11-29-2010

The Data Curation Profiles Toolkit: The Profile Template

Jake Carlson
Purdue University, jakecar@umich.edu

Recommended Citation
Carlson, Jake, "The Data Curation Profiles Toolkit: The Profile Template" (2010). *Data Curation Profiles Toolkit.* Paper 4.
http://dx.doi.org/10.5703/1288284315653

The Data Curation Profiles Toolkit

The Profile Template

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. cc BY NC SA

Introduction

The Data Curation Profile is composed of the following sections and sub-sections, each of which is defined below. The information used to populate the profile will largely come from the information gathered through the use of the Interview Worksheet and the Interviewer's Manual. The following table provides a broad overview of how the interview modules correspond to the sections of the Data Curation Profile template:

Interview Module		Data Curation Profile Section
Background / Demographic Questions (Interviewer's Manual)		Section 2 - Overview of the research
Module 1 – The Data Set		Section 3 - Data kinds and stages
Module 2 – The Lifecycle of the Data Set		
Module 3 – Sharing		Section 7 – Sharing & Access
Module 4 – Access		
Module 5 – Transfer of Data / Ingest into a Repository		Section 6 - Ingest / Transfer
Module 6 – Organization and Description of Data		Section 5 - Organization and description of data (incl. metadata)
Module 7 – Discovery		Section 8 - Discovery
Module 8 – Intellectual property		Section 4 - Intellectual property context and information
Module 9 – Tools		Section 9 - Tools
Module 10 – Linking / Interoperability		Section 10 – Linking / Interoperability
Module 11 – Measuring Impact		Section 11 - Measuring Impact
Module 12 – Data Management		Section 12 – Data Management
Module 13 – Data Preservation		Section 13 - Preservation

Keep in mind, however, that the participating researcher may have discussed relevant aspects of the data and their needs for the data outside of a particular module. For example, the participating researcher may have brought up an issue relevant to the discovery of his/her data while answering questions within the "access" module. Therefore it is important to review the interviews as a whole for information to populate the sections of a Data Curation Profile rather than relying solely on information from a particular module.

A Data Curation Profile is meant to incorporate and convey the voice of the researcher. Therefore, although the person constructing the Profile will naturally be the one to craft the language used in populating the sections and sub-sections, care should be taken to provide an authentic representation of the researcher's voice in the Data Curation Profile.

Do not drop sections or sub-sections from the Data Curation Profile (unless of course the decision to exclude a particular module was made before the interviews were conducted). It is likely that there will be some gaps in the information gathered that will lead to some of the sections or sub-sections in the Data Curation Profile containing little or no content. If this happens, do not delete the section or sub-section from the Profile or leave it blank. Instead, indicate that not enough information was available from the information collected to make an entry by writing "Not discussed by the researcher" or a similar statement.

Not every section contains defined sub-sections. Additional sections and sub-sections can be added to individual Data Curation Profiles as warranted and at the profile author's discretion. Relevant information that does not easily fit into a sub-category may be included directly beneath a section heading before the sub-sections are listed.

The amount of time needed to construct a Data Curation Profile will naturally vary according to the amount of information collected, the complexity of the data and the researcher's associated needs for the data, and other factors. Generally it should take approximately 5 – 10 hours to complete a Data Curation Profile.

After the Data Curation Profile has been drafted you may want to consider asking the researcher(s) who were interviewed to review the draft and provide feedback.

Examples of completed Profiles are available at: http://www.datacurationprofiles.org

Definitions:

- **Ancillary Data:** The supplementary data that are used by the data client to better understand or interpret the primary data. For example, the data client may collect weather data to explain unusual readings from sensors.

- **Data:** The particular data set or collection that was selected by the data client(s) to be profiled. This data set is the focus of the interviews and of the resulting data curation profile.

- **Data Author:** The person (or persons) who generated the data.

