Who Pays? Implications of Value In Research Data Sustainability

Myron Gutmann, University of Colorado Boulder
Francine Berman, Rensselaer Polytechnic Institute
Jeremy York, University of Colorado Boulder
Organization of the Session

- Describe our research project, findings so far, conclusions drawn
- Discussion of implications and future developments
- Panelists:
  - Donald York, Founding Director, Sloan Digital Sky Survey
  - James Hilton, University Librarian, Dean of Libraries, & Vice Provost for Academic Innovation at the University of Michigan
  - Amy Walton, Program Director, CISE, National Science Foundation
Stewardship Gap Problem

- **Research data → innovation.**
  - Research increasingly expected to be available to the broader research community and general public now and in the future.

- **Preservation and stewardship of research data often ad hoc with much of it at risk**
  - How much is sustainable?
  - What data is at risk?
  - What should we do about it?

- **Lack of understanding about the sustainable stewardship gap hampers evidence-based discussion, prioritization and potential strategic investments.**
Is there a Stewardship Gap?

- **NIH estimates** for 2011 PubMed Central publications:
  - 12% of publication data sets deposited in recognized repositories, 88% of the data sets were invisible
  - Estimated approximately **200,000-235,000 invisible data sets** generated NIH work published in 2011
  - 87% of the invisible are new, 13% reflect data re-use
  - More than 50% of the datasets based on live human/animal subjects

- **Lack of comprehensive understanding about the broader sustainable stewardship gap hampers evidence-based discussion, prioritization and potential strategic investments.**
How would knowing the size and nature of the Stewardship Gap help?

“Funders, and particularly public funders, are under great pressure to show how their funding contributes to broad economic growth, how it addresses the needs of society, and to demonstrate that the requirements that they impose on the work they fund makes discovery ever more rapid, extensive, and cost-effective. From this perspective, they are not interested in data preservation or even data sharing other than as a necessary precondition to data reuse; they are interested in conformance to their data management and sharing policies because it is the only way they can create the preconditions for data reuse. They are hungry for examples of how data reuse has improved the processes of scholarship and discovery, or contributed to economic growth, job creation, control of health care costs, or public policy.”

The Stewardship Gap Project

- Understand the gap between valuable digital data and the amount responsibly stewarded
- Address the question: “So what if there is a stewardship gap?”

Who’s Involved? [Planning Group]
- Myron Gutmann, U. of Colorado (PI, co-lead)
- Fran Berman, RPI (co-lead)
- Jeremy York (Project Manager)
- George Alter, ICPSR
- Chris Borgman, UCLA
- Phil Bourne, NIH
- Vint Cerf, Google
- Sayeed Choudhury, Johns Hopkins University
- Elizabeth Cohen, Stanford University
- Trisha Cruse, DataONE
- Peter Fox, RPI
- John Gantz, IDC
- Margaret Hedstrom, U. of Michigan
- Brian Lavoie, OCLC
- Cliff Lynch, CNI
- Andy Maltz, Science and Technology Council, Academy of Motion Picture Arts and Sciences
- Guha Ramanathan, Google
Not One Gap But Many

- Many kinds of gaps
- Different gaps require different measurements
- Need to connect future policy and strategies--investment and otherwise--to the measurable gaps

Method
- Read Literature: The Stewardship literature identifies many kinds of gaps, which we explore in this research
- Interview members of the community to learn what’s being done and how they perceive the stewardship of their data.
Six Stewardship Gaps

- **Culture**: Gaps arising from differences in community attitudes, norms, and goals that affect data stewardship.

- **Knowledge**: Gap between the knowledge needed to effectively steward data and what is currently known.

- **Responsibility**: Gap between who has responsibility for stewardship and who is best placed to steward data over time.

- **Commitment**: Gap between the commitments that exist for valuable data and those necessary to ensure long-term stewardship.

- **Resources**: Gap between the people, money, infrastructure, and tools needed to steward data and what is now available.

- **Actions**: Gap between the actions taken to facilitate stewardship of data and the actions needed.
Six Stewardship Gaps

- Gaps arising from differences in community attitudes norms and goals that affect data stewardship
- Gap between the knowledge needed to effectively steward data, and what is currently known
- Gap between who has responsibility for stewardship and who is best placed to steward data over time
- Gap between the commitments that exist for valuable data and those necessary to ensure long-term stewardship
- Gap between the people, money, infrastructure, and tools needed to steward data, and what is now available
- Gap between the actions taken to facilitate stewardship of data and the actions needed
The Critical Importance of Value

- Value is an overarching theme
- Articulated or not, the value of data should determine the extent of stewardship
- Value is measured multiple ways, to the original researcher and others, in one field of study as opposed to others, now and in the future
- The hardest question to answer is the tradeoff between value and investment. **What value of data is worth what amount of stewardship investment?**
Researcher Agreement with Type of Value

