Who is Responsible for CENS Data?

Jillian C. Wallis, Christine L. Borgman, & Matthew S. Mayernik

Data Practices - Monitoring, Modeling, & Memory

Introduction: Data availability relies on someone being responsible for the data

Making Data Available

- CENS data need to be made available
  - As requested by the NSF and other funding agencies for accountability
  - To make data reusable for other research
- Someone is responsible for CENS data
  - The curation of data is predicated on the availability of data. If data are available then someone has made them available, and this person is presumably responsible for the data.
- Problem of defining terms
  - What does it mean to make data available?
  - What does it mean to be responsible for data?

Impacts of “Responsibility”

- 2008 RIN Stewardship of Digital Research Data Guidelines
  - “establish a framework of codes of practice to ensure that creators and users of research data are aware of and fulfill their responsibilities in accordance with these principles”
- Lack of data reporting during the Annual Report
  - CENS groups were asked to report data that had been collected during the reporting cycle
  - Of the 60+ groups, only 11 reported data
- We can not assume that CENS researchers are responsible for their data, because we have no shared definition of responsibility

Problem Description: There is no shared understanding of who is responsible for data

- Being a responsible researcher has in the past indicated:
  - Choosing appropriate methods
  - Designing effective experiments
  - Only making assertions that could be supported by the data
- Responsibility in metadata (library usage)
  - When an item is catalogued, metadata about the item is recorded to provide various access points, such as title, genre, or author
  - “[Author fields] are made for persons, corporate bodies, and meetings having some form of responsibility for the creation of the work, including intellectual and publishing responsibilities.” - Marc 21

Proposed Solution: Attempt to capture the various ways to interpret responsibility

Approach

- Interview studies
  - Round 1 (2006): 24 participants were chosen from a stratified random sample stratified by research was part of technology or an application science
  - Round 2 (2010): 20 participants were chosen from a stratifies random sample stratified on the order of magnitude of each researcher’s betweenness centrality in the CENS bibliometric network
  - Interviews were recorded with each participant following an approved interview question protocol that covered many aspects of data practices at CENS, including variables collected, data sharing, etc.
  - Interviews averaged 60 minutes apiece, which were transcribed, and then coded in NVivo to identify emergent themes
  - Part of the data from this larger data practices study was then mined for insight into the problem of understanding different concepts of responsibility for data
- Interview questions of interest
  - Would you consider yourself the author of a dataset? Or what are your criteria for determining authorship of data?
  - Who would you consider to be the owner of a dataset?
  - Who is responsible for a dataset?
- Data practices study participants

<table>
<thead>
<tr>
<th>Interviews</th>
<th>Terrestrial</th>
<th>Contam</th>
<th>Aquatic</th>
<th>Seismic</th>
<th>Urban</th>
<th>General</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Application Scientists</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Faculty</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Staff</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Students</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Technology Scientists</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Faculty</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Staff</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Students</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>Totals</td>
<td>14</td>
<td>5</td>
<td>4</td>
<td>6</td>
<td>3</td>
<td>3</td>
<td>20</td>
</tr>
</tbody>
</table>

Figure 1: Distribution of participants interviewed as part of round 1 and round 2.

It should be noted that the first interview study was largely focused on the ecological sciences (Terrestrial, Contam, and Aquatic) and as a result did not solicit any participants from Seismic, Urban Sensing, or any non-application affiliated technology researchers. Six researchers were interviewed during both interview rounds, and have been counted twice within this table.

Figure 2: Variety of responses of who would be considered the data author, data owner, or entity responsible for the data.

Conclusions

- Response variety
  - Even more nuanced than the figure above can capture
  - An Indication of no shared understanding of responsibility for data
  - There is no clear entity responsible for the data from any interpretation of the terminology
  - This needs to be remedied if we want a CENS data repository to succeed
- Problematic concepts of responsibility
  - Data Authorship: “authoring” of data does not fit within researchers’ conception of what it means to author
  - Data Ownership: a more common concept, but often conflated with the concept of access
  - Data Responsibility: the variety of interpretations of responsibility makes responses to this question difficult to compare

Future research

- As an exploratory study this has given us a good idea of where future research is needed in order to capture more consistent answers
- Follow-up interviews need to request definitions of terms used, solicit alternate terms, and discuss why one term would be more applicable than another