- **Data Client:** The person (or persons) who were interviewed and whose perspectives are represented in the data curation profile. The completed profile will represent the data client's particular perspectives and needs with regards to the data. This person(s) may or may not be the data author(s), as they may be taking data from a collaborator, a researcher not directly affiliated with the project, a data repository or another other external source.

- **Data Stage:** A discrete activity or set of activities in the data lifecycle as defined by the data client.

- **Primary Data:** The data that serve as the core elements of the data client's research activities. (Compare with "Ancillary Data")

Data Curation Profile – <name – (see discipline/sub-discipline)>

Profile Author	<Name of the person who constructed this data curation profile>
Author's Institution	<Name of the author's institution>
Contact	<Name of a contact person (if different from the profile author) and email>
Researcher(s) Interviewed	<Name(s), and title(s) of the researcher(s) and any other personnel who were interviewed.>
Researcher's Institution	<Name of the researcher's institution>
Date of Creation	<Date the data curation profile was created>
Date of Last Update	<Date the data curation profile was last updated. Leave blank if no updates have been made.>
Version of the Tool	<Version of the toolkit used to construct this data curation profile. The current version of the Data Curation Profile Toolkit is 1.0>
Version of the Content	<Version of content within the Data Curation Profile itself>
Discipline / Sub-Discipline	<The discipline / sub-discipline or specialty of the researcher interviewed. For example: "Civil Engineering / Traffic Flow" The sub-discipline or specialty will serve as the name of the data curation profile to be listed at the top of the first page and in the header of the document. If more than one researcher was interviewed in constructing this profile then use the discipline / sub-discipline of the researcher whose responses contributed the most to the construction of the data curation profile.>
Sources of Information	<List the sources of information used in constructing the data curation profile. Typically this will include an initial interview, a second interview and the interview worksheet completed by the researcher. Be sure to include the dates of the interviews. For example: • An initial interview conducted on September 8, 2010. • A second interview conducted on September 23, 2010. • A worksheet completed by the scientist as a part of the interviews. • A sample of the profiled data.>
Notes	<Any information that would help reader's understand the information presented in the data curation profile. This includes any modifications or additions made to the interview worksheet, interviewer's manual or the data curation profile structure. Leave blank if there are no notes to include.>
URL	<If the data curation profile is available online then post the URL of its location here.>
Licensing	<List the license and conditions (if any) for sharing your data curation profile with others. Creative Commons license permitting non-commercial by others is recommended.>

Section 1 - Brief summary of data curation needs

Include on the first page beneath the table if possible.

This section is for providing a brief summary of the data client's needs as they relate to data curation. It is meant to give the reader an overview of the issues that are most pertinent or important to the data client in working with his or her data

Section 2 - Overview of the research

This section is meant to provide a summary of the nature of the research being performed by the data client in the context of the data being generated or used.

2.1 - Research area focus

This sub-section should include a brief description of the data clients current overall research area or focus as well as an overview of the specific research project associated with the data. Include the research goals and methodologies used in this research area by the data client.

2.2 - Intended audiences

The information needed to populate this sub-section will likely come from Module 3 – Sharing.

This sub-section is meant to identify who the potential audiences for the data (not the research as a whole) are or might be according to the data client. The audience types listed may be specific ("Researchers studying the effects of climate change on plant growth during the Mesozoic era") or broad ("Climate Change Researchers") as dictated by the data client. Audience types may be those with whom the data client is currently sharing his or her data or audience types the data client imagines would be interested in this data.

2.3 - Funding sources

The information needed to populate this sub-section will likely come from Module 8 – Intellectual Property.

A listing of the data client's sources of funding for his or her research that generated or employed the data being discussed. Include any information pertaining about the disposition of the funding source towards managing, sharing or preserving the data. For example: Is a data management plan required as a condition of award? Is the sharing of data with others encouraged by the funding source?

Section 3 - Data kinds and stages

This section is meant to capture and convey information about the data itself.

3.1 - Data narrative

A description of the data lifecycle, and the stages within the lifecycle, within the context of how the data is used in the data client's research. This description should provide background and context for the data table in 3.2.

