Services for Big Data...
and the Long Tail

Patricia Cruse
Stephen Abrams
Carly Strasser
Perry Willett

University of California Curation Center
California Digital Library
http://www.cdlib.org/uc3
OSTP March 2012
Big Data Effort Launched
The Little Guys?
The Long Tail

Size of dataset

# datasets

Hampton et al. *Forthcoming* 2012
The Long Tail

Size of dataset

# datasets
# researchers

Hampton et al. *Forthcoming* 2012
The Long Tail

Size of dataset

#datasets
#researchers
#grants

Hampton et al. *Forthcoming* 2012
The Long Tail

Size of dataset grant ($)

# datasets
# researchers
# grants

Data Managers

All alone

Hampton et al. *Forthcoming* 2012
UGLY TRUTH

Many (most?) researchers...

are not taught data management
don’t know what metadata are
can’t name data centers or repositories
don’t share data publicly or store it in an archive
aren’t convinced they should share data
The Fallout: Where data end up
Barriers

- Cost: time, personnel, software & hardware
- Confusion about standards, how to start
- Disparate datasets
- Lack of training
- Fear of lost rights/benefits
- No incentives
Two Services to Mitigate Long Tail Challenges

1. DMP Tool
   Build your Data Management Plan

2. DataUp
Data Management Planning – DMPTool
Meeting funding agencies data management plan requirements

- Seven institutions
- Helps researchers meet funding agency requirements
- Guide researchers through the process of creating a data management plan
- Connects researchers to local resources
- Open to everyone

http://dmptool.org
**Customize resources for your institution**

**DMP Tool**
Guidance and Resources for your Data Management Plan

---

**NSF-BIO: Biological Sciences: 2. Data Storage and Preservation**

What physical and/or cyber resources and facilities (including third party resources) will be used to store and preserve the data?

---

**Progress**

Click on a section below to edit it at any time.

- = complete

**Plan description**

1. Products of Research
2. Data Storage and Preservation
3. Data Formats and Metadata

---

**Help**

Use this section to describe your long-term strategy for storing, archiving and preserving your data. Describe how and where you will make these data and metadata available to the community. Consider these questions:

- What is the long-term strategy for maintaining, curating and archiving the data?

---

**Resources**

- University of California, Office of the President
  - UC3: Data management plans
  - UC3: Merritt Repository Service
  - UC3: Security, storage and backups

**General**

- NSF Data Sharing Policy
- NSF Data Management Plan Requirements
- NSF BIO Directorate Data Management Plan Guidelines
In many cases, get data management advice and resources for your specific institution.

The DMP Tool allows you to:

1. [Description]
2. [Description]
3. [Description]
4. [Description]

Recent DMP News:
- DMPTool workshop at the DLF Fall Forum
- DMPTool demo: Wed Oct 19
- Importance of Data Management Education

See a plan created with the DMP Tool

Photo courtesy of the International Maize and Wheat Improvement Center
Growth rate since launch in Oct 2011

<table>
<thead>
<tr>
<th></th>
<th>Oct-11</th>
<th>Nov-11</th>
<th>Dec-11</th>
<th>Jan-12</th>
<th>Feb-12</th>
<th>Mar-12</th>
<th>Apr-12</th>
<th>May-12</th>
<th>Jun-12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unique Users</td>
<td>180</td>
<td>548</td>
<td>772</td>
<td>1146</td>
<td>1399</td>
<td>1624</td>
<td>1809</td>
<td>2092</td>
<td>2302</td>
</tr>
<tr>
<td>Plans</td>
<td>114</td>
<td>398</td>
<td>582</td>
<td>882</td>
<td>1090</td>
<td>1291</td>
<td>1479</td>
<td>1711</td>
<td>1887</td>
</tr>
<tr>
<td>Institutions</td>
<td>65</td>
<td>155</td>
<td>206</td>
<td>288</td>
<td>331</td>
<td>381</td>
<td>405</td>
<td>419</td>
<td>452</td>
</tr>
</tbody>
</table>
New additions

- Launched an open organization to build a community around the data management issues!

On the horizon

- Adding new functionality
- Tighter community integration and educational components
- Integration with other systems and initiatives
- Routing of DMPs to support personnel
- Add analytics

http://dmptool.org
Open Source Tool

Facilitate sharing, preserving, using, citing, re-using tabular data

Add-in & Web Application

DataUp
Features
Best practices check
Generate metadata
Generate citation
Post data to repository
### Best Practices Check

![Spreadsheet Image]

