2. What are your funding sources for this research?

_____

3. Do any of your funding sources require that you: (please circle your response)

   - Draft a data management plan as a condition of funding?

     Yes No I don't know

   - Share your data with others, publish your data, or deposit your data into a data repository?

     Yes No I don't know

   - Preserve your data beyond the life of the funding?

     Yes No I don't know

4. Is this dataset bound by any privacy or confidentiality concerns?

   Yes No I don't know

   - If "yes" please explain:

   _____

   _____

## Module 9 – Tools

1. What tools – software or hardware – are used in generating the data?

(e.g., “a data logger”, “a remote sensor”, etc. – please describe.)

_____

_____

_____

2. What tools - software or hardware - are required to utilize the data?

(e.g., "Microsoft Excel 2003", “GIS”, "I wrote my own program”, etc. – please describe.)

_____

_____

_____

_____

3. Please prioritize your need for the following types of services for your data.

| | Not a priority | Low Priority | Medium Priority | High Priority | I Don't Know or N/A |
|---|---|---|---|---|---|
| The ability to connect the data set to visualization or analytical tools. | | | | | |
| The ability of others to comment on or annotate the data set. | | | | | |

## Module 10 – Linking / Interoperability

1. In the journals, or places you publish most often, are data, or other supplemental information, accepted for publication?

   Yes No I don't know

   - If "yes" please list the journal or place of publication:

   _____

   _____

2. Please prioritize your need for the following types of services for your data.

| | Not a priority | Low Priority | Medium Priority | High Priority | I Don't Know or N/A |
|---|---|---|---|---|---|
| The ability to connect your data with publications or other outputs. | | | | | |
| The ability to support the use of web services APIs. | | | | | |
| The ability to connect or merge your data with other data sets. | | | | | |

## Module 11 – Measuring Impact

1. Please prioritize your need for the following types of services for your data.

| | Not a priority | Low Priority | Medium Priority | High Priority | I Don't Know or N/A |
|---|---|---|---|---|---|
| The ability to see usage statistics on how many people have accessed this data. | | | | | |
| The ability to gather information about the people who have accessed or made use of this data. | | | | | |

2. Beyond usage statistics, are there any other measurements or analytics that you would like to apply to your data?

_____

_____

_____

## Module 12 – Data Management

1. What are the primary ways that you currently manage your data? Please include the storage media(s) and any tools used in your management of the data:

   _____

   _____

   _____

   _____

2. Do you currently make back-up copies of your data?

   Yes No I don't know

   - If you answered "yes" to this question, approximately how often do you make back-up copies?

     _____

3. Do you currently take any security measures to protect your data?

   Yes No I don't know

   - If you answered "yes" to this question, what security measures have you taken?

     _____

4. Please prioritize your need for the following types of services for your data.

| | Not a priority | Low Priority | Medium Priority | High Priority | I Don't Know or N/A |
|---|---|---|---|---|---|
| The ability to enable version control for this dataset. | | | | | |

## Module 13 – Data Preservation

1. What are the most important parts of the data to preserve (manage and maintain over time)?

   _____

   _____

   _____

   _____

2. How long would your data set be useful or have value for you or others if it were to be preserved?

   ___ My dataset does not need to be preserved.
   ___ Less than 3 years.
   ___ 3 years or more but less than 5 years.
   ___ 5 years or more but less than 10 years.
   ___ 10 years or more but less than 20 years.
   ___ 20 years or more but less than 50 years.
   ___ 50 years or more but less than 100 years.
   ___ Indefinitely.

3. If your dataset were to be preserved and maintained for long term access, please prioritize your need for the following types of services.

| | Not a priority | Low Priority | Medium Priority | High Priority | I Don't Know or NA |
|---|---|---|---|---|---|
| The ability to audit this dataset to ensure its structural integrity over time. | | | | | |
| The ability to migrate datasets into new formats over time. | | | | | |
| A secondary storage site for the data set. | | | | | |
| A secondary storage site for the data set at a different geographic location. | | | | | |
| Documentation of any and all changes that were made to the dataset over time. | | | | | |