3.2 – The data table

The data table provides a summary of key characteristics about the data in tabular form. The content and the structure of the data table will be different for each profile, based on the nature of the data under discussion. The categories listed below are common to many data lifecycles, but are meant only to serve as placeholders. The basis for the actual categories in the data lifecycle will be identified by the researcher in Module 2 – The Lifecycle of the Data Set. However, the broad categories listed in the top row should appear in every profile.

Information in the data table is meant to refer to the primary data being used by the data client. Information about additional data sets that are used to interpret, or better understand the primary data or serves to augment the utility of the primary data may be included in the rows beneath the primary data table as "Ancillary data". Ancillary data and its use should also be described in "Section 3.1 - data narrative" above. (see "definitions" section at the beginning of this document).

If more than one primary data set exits, data tables may be repeated to capture information about both data sets. However, if the additional data sets are different enough to warrant separate treatment in their handling, organization, management or curation in several areas a separate data curation profile should be generated for these data.

In cases where information was not collected or the scientist could not provide a response, the table cell should be left blank.

Data Table Categories:

- Data Stage –

  The data stage category serves to breakdown the data lifecycle for a particular data set into discrete stages.

  The default stages within the "data stage" column are:

  - **Raw:** The data is newly created, generated or acquired.
  - **Processed:** The raw data is reviewed, refined or revised to better enable its use in the research. This may include reducing "noise" in the data, removing elements in the data that are superfluous to its use, or checking for errors. Processing data may also include adding additional or supplementary information including metadata to the data set.
  - **Analyzed:** The stage in which data are critically examined by the researcher(s) to provide information or answers to their research questions. The process of analyzing the data may produce new data sets, by-products, or other outputs that should be accounted for.
  - **Finalized:** The last stage in the data lifecycle in which all re-workings and manipulations of the data by the researcher have ceased.

  The stages within the data lifecycle will be identified by the data client in the Interview Worksheet. These stages may deviate from the default categories listed above depending on the nature of the data, the research being conducted, and the characterization of the data stage by the data client. For example, there may be a stage in which a data set is interpolated with another data set.

  The data client will provide a descriptive title for each data stage as a part of filling out "Module 2 – The Lifecycle of the Data Set" in the interview worksheet. Unless these titles are overly long, laden with jargon, or do not sufficiently capture or describe the work being done at this stage, the descriptive title assigned to the data stage by the data client should be used in the data table. Where appropriate, a sentence or two about the lifecycle stage should be included in the "other/notes" column (to compliment a richer description in the "data narrative" section).

  Data stages should be fairly broad in scope yet distinctive from each other, and sufficient in number to capture all events in the data's lifecycle. It is better to group like events in the lifecycle (such as different types of processing or analysis) together rather than to list them separately provided they occur sequentially. Listing between 4 to 8 data stages will generally be sufficient.

  The data specifically designated by the data client to make publicly available are to be indicated by shading the row representing the corresponding data stage in gray.

- **Output –**

  The data type and/or main characteristics of the data at each stage.

- **# of Files/Typical Size –**

  The approximate number of data files and the size of the average file within the data set including the unit of measurement used (KB, MB, etc.). The more precise this information can be the better, but rough estimations are acceptable if they are the best information available.

- **Format –**
  The format of the data files. Listing the file extension only is acceptable for common formats – (for example: .xls)

- **Other / Notes –**
  Any information that helps to clarify or convey information about the data at this particular stage in its lifecycle should be included here.

Data Stage	Output	# of Files / Typical Size	Format	Other / Notes
		Primary Data		
Raw				
Processed				
Analyzed				
Finalized				
		Ancillary Data		
Ancillary Data #1				
Ancillary Data #2				

Note: The data specifically designated by the scientist to make publicly available are indicated by the rows shaded in gray (the "processed" row is shaded here as an example). Empty cells represent cases in which information was not collected or the scientist could not provide a response.