<table>
<thead>
<tr>
<th>A</th>
<th>L</th>
<th>M</th>
<th>N</th>
<th>O</th>
<th>P</th>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>300</td>
<td>380</td>
<td>200+</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>450</td>
<td>900</td>
<td>200+</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>500</td>
<td>810</td>
<td>755</td>
<td>450+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>400</td>
<td>1450</td>
<td>1100</td>
<td>550</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>350</td>
<td>200</td>
<td>1150</td>
<td>550</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>1000</td>
<td>100</td>
<td>800</td>
<td>850</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>2425</td>
<td>1600</td>
<td>550</td>
<td>2080</td>
<td>250</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>1300</td>
<td>165</td>
<td>800</td>
<td>760</td>
<td>1640</td>
<td>300</td>
</tr>
<tr>
<td>20</td>
<td>1450</td>
<td>1600</td>
<td>600</td>
<td>1440</td>
<td>180</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>1050</td>
<td>2120</td>
<td>500</td>
<td>2200</td>
<td>150</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>1320</td>
<td>1400</td>
<td>950</td>
<td>1500</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>1240</td>
<td>680</td>
<td>400</td>
<td>1640</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>1560</td>
<td>630</td>
<td>840</td>
<td>450</td>
<td>760</td>
<td>180</td>
</tr>
<tr>
<td>25</td>
<td>1000</td>
<td>420</td>
<td>880</td>
<td>700</td>
<td>600</td>
<td>80</td>
</tr>
<tr>
<td>26</td>
<td>0</td>
<td>440</td>
<td>200</td>
<td>480</td>
<td>&lt;50</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>90</td>
<td>350</td>
<td>510</td>
<td>250</td>
<td>180</td>
<td>200</td>
</tr>
<tr>
<td>28</td>
<td>30</td>
<td>350</td>
<td>480</td>
<td>100</td>
<td>60</td>
<td>200</td>
</tr>
<tr>
<td>29</td>
<td>90</td>
<td>180</td>
<td>320</td>
<td>100</td>
<td>240</td>
<td>&lt;100</td>
</tr>
<tr>
<td>30</td>
<td>60</td>
<td>&lt;100</td>
<td>330</td>
<td>&lt;100</td>
<td>0</td>
<td>&lt;100</td>
</tr>
<tr>
<td>31</td>
<td>30</td>
<td>&lt;50</td>
<td>120</td>
<td>&lt;50</td>
<td>30</td>
<td>&lt;50</td>
</tr>
<tr>
<td>32</td>
<td>0</td>
<td>&lt;50</td>
<td>90</td>
<td>&lt;25</td>
<td>0</td>
<td>&lt;25</td>
</tr>
<tr>
<td>33</td>
<td>0</td>
<td>&lt;25</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**Error Details**

- Oct_7m15(6)
- Oct17to26(4)
- Oct27toNov16(6)
- scope_data_No...(6)
- Notes (6)
- red_days (6)

**Remove Selected**

- Fixable Errors
- Non fixable Errors

- Embedded comments
  - Remove
  - Advice
- Embedded charts, tables, pictures
  - Remove
  - Advice
- Color coded text or cell shading
  - Remove
  - Advice
- Non-contiguous data
  - Remove
  - Advice
- Blank Cells
  - Remove
  - Advice
- Special Characters
  - Remove
  - Advice
Generate Metadata

CREATE METADATA

It generates and applies metadata based on a pre-defined schema. Some of the metadata values are automatically populated by the system – the remaining must be supplied by you. Some of the metadata fields are required, designated by red asterisk, and some are optional. Please make sure you provide at least all the required metadata values.

Data descriptions

- Creator:First name
- Creator:Last name
- Creator:Email
- Creator:Phone
- Creator:Organization
- Title of dataset
- Today's date
- Abstract
- Keyword(s)
- Date Contact Person: First name
- Date Contact Person: Last name
- Date Contact Person: Address
- Date Contact Person: City
- Date Contact Person: State/province
- Date Contact Person: Postal code
- Date Contact Person: Country
- Date Contact Person: Phone
- Date Contact Person: Email
- Date Contact Person: Organization
- Geographic coverage: Description
- Geographic coverage: West bounding coordinate
- Geographic coverage: East bounding coordinate
- Geographic coverage: North bounding coordinate
- Geographic coverage: South bounding coordinate
- Temporal coverage: Description
- Temporal coverage: Beginning date
- Temporal coverage: Ending date
- Project title
- Project description
- Intellectual rights
- Data table name
- Data table description
- Identifier
- Citation

Column descriptions

- Creator:First name
- Creator:Last name
- Creator:Email
- Creator:Phone
- Creator:Organization
- Title of dataset

SAVE CANCEL
Create a Data Citation

**CREATE CITATION**

Steps:
1. Retrieve an identifier from the repository for your dataset (Read this important information about your unique identifier here)
2. Check the rest of the information below for accuracy
3. Make any necessary changes or edits to the information
4. Click ‘Generate Citation’ to create your data citation

- **Unique Identifier**: ark://99999/fk40c5cf5
- **Publication year**: 2009
- **Dataset title**: Title here
- **Dataset version**: V1
- **Publisher (Data owner)**: Strasser Carly

**Citation**:
Upload to a Repository

Launching in September 2012
Meeting the Long Tail Challenge

- Engage with the research community early in the lifecycle?
- Deploy simple and flexible infrastructures and services that can be used in diverse ways
- Start small: provide simple solutions that build up to solve more complex problems
- Collaborate now more than ever!
UC3’s Service Landscape

UC Curation Center
http://www.cdlib.org/uc3
uc3@ucop.edu

Stephen Abrams
Patricia Cruse  Joan Starr
Scott Fisher  Tracy Seneca
Erik Hetzner  Carly Strasser
Greg Janée  Marisa Strong
John Kunze  Adrian Turner
David Loy  Perry Willett
Mark Reyes