The Data Census: Assessing Data Services at MSU

Sara Mannheimer
Data Management Librarian, Montana State University
sara.mannheimer@montana.edu
@saramannheimer

CLIR/DLF E-Research Network
October 14, 2015
Survey Development

Collaboration between Library and Campus IT (with input from Office of Planning and Analysis)

Informed by other institutional surveys and assessment frameworks (see resources slide)
Survey Development

Size and types of research data

Data storage and backup behaviors

Data sharing behaviors
Survey Goal

Gauge interest in and need for institutional data curation, storage, and publication services
Distributing the Census

Distributed in January 2015

Email to academic faculty and staff was cosigned by Vice President for Research, Dean of the Library, and CIO

Vice President for Research, Dean of the Library, and CIO helped with publicity and encouraged responses

Qualtrics survey tool
Assumptions
Assumptions

Most researchers store their data digitally, on a personal computer or external hard drive.

Researchers create a wide variety of data formats.

Researchers have relatively small amounts of data—less than 100 Gb.
Assumptions

Researchers share data within research teams and labs

Researchers share data after article publication or grant completion

Researchers with bigger data are more likely to need campus data services
Survey Results
Demographics
Demographics
Survey Results:
Data Storage, Format, and Size
Assumptions

Most researchers store their data digitally, on a personal computer or external hard drive.
Data Storage

<table>
<thead>
<tr>
<th></th>
<th>number</th>
<th>percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paper copies</td>
<td>79</td>
<td>46.20%</td>
</tr>
<tr>
<td>Computer</td>
<td>131</td>
<td>76.61%</td>
</tr>
<tr>
<td>External Drive</td>
<td>108</td>
<td>63.16%</td>
</tr>
<tr>
<td>MSU Resource</td>
<td>28</td>
<td>16.37%</td>
</tr>
<tr>
<td>Cloud service</td>
<td>44</td>
<td>25.73%</td>
</tr>
<tr>
<td>Off Campus</td>
<td>22</td>
<td>12.87%</td>
</tr>
<tr>
<td>Other</td>
<td>17</td>
<td>9.94%</td>
</tr>
</tbody>
</table>
Assumptions

Researchers create a wide variety of data formats
Data Format

<table>
<thead>
<tr>
<th></th>
<th>number</th>
<th>percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spreadsheet</td>
<td>140</td>
<td>81.87%</td>
</tr>
<tr>
<td>Text</td>
<td>125</td>
<td>73.10%</td>
</tr>
<tr>
<td>Audio</td>
<td>32</td>
<td>18.71%</td>
</tr>
<tr>
<td>Video</td>
<td>48</td>
<td>28.07%</td>
</tr>
<tr>
<td>Images</td>
<td>91</td>
<td>53.22%</td>
</tr>
<tr>
<td>GIS</td>
<td>26</td>
<td>15.20%</td>
</tr>
<tr>
<td>Timeseries</td>
<td>42</td>
<td>24.56%</td>
</tr>
<tr>
<td>Sensor Feeds</td>
<td>27</td>
<td>15.79%</td>
</tr>
<tr>
<td>Other</td>
<td>35</td>
<td>20.47%</td>
</tr>
</tbody>
</table>
Assumptions

Researchers have relatively small amounts of data—less than 100 Gb.
## Data Size

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>blank</td>
<td>2</td>
</tr>
<tr>
<td>Don’t Know</td>
<td>13</td>
</tr>
<tr>
<td>&lt; 10GB</td>
<td>32</td>
</tr>
<tr>
<td>10 - 100 GB</td>
<td>57</td>
</tr>
<tr>
<td>100 - 1000 GB</td>
<td>38</td>
</tr>
<tr>
<td>&gt; 1000 GB</td>
<td>29</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>171</strong></td>
</tr>
</tbody>
</table>

![Pie chart showing data size distribution](chart.png)
Survey Results: Data Sharing
Assumptions

Researchers share data within research teams and labs
Data Sharing

<table>
<thead>
<tr>
<th></th>
<th>number</th>
<th>% of Total</th>
<th>% of Sharers</th>
</tr>
</thead>
<tbody>
<tr>
<td>University</td>
<td>89</td>
<td>52.05%</td>
<td>74.17%</td>
</tr>
<tr>
<td>Federal</td>
<td>33</td>
<td>19.30%</td>
<td>27.50%</td>
</tr>
<tr>
<td>Lab</td>
<td>28</td>
<td>16.37%</td>
<td>23.33%</td>
</tr>
<tr>
<td>K-12</td>
<td>2</td>
<td>1.17%</td>
<td>1.67%</td>
</tr>
<tr>
<td>Open</td>
<td>16</td>
<td>9.36%</td>
<td>13.33%</td>
</tr>
</tbody>
</table>
Assumptions