3.3. - Target data for sharing

Much of the information needed to populate this sub-section will likely come from Module 3 – Sharing.

A statement about the data to be shared with others as identified by the data client. At minimum, this statement should identify at what stage in the data lifecycle the data should be shared. The data should also be identified in the profile by shading the corresponding data stage in "Section 3.2 – The data table" in grey.

3.4 - Value of the data

A statement about the real or potential value of the data, both in general and for specific audiences, as seen from the perspective of the data client.

3.5 - Contextual narrative

Any additional information about the data, the data workflow, or the nature of the data that would help to inform its handling, management or curation.

In addition, descriptions of or information about any disciplinary or sub-disciplinary norms, perspectives, or practices in the handling, managing, or curating data provided by the data client for context may be listed here. For example, "The scientist stated that the "culture" of Earthquake

Engineering community is beginning to move towards a greater understanding and acceptance of sharing of research data."

Section 4 - Intellectual property context and information

This section exists to capture information regarding issues of data ownership, intellectual property and attribution of credit about the data set under discussion.

4.1 - Data owner(s)

A statement of ownership with regards to the data according to the perspective of the data client. If more than one possible owner exists (ex. data author and his/her employing institution), each author should be listed along with a statement regarding their perceived domains of ownership and any potential overlap.

4.2 - Stakeholders

A statement listing any groups, organizations, individuals or others that were identified by the data client as having made an investment in the data or that the data client would feel the need to consult with regarding the data's release and use. For example, graduate students, funding agencies, or publishers (if they accepted data as supplementary files) might be listed as stakeholders by the data client.

4.3 - Terms of use (conditions for access and (re)use)

A statement detailing any conditions or policies for accessing and/or re-using the data that would need to be satisfied in making the data available.

4.4 - Attribution

Much of the information needed to populate this sub-section will likely come from the priority questions on citing data in "Module 4 – Access".

A statement regarding the need for attribution when the data is used by others, and, if appropriate, a statement about the nature or form of the desired attribution.

Section 5 - Organization and description of data (incl. metadata)

This section is meant to provide an overview about how the data are currently organized and described. A secondary purpose is to identify the shortcomings in the description and organization of the data (particularly those brought up by the data client) as well as possible options to address these shortcomings.

5.1 - Overview of data organization and description (metadata)

A broad summary of how the data are currently organized and described. Any metadata associated with the data set should be listed here.

This entry should also contain information about the needs and desires of the data client relating to the organization and description of the data. This includes any stated needs by the data client to make the data accessible in a particular format or in multiple formats for different purposes.

5.2 - Formal standards used

Any application of formal metadata, ontological, or other standards pertaining to the organization or description of the data are to be listed here.

In addition, if the data client is aware of any field-specific standards regarding metadata, ontologies, etc. that could or should be applied; these should be listed in this sub-section as well.

5.3 - Locally developed standards

If the data client or others working with the data have developed any in-house or local standards, these should be described in this sub-section.

5.4 - Crosswalks

This sub-section provides a space to indicate if crosswalks between standards (local or formal) are in place, or if they will be needed in sharing, curating, or preserving the data.

5.5 - Documentation of data organization/description

This sub-section should contain information about any existing documentation and/or documentation practices in place regarding the description and/or organization of the data. Any existing papers or documents that provide such a description--e.g., standard operating procedure documents or papers describing methodologies that include organizational practices—should be included.

Section 6 - Ingest / Transfer

This section should list any needs or functionalities described by the data client with regards to the details of ingesting the data into a repository or transferring it to the data curator. This includes any preparations or actions needed before the ingestion or transfer of data would take place, as well as data client preferences in submitting the data.

Section 7 – Sharing & Access

This section describes the needs and opinions of the data client regarding sharing the data publicly and making it accessible to others.

7.1 - Willingness / Motivations to share

Statements regarding the data client's feelings/reservations/willingness towards sharing their data with others at various stages of the data lifecycle.