- **Reuse outside immediate community**: 34 Agree, 11 Neutral, 49 Disagree
- **Timeless (will never lose value)**: 47 Agree, 9 Neutral, 56 Disagree
- **Longitudinal value**: 40 Agree, 12 Neutral, 25 Disagree
- **Reuse in immediate community**: 52 Agree, 16 Neutral, 26 Disagree
- **Inclusion in Reference collection**: 64 Agree, 16 Neutral, 27 Disagree
- **Current or Potential Impact**: 73 Agree, 15 Neutral, 25 Disagree
- **Data organization**: 80 Agree, 11 Neutral, 18 Disagree
- **Difficult to recreate**: 91 Agree, 4 Neutral, 18 Disagree
- **Own Research**: 107 Agree, 5 Neutral, 1 Disagree
Reasons for Value with Greatest Impact on Preservation Decisions

#1: First Choice
#2: Second Choice
#3: Third Choice

- Reuse by others
- Own research
- Potential reuse
- Difficult to Re-create
- Longitudinal Value
- Accountability
- Current or potential impact
- Uniqueness
- Mission
Researchers want to keep data for a long time, but the desire is not matched by commitment

- 95 out of 120 of datasets (79%) have an intention to preserve
- For 85 of these (71%), the intention is 10+ years
- 4 of 89 10+ year datasets (5%) have a commitment

Do intentions translate into preserved data?
## Term of Commitment or Intention and Term of Value

<table>
<thead>
<tr>
<th>Term of Commitment or Intention</th>
<th>Term of Value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Indefinite</td>
</tr>
<tr>
<td>Indefinite</td>
<td>31</td>
</tr>
<tr>
<td>&gt; 10 years</td>
<td></td>
</tr>
<tr>
<td>&lt;= 10 years</td>
<td>10</td>
</tr>
<tr>
<td>&lt;= 5 years</td>
<td>3</td>
</tr>
<tr>
<td>Undetermined</td>
<td>3</td>
</tr>
</tbody>
</table>
### But How Much Commitment Is There?

<table>
<thead>
<tr>
<th>Term of Commitment or Intention</th>
<th>Term of Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indefinite</td>
<td>&lt;= 10 years</td>
</tr>
<tr>
<td>Indefinite</td>
<td>2</td>
</tr>
<tr>
<td>&gt; 10 years</td>
<td>1</td>
</tr>
<tr>
<td>&lt;= 10 years</td>
<td>1</td>
</tr>
<tr>
<td>&lt;= 5 years</td>
<td>2</td>
</tr>
<tr>
<td>Undetermined</td>
<td></td>
</tr>
</tbody>
</table>

**Term of Value:**
- Indefinite
- <= 10 years
- <= 5 years
- Undetermined
- Unsure
Type of Value with Greatest Impact on Preservation Decisions

**Reuse by others** was most often cited as having an impact on preservation decisions.

Where $\text{Term CI} = \text{Term V}$, the most common types of value are:

1. Difficult to re-create
2. Longitudinal
3. Own research
4. Uniqueness
Type of Value with Greatest Impact on Preservation Decisions

Where \( \text{Term CI} > \text{Term V} \), the most common reasons for value are

- Good scholarly practice
- Own research
- Potential reuse
- Difficult to re-create

Datasets did not have value due to

- Uniqueness
Type of Value with Greatest Impact on Preservation Decisions

Where Term CI < Term V, the most common types of value were
• Longitudinal
• Own research
• Potential reuse
• Accountability

There was no value due to
• Difficult to re-create
• Uniqueness of data
### Type of Value with Greatest Impact on Preservation Decisions

<table>
<thead>
<tr>
<th></th>
<th>Term CI = Term V</th>
<th>Term CI &gt; Term V</th>
<th>Term CI &lt; Term V</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reuse</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Difficult to re-create</td>
<td></td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Longitudinal</td>
<td>2</td>
<td>×</td>
<td>1</td>
</tr>
<tr>
<td>Own research</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Uniqueness</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Potential reuse</td>
<td>×</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Accountability</td>
<td>×</td>
<td>×</td>
<td>4</td>
</tr>
<tr>
<td>Good scholarly practice</td>
<td>×</td>
<td>1</td>
<td>×</td>
</tr>
<tr>
<td>Impact</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Mission</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
</tbody>
</table>
Questions for Discussion

- What role do perceptions of value play in decisions about funding the production, management, and care of research data?
- Are there types of value for which data stewardship investments should be prioritized?
- What barriers exist to identifying data value and what strategies or interventions could provide insight into the value data may hold?
- What implications might the ability to identify types of value have for who should have financial and management responsibilities for data stewardship?
- What is the state of the art of policy and practice, and what different policies and practices would lead to more sustainability for valued research data?