The Data Curation Profile

IASSIST 2010

Jake Carlson
Data Research Scientist
Purdue University Libraries
Investigating Data Curation Profiles Across Research Domains

Goals:

- A better understanding of the practices, attitudes and needs of researchers in managing and sharing their data.
- Identify possible roles for librarians and the skill sets they will need to facilitate data sharing and curation.
- Develop “data curation profiles” – a tool for librarians and others to gather information on researcher needs for their data and to inform curation services.
Interviews: Background

Agronomy & Soil Science (Purdue & UIUC),
Anthropology (UIUC), Biochemistry (Purdue),
Biology (Purdue), Civil Engineering (Purdue),
Earth & Atmospheric Sciences (Purdue & UIUC),
Electrical & Computer Engineering (Purdue),
Food Science (Purdue), Geology (UIUC),
Horticulture & Plant Science (Purdue & UIUC),
Kinesiology (UIUC), Speech and Hearing (UIUC)
First interviews: workflow

- Research Data Lifecycle (what’s the story of the data)
- Data Management / Storage
- Disposition of the Data
- Data Dissemination and Sharing
- Data Preservation and Repositories
- Roles for Libraries and Librarians
Interviews: Coding

- Data
- Data Management
- Intellectual Property
- Library Services
- Preservation
- Scientific Work Practices

- Sharing and Access
- Stakeholders
- Storage
- Use and Re-use
- Needs
- Field Needs
Examples of Needs

- Need - “Yeah, we’ll make the raw data available at some point, or at least the raw data associated with the manuscript that’s reported. So we have the credit…” (Biologist)

- Field Need – “I would like to see us move towards a community, collaborative world where the boundary conditions are a community thing, and community modifiable… If you tied it in with the [sub-discipline] modeling community, all of the sudden you’d have this kind of interesting interaction.” (Earth & Atmospheric Scientist)
Second interviews: detailed needs

- Follow up questions to close information gaps
- Worksheet for subjects to fill out to enable us to make comparisons
“How long should your data be preserved?”

- No preservation: 6
- Less than 3 years: 1
- 3-5 years: 4
- 5-10 years: 4
- 10-20 years: 0
- 20-50 years: 0
- 50-100 years: 2
- Indefinitely: 0
- No answer: 1

N=19
Preservation Services

- Audit dataset
- Migration to new formats
- Secondary storage site
- Offsite secondary storage site
- Document changes made over time

Priority Levels:
- Not a Priority
- Low Priority
- Medium Priority
- High Priority
- Don't Know or NA
The Data Curation Profile

- A means to capture requirements for specific data generated by a single scientist or lab, based on their reported needs and preferences for these data.
- A concise, structured document suitable for sharing and annotation.
- A resource for Librarians, Archivists, IT Professionals, Data Mangers, and others.
DCP Sections

- Information about the Data and its Context
  - Overview of the Research
    - Focus
    - Intended Audience
    - Funding
  - Data Kinds and Stages
    - Data Narrative (data lifecycle)
    - Target Data for Sharing
    - Use/re-use Value
    - Contextual Narrative
# Data Table – Traffic Flow

<table>
<thead>
<tr>
<th>Data Stage</th>
<th>Output</th>
<th>Typical File Size</th>
<th>Format</th>
<th>Other / Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Primary Data</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Raw</td>
<td>Sensor data</td>
<td>100k in 1 file per day</td>
<td>proprietary to the sensor</td>
<td>FTP downloads are mostly automated.</td>
</tr>
<tr>
<td>Processing Stage 1</td>
<td>Sensor data – open/accessible format</td>
<td>Roughly 6kb</td>
<td>.csv / .xls</td>
<td>Data are formatted into .csv before bring reformatted into a mySQL database.</td>
</tr>
<tr>
<td>Processed</td>
<td>Data vectors</td>
<td>800 records per intersection per day.</td>
<td>SQL / .xls</td>
<td>Data are extracted from the mySQL database for analysis purposes.</td>
</tr>
<tr>
<td>Analyzed</td>
<td>charts/Graphs</td>
<td></td>
<td>.xls / .emf</td>
<td>charts and graphs used for interpretation.</td>
</tr>
<tr>
<td>Published</td>
<td>charts/graphs</td>
<td></td>
<td>.ppt</td>
<td>Data are presented via power point.</td>
</tr>
<tr>
<td><strong>Augmentative Data</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Image</td>
<td>Stills taken from video</td>
<td>.gif /.jpg / .ppt</td>
<td></td>
<td>Images generated from video.</td>
</tr>
</tbody>
</table>
DCP Sections

- Information about Needs
  - Intellectual Property
  - Organization and Description of Data
  - Ingest
  - Access
  - Discovery
  - Tools
  - Interoperability
  - Measuring Impact
  - Data Management
  - Preservation
Next Steps

- Going from research project to operational tool for use by others

A Data Curation Profile would be developed through 3 stages:

- Stage 1 – Preparation
- Stage 2 – Worksheet & Interviews
- Stage 3 – Constructing the Profile
Next Steps

The Data Curation Profile Tool will consist of 4 components:

- User Guide
- Interview Guide
- Interview Worksheet
- Template
Potential Uses of the DCP

At an individual level, the Data Curation Profile:

- Provides a structure for conducting a data interview between an information professional and a researcher or research group.
- Provides a means for a researcher or a research group to go beyond a cursory introduction of data curation concepts and practices to get at more the granular issues and decisions that will need to be made.
Potential Uses of the DCP

At an institutional level, the Data Curation Profile:

- Serves as a foundational document to guide the management and/or curation of a particular data set.
- May be used to inform the development of data services to be offered by the institution, as well as to help to identify the types of tools, infrastructure and responsibilities for data services staff.
Potential Uses of the DCP

At a public level, the Data Curation Profile:

- May be used by others as a guide in developing data services at their own institutions.
- May be used as objects of research to further a better understanding of data types researchers want or need to share, curate or preserve, and the needs of researchers in doing so.
http://www.ijdc.net/index.php/ijdc/article/view/137

Project wiki - http://datacurationprofiles.org
Questions?

IASSIST 2010

Jake Carlson
Data Research Scientist
Purdue University Libraries