7.2 - Embargo

Indicate if the data client requires an embargo period before releasing the data publicly, and if so how long an embargo would be needed. This section also includes the nature of and reason for that embargo.

7.3 - Access control

A narrative regarding the data client's feelings about the need to restrict or control access to the data to or from particular parties.

7.4 Secondary (Mirror) site

Information regarding the data client's priority and thoughts about the need for a mirror site and whether the site is needed in a separate geographical location.

Section 8 - Discovery

A narrative regarding general needs/desires of the researcher with regards to the discovery of the data, both within the specific context of a repository (if discussed) and in general. This includes the perceived need for researchers in the same discipline as the data client, researchers in other disciplines and the general public to discover the data, through internet search engines or other methods.

Section 9 - Tools

This section is designed to provide information about any tools that were used to generate the data (if applicable), as well as tools that would be needed by users of the repository to access, use, visualize, interpret or interact with the data as reported by the data client.

Section 10 – Linking / Interoperability

A statement describing needs for the data to interoperate with other datasets or to be linked / connected with other documents. This includes the need for web APIs, citation in publications, or any comments about integration of/dependency on 3rd party data, etc.

Section 11 - Measuring Impact

This section is meant to provide information regarding any needed or desired metrics that demonstrate the data's impact.

11.1 - Usage statistics & other identified metrics

A statement on the need for usage statistics and/or other metrics identified by the data client for the data, as well as any details on how usage could/should be measured according to the data client.

11.2 - Gathering information about users

If the data client states a need to gather information about the individuals who are accessing and/or using his/her data, details of this need should be listed here.

Section 12 – Data Management

This section should contain any pertinent information about how the data is currently being managed, as well as any needs relating to data management in making the data accessible to others.

If needed, a general statement about the researcher's current or anticipated data management needs can be inserted here.

12.1 - Security / Back-ups

A statement about the data client's current and desired security and back-up practices. Be sure to clear differentiate between the current and desired practices in the profile.

12.2 - Secondary storage sites

Information about the data client's current use of and anticipated need for off-site storage in managing and/or curating this data, both in the same location and at different geographical locations.

12.3 - Version control

A narrative about current practices and needs/desires for version control of the data set in managing and/or curating the data.

Section 13 - Preservation

This section contains information about the needs / desires of the data client regarding the preservation of the data set under discussion.

If needed, a general statement about the researcher's preservation needs can be inserted here.

13.1 - Duration of preservation

A statement about the length of time the data is to be preserved. The duration may be event based rather than time based, though estimation for the length of time related to the event should be noted where possible.

13.2 - Data provenance

A statement about what information needs to be captured to establish and maintain data provenance as well as the priority placed on capturing data provenance by the data client as a part of curating the data.

13.3 - Data audits

A statement regarding the importance of conducting periodic data audits for the data set. This sub-section should also include the priority level for conducting data audits according to the data client and any information pertaining to what the data audit should consist of.

13.4 - Format migration

A statement about past, current, and future needs/desires regarding migrating the data from its original format to another migration.

Section 14 – Personnel

This section is to be used to document roles and responsibilities of the people involved in the stewardship of this data.

This information can be withheld if the data curation profile is to be made publicly accessible. If the profile author chooses to withhold this information then enter "(Withheld)"

14.1 - Primary data contact (data author or designate)

Identifies the data client and provides contact information for this person.

14.2 - Data steward (ex. library / archive personnel)

Identifies the primary contact responsible for stewarding / curating the data, and provides contact information for this person.

14.3 - Campus IT contact

Identifies the primary contact responsible for the Information Technology aspects of stewarding / curating the data, and provides contact information for this person.

14.4 - Other contacts

The names, title, and contact information for other personnel who are persons of interest in working with the data, such as graduate students, lab assistants, and other personnel involved in the research.

This subsection may also include the names, title, and contact information for other personnel who are persons of interest in curating the data, such as relevant subject librarian, the repository manager, and other library/archival personnel involved in the curation process.