The Stewardship Gap

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

http://bit.ly/stewardshipgap

Coalition for Networked Information
April 4-5, 2016
Austin, Texas

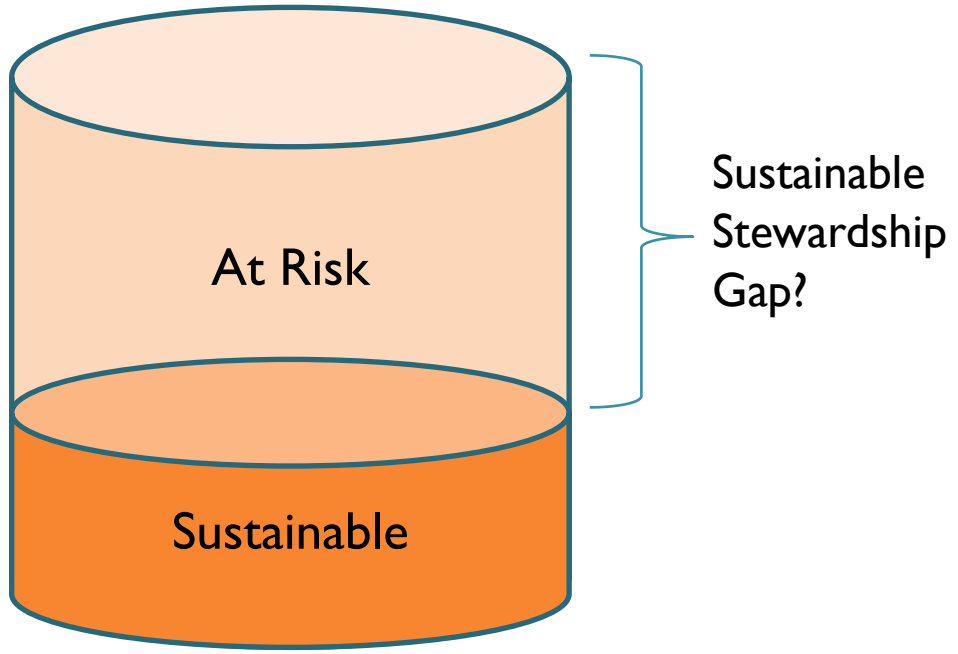
Rensselaer

University of Colorado
Boulder

INTRODUCTION

Stewardship Gap Problem

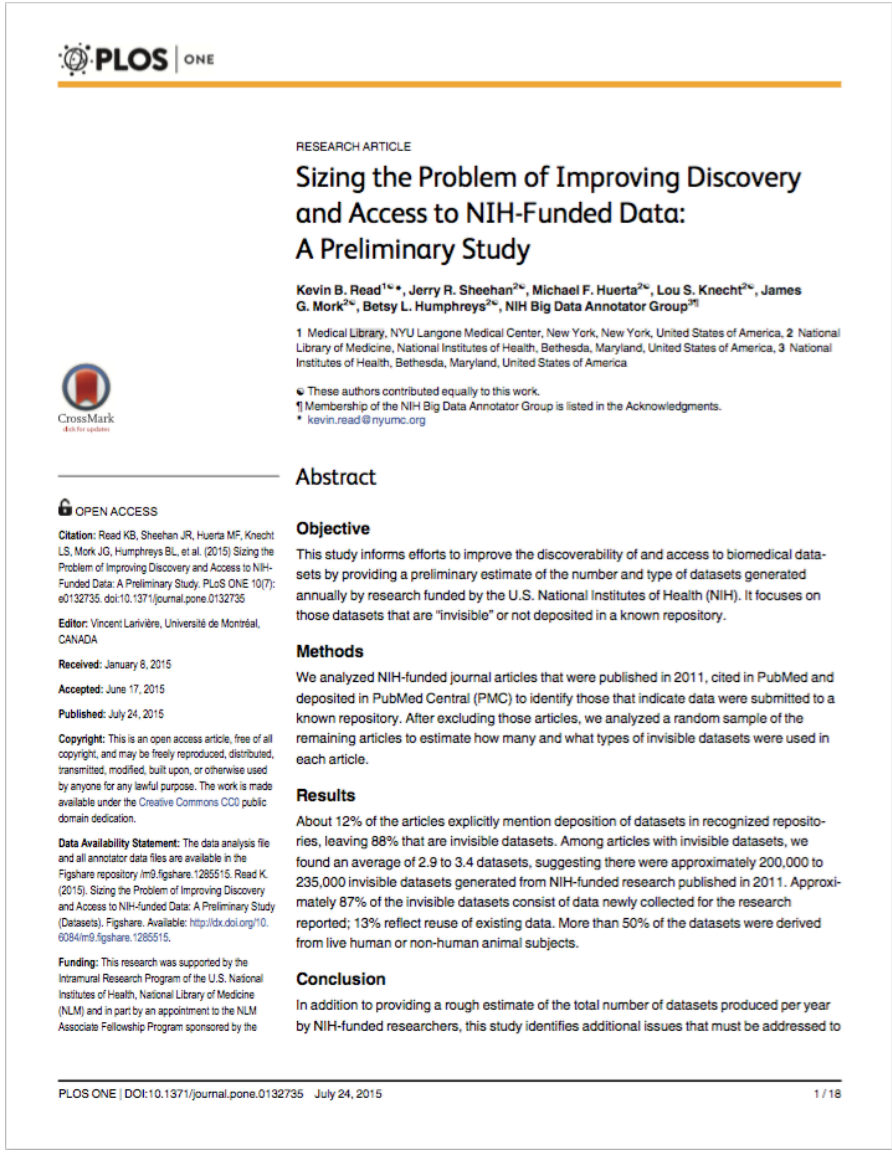
- **Research data → innovation.**
  - Research increasingly expected to be available to the broader research community and general public *now* and *in the future*.
- **Preservation and stewardship of research data often ad hoc with much of it at risk**
  - *How much is sustainable?*
  - *What data is 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.**



