Data management: The new frontier for libraries

Kathleen Shearer
What is driving this trend?

1. Verification, reproducibility and transparency of scientific results
2. New scientific discoveries through re-use and integration of datasets
3. Greater social and economic benefits through application of research outputs
4. And because we can...

Figure ES.1 Average data storage cost for consumers
1998-2012, per Gbit

BIG DATA!

90% of world's data has been generated over last two years

http://www.sciencedaily.com/releases/2013/05/130522085217.htm
The Poster Child...

- Sequencing centres agreed to make results available the same day
- (1990-2013) Finished coding the human genome 2 years ahead of time
- Contributes to treating hundreds of diseases
- Billions of dollars in economic spin-off
- Transformed the way biomedical research is conducted
Other examples of data sharing
Research data management is like a three-legged stool

- Policies
- Culture
- Infrastructure and Services
Policies and laws are important levers to move open science forward
Policies: moving towards openness...

Open science: a hot issue for OECD and non-OECD countries

Note: Simple counts do not account for the magnitude and impact of policy changes.
Source: Country responses to the STI Outlook policy questionnaire 2014.
Funders’ data policies

World Funders by Data Archiving Policy Type

- Required (42 = 27%)
- Encouraged (19 = 12%)
- (No Policy) (93 = 59%)

Total = 158 funders
## Typical Elements of a Data Policy

<table>
<thead>
<tr>
<th>Policy requirements</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data quality and standards</strong></td>
<td>Investigators are required to adhere to international standards to enable access and reuse in the discipline.</td>
</tr>
<tr>
<td></td>
<td>Data documentation and metadata must accompany data so that the data is understandable by others.</td>
</tr>
<tr>
<td><strong>Data access and sharing</strong></td>
<td>Investigators are required to make data available to be shared (usually upon publication of results or shortly thereafter, although some agencies do allow embargo periods).</td>
</tr>
<tr>
<td></td>
<td>Requirements for deposit of metadata into a local or national catalogue</td>
</tr>
<tr>
<td><strong>Data retention and preservation</strong></td>
<td>Data should be retained for a certain time limit, where possible, investigators must deposit their data in a long-term archive to ensure the preservation of their data.</td>
</tr>
<tr>
<td><strong>Data management plans</strong></td>
<td>Research proposals must include a Data Management Plan in proposal.</td>
</tr>
<tr>
<td>Common exceptions to policies</td>
<td></td>
</tr>
<tr>
<td>------------------------------</td>
<td></td>
</tr>
<tr>
<td><strong>Privacy</strong></td>
<td>The rights and privacy of individuals who participate in research must be protected at all times. Thus, data intended for broader use should be free of identifiers that would permit linkages to individual research participants and variables that could lead to deductive disclosure of the identity of individual subjects.</td>
</tr>
<tr>
<td><strong>Traditional knowledge</strong></td>
<td>Where local and traditional knowledge is concerned, rights of the knowledge holders shall not be compromised.</td>
</tr>
<tr>
<td><strong>Data of a sensitive nature</strong></td>
<td>Where data release may cause harm, specific aspects of the data may need to be kept protected (for example, locations of nests of endangered birds or locations of sacred sites).</td>
</tr>
<tr>
<td><strong>Intellectual property/Data ownership</strong></td>
<td>It may be necessary on occasion to delay publication for a short period to allow time for applications to be drafted.</td>
</tr>
</tbody>
</table>
Culture: RDM will require changes in researcher attitudes and practices.
Many researchers would rather share their toothbrush than there data...
Data sharing practices

<table>
<thead>
<tr>
<th>Country</th>
<th>Data Sharing Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>46% SHARING, 54% NOT SHARING</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>43% SHARING, 57% NOT SHARING</td>
</tr>
<tr>
<td>Japan</td>
<td>44% SHARING, 56% NOT SHARING</td>
</tr>
<tr>
<td>China</td>
<td>36% SHARING, 64% NOT SHARING</td>
</tr>
<tr>
<td>Brazil</td>
<td>52% SHARING, 48% NOT SHARING</td>
</tr>
<tr>
<td>Australia</td>
<td>41% SHARING, 59% NOT SHARING</td>
</tr>
<tr>
<td>Germany</td>
<td>55% SHARING, 45% NOT SHARING</td>
</tr>
</tbody>
</table>

**DATA SHARING TRENDS BY COUNTRY**

**UNITED STATES**
Among researchers in the US sharing their data publicly, two out of three do so because it is standard practice in the communities and because they believe it benefits the public. Similar to their counterparts in the UK, the majority of US-based researchers also share data to increase the impact or visibility of their research.

**UNITED KINGDOM**
While more than 40% of UK researchers are sharing data, only about 14% are using discipline-specific or other public repositories like Dryad and figshare. The two key drivers that motivate UK researchers to share their data are the prospect of gaining increased impact or visibility for their work and to satisfy funder requirements.