Researchers share data after article publication or upon grant completion.
### Public Data Sharing

<table>
<thead>
<tr>
<th></th>
<th>number</th>
<th>% of Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>After Publication</td>
<td>72</td>
<td>42.11%</td>
</tr>
<tr>
<td>After Grant End</td>
<td>9</td>
<td>5.26%</td>
</tr>
<tr>
<td>Immediately</td>
<td>19</td>
<td>11.11%</td>
</tr>
<tr>
<td>No</td>
<td>70</td>
<td>40.94%</td>
</tr>
</tbody>
</table>

![Bar chart showing public data sharing choices]
Survey Results:
Interest in Data Services
Assumptions

Researchers with bigger data are more likely to need campus data services
Interest in Data Services

<table>
<thead>
<tr>
<th>Number</th>
<th>Want Info</th>
<th>% Want Info</th>
</tr>
</thead>
<tbody>
<tr>
<td>Don’t Know</td>
<td>13</td>
<td>7</td>
</tr>
<tr>
<td>&lt; 10GB</td>
<td>32</td>
<td>19</td>
</tr>
<tr>
<td>10 - 100 GB</td>
<td>57</td>
<td>44</td>
</tr>
<tr>
<td>100 - 1000 GB</td>
<td>38</td>
<td>29</td>
</tr>
<tr>
<td>&gt; 1000 GB</td>
<td>29</td>
<td>24</td>
</tr>
</tbody>
</table>
Survey Results:
Correlated Responses
Those who share data tend to use other people’s data
Those with larger amounts of data tend to use other people’s data
Those who back up their data also annotate their data
Those with large datasets are more likely to annotate their data.
Follow-up Interviews
Follow-up interviews

Interviews were scheduled with researchers from a cross-section of departments and research centers:

- Electrical & Computer Engineering
- Plant Sciences & Plant Pathology
- Agricultural Economics & Economics
- History
- Land Resources & Environmental Science
- Ecology
- College of Engineering
- Physics
- Spatial Sciences Center
Interview Results:
Data Sharing
Data Sharing

Researchers weren’t always sure how large their data was in terms of file size, but could describe the tools and methods they used to work with the data.

Can the data be shared through e-mail attachments?  
Stored on external hard drives?  
Require deposit in institutional clusters or repositories?
Data Sharing

“It's important for data to be public. There should be no black boxes. We need systems that promote transparency and reproducibility in science.”

“The journals where I typically publish have no data archiving requirements, so that's not [an activity] I end up doing.”

“Journals with a high impact factor [in my field] require data.”
Interview Results:
“Big” Data
“Big” Data

History faculty member: a 50 Kb dataset is “big” data.

Another researcher working with 10 Gb of data: not “big” data, since it is smaller than previous projects.

"Managing multiple 'tiny data' sets is just as labor-intensive as [working with] big data."
Interview Results:
Collaboration
Collaboration

All interviewees described challenges when collaborating and working with collaborative data across institutions:

different IT infrastructures

different data handling routines

different storage and access requirements

interdisciplinary cultural practices
Collaboration

“What we need to create is an interdisciplinary ‘Rosetta Stone’ to make data shareable”

“There is no language to talk about data processes that facilitate sharing”
Interview Results:
Library Love
Library Love

All interviewees said they would use MSU Library data services:

- Dataset annotation and metadata markup
- Assistance with deposit in relevant data repositories
- Educational programs and training on campus IT resources
- Data outreach and publicity
Discoveries
Discoveries

The data census created a foundation of knowledge about data on campus
Discoveries

The census will help us improve our understanding of the local research data environment
Discoveries

The census will inform Campus IT and Library priorities
Discoveries

The census confirmed that Campus IT and the Library must work closely together to meet data needs at MSU
Lessons Learned

Research other institutional data surveys and assessment frameworks

Balance length of survey with depth of information gathered

Advocate for the information the library wants to collect.

Seek IRB approval to enable data sharing

(Survey instrument: http://doi.org/10.15788/m2h59m)
Next steps

Use the data census to validate other research (e.g. Sentiment Analysis project)

Revise and redistribute the survey in early 2016
Resources