(Valuable) Sponsored Research Data

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.**



PLOS ONE

RESEARCH ARTICLE

Sizing the Problem of Improving Discovery and Access to NIH-Funded Data: A Preliminary Study

Kevin B. Read[1☯*], Jerry R. Sheehan[2☯], Michael F. Huerta[2☯], Lou S. Knecht[2☯], James G. Mork[2☯], Betsy L. Humphreys[2☯], NIH Big Data Annotator Group[3¶]

1 Medical Library, NYU Langone Medical Center, New York, New York, United States of America, 2 National Library of Medicine, National Institutes of Health, Bethesda, Maryland, United States of America, 3 National Institutes of Health, Bethesda, Maryland, United States of America

☯ These authors contributed equally to this work.
¶ Membership of the NIH Big Data Annotator Group is listed in the Acknowledgments.
* kevin.read@nyumc.org

OPEN ACCESS

Citation: Read KB, Sheehan JR, Huerta MF, Knecht LS, Mork JG, Humphreys BL, et al. (2015) Sizing the Problem of Improving Discovery and Access to NIH-Funded Data: A Preliminary Study. PLoS ONE 10(7): e0132735. doi:10.1371/journal.pone.0132735

Editor: Vincent Larivière, Université de Montréal, CANADA

Received: January 8, 2015

Accepted: June 17, 2015

Published: July 24, 2015

Copyright: This is an open access article, free of all copyright, and may be freely reproduced, distributed, transmitted, modified, built upon, or otherwise used by anyone for any lawful purpose. The work is made available under the Creative Commons CC0 public domain dedication.

Data Availability Statement: The data analysis file and all annotator data files are available in the Figshare repository /m9.figshare.1285515. Read K. (2015). Sizing the Problem of Improving Discovery and Access to NIH-funded Data: A Preliminary Study (Datasets). Figshare. Available: http://dx.doi.org/10.6084/m9.figshare.1285515.

Funding: This research was supported by the Intramural Research Program of the U.S. National Institutes of Health, National Library of Medicine (NLM) and in part by an appointment to the NLM Associate Fellowship Program sponsored by the

Abstract

Objective

This study informs efforts to improve the discoverability of and access to biomedical datasets by providing a preliminary estimate of the number and type of datasets generated annually by research funded by the U.S. National Institutes of Health (NIH). It focuses on those datasets that are "invisible" or not deposited in a known repository.

Methods

We analyzed NIH-funded journal articles that were published in 2011, cited in PubMed and deposited in PubMed Central (PMC) to identify those that indicate data were submitted to a known repository. After excluding those articles, we analyzed a random sample of the remaining articles to estimate how many and what types of invisible datasets were used in each article.

Results

About 12% of the articles explicitly mention deposition of datasets in recognized repositories, leaving 88% that are invisible datasets. Among articles with invisible datasets, we found an average of 2.9 to 3.4 datasets, suggesting there were approximately 200,000 to 235,000 invisible datasets generated from NIH-funded research published in 2011. Approximately 87% of the invisible datasets consist of data newly collected for the research reported; 13% reflect reuse of existing data. More than 50% of the datasets were derived from live human or non-human animal subjects.