**JAPAN**
Compared with their counterparts around the world, researchers in Japan cite concerns about being scooped as a reason for not sharing data more frequently. Nearly five out of ten Japanese researchers point to this as a reason for not sharing their data, roughly double the global average.

**CHINA**
Nearly five in ten Chinese researchers say they are not sharing data because they are not required to do so by their funders or institutions. They are more likely than their global counterparts to say that they do not see data sharing as a personal responsibility and plan to take direction from funders to guide their data sharing decisions in the future.

**BRAZIL**
Two out of three researchers in Brazil say that a guarantee of proper credit or attribution would compel them to share more of their data publicly in the future.

**AUSTRALIA**
Researchers in Australia say they would be most incentivized to make their data accessible in the future to ensure preservation as well as transparency and re-use. The majority of researchers also ranked funder requirements among top reasons to share in the future.

**GERMANY**
Among German researchers sharing their data publicly, three out of four are driven to share data because they believe it will increase the visibility of their research and want to ensure public transparency and re-use. About 20% of German researchers are making use of general purpose repositories (like figshare and Dryad), significantly more than their counterparts around the world, including those in the US and UK.

From Wiley’s Research Data Insights Survey, 2014
2,250 responses from around the world
Both policy requirements and incentives are critical for cultural change.

From the cover of “Cultural Change Through Measurable Management by Robin Byrne
These services still only support a small portion of the research datasets produced by researchers around the world!
The data landscape

The 2011 survey by *Science*, found that 48.3% of respondents were working with datasets that were less than 1GB in size and over half of those polled store their data only in their laboratories. *Science* 11 February 2011: Vol. 331 no. 6018 pp. 692-693 *DOI*: 10.1126/science.331.6018.692
Libraries can close this huge gap in RDM
Challenges of managing research data

• Data is very diverse and can be complex
• Need new tools and technologies earlier on in the lifecycle
• Privacy issues: not all data can be open by default
• Domain expertise and thematic collections have more value
• Must work closely with the data creator and users
• Good metadata is essential!
Data sharing and re-use requires active data management across the lifecycle

The Research Data Management Lifecycle

URL: http://guides.library.ucsc.edu/datamanagement
## (potential) library services in RDM

<table>
<thead>
<tr>
<th>Stage</th>
<th>Services</th>
<th>Other stakeholders</th>
</tr>
</thead>
<tbody>
<tr>
<td>DMP Planning</td>
<td>• Assistance with data management planning</td>
<td>• Researchers</td>
</tr>
<tr>
<td></td>
<td>• Training graduate students and researchers about data management</td>
<td>• IT department (if DMP Online)</td>
</tr>
<tr>
<td></td>
<td>• IT department (if DMP Online)</td>
<td>• Administrators</td>
</tr>
<tr>
<td>Collection/prod uction of data</td>
<td>• Help with organizing data, adopting standards and ensuring comprehensive metadata</td>
<td>• Researchers</td>
</tr>
<tr>
<td></td>
<td>• Local storage</td>
<td>• IT department (if big data)</td>
</tr>
<tr>
<td>Analysis</td>
<td>• Tools for active data management (e.g. Open Science Framework or <a href="http://www.eric-project.org">http://www.eric-project.org</a>)</td>
<td>• Researchers</td>
</tr>
<tr>
<td></td>
<td>• IT department (if big data)</td>
<td>• IT department (if big data)</td>
</tr>
<tr>
<td>Data dissemination and sharing</td>
<td>• Managing a data repository</td>
<td>• Researchers</td>
</tr>
<tr>
<td></td>
<td>• Minting/assigning DOIs</td>
<td>• IT department</td>
</tr>
<tr>
<td></td>
<td>• Indexing and registering datasets</td>
<td>• Administrators</td>
</tr>
<tr>
<td></td>
<td>• Licensing and privacy issues</td>
<td>• Ethics boards</td>
</tr>
<tr>
<td>Re-use</td>
<td>• Data discovery</td>
<td>• Researchers</td>
</tr>
</tbody>
</table>
RDM requires working with many stakeholders

- Individual researchers
- Domain communities
- Policy makers
- IT and compute facilities
- University administrators
- Ethics boards
- Publishers
- Libraries
Coherent collection development policies

- Collecting data sets related to publications
- Supporting the adherence to specific funders data policies
- Building domain collections
- Anything produced at the institution

Should every research library really have a data repository?

(image credit: tjinternational.ltd.uk)
The role of research library associations

- Organizing shared services and infrastructure
- Defining RDM service models
- Identifying competencies and skills (e.g. Librarians’ Competencies for E-Research and Scholarly Communication)
- Promoting the role of the library in RDM with other stakeholders
Thank you

Kathleen Shearer, Executive Director, COAR
m.kathleen.shearer@gmail.com
@kathleeshearer