Conclusion

In addition to providing a rough estimate of the total number of datasets produced per year by NIH-funded researchers, this study identifies additional issues that must be addressed to

PLOS ONE | DOI:10.1371/journal.pone.0132735 July 24, 2015 1 / 18

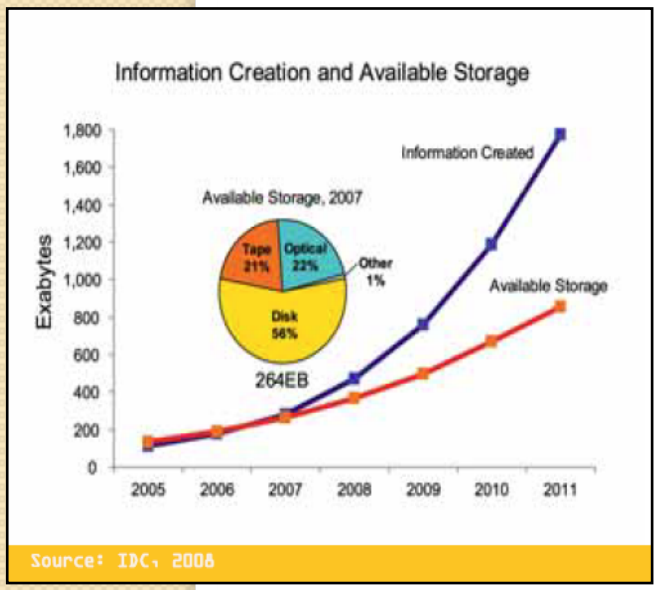
* From PLOS ONE
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0132735

How would knowing the size and nature of the Stewardship Gap help?

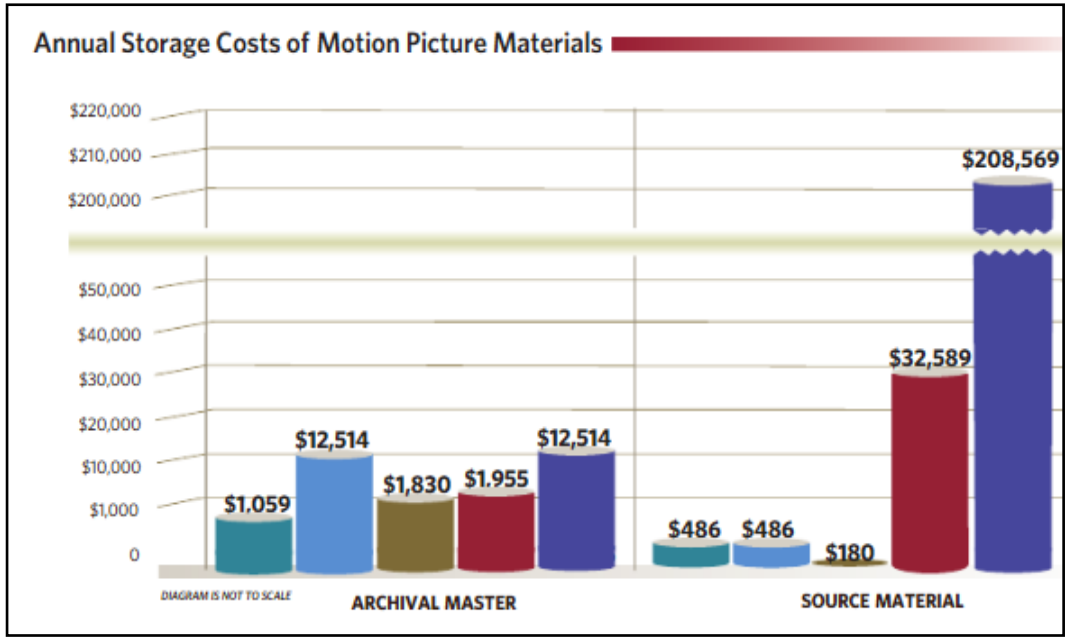
"Funders, and particularly public funders, are under great pressure to show how their funding contributes to broad economic growth, how it addresses the needs of society, and to demonstrate that the requirements that they impose on the work they fund makes discovery ever more rapid, extensive, and cost-effective.

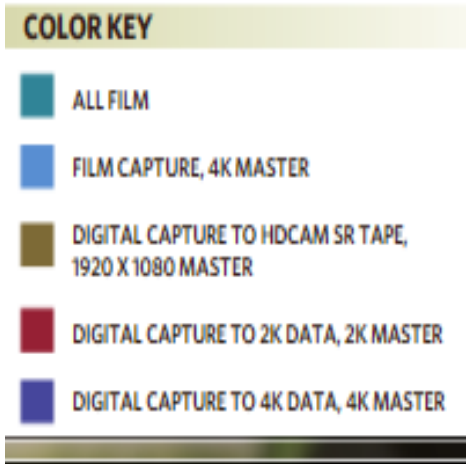
From this perspective, they are not interested in data preservation or even data sharing other than as a necessary precondition to data reuse; they are interested in conformance to their data management and sharing policies because it is the only way they can create the preconditions for data reuse. 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."

Clifford Lynch, The Next Generation of Challenges in the Curation of Scholarly Data," Research Data Management: Practical Strategies for Information Professionals, edited by Joyce M. Ray. West Lafayette, IN: Purdue University Press, 2013.



IDC reports on the Digital Universe, http://www.emc.com/leadership/digital-universe/index.htm#Archive





AMPAS report on the Digital Dilemma, http://www.scribd.com/doc/55498058/The-Digital-Dilemma

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

Specific Tasks

- Identify a ***sampling frame and strategic case studies***
- Develop a robust ***evaluation instrument***
- Produce a set of ***actionable recommendations*** and summary ***reports*** that can help guide strategic decisions about the stewardship gap

Not One Gap But Many

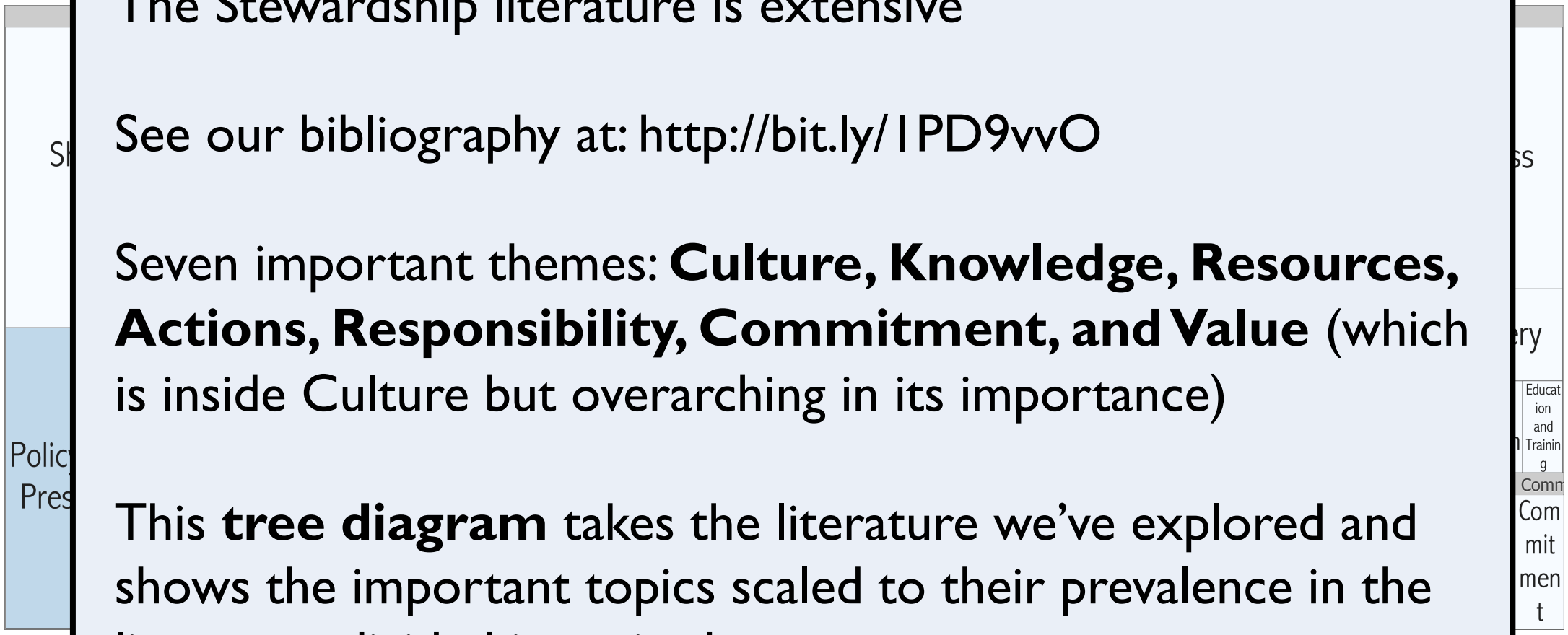
- Many kinds of gaps
- Different gaps require different measurements
- Need to connect future policy and strategies--investment and otherwise--to the measurable gaps
- Method
  - Read Literature: The Stewardship literature identifies many kinds of gaps, which we explore in this research
  - Interview members of the community to learn what's being done and how they perceive the stewardship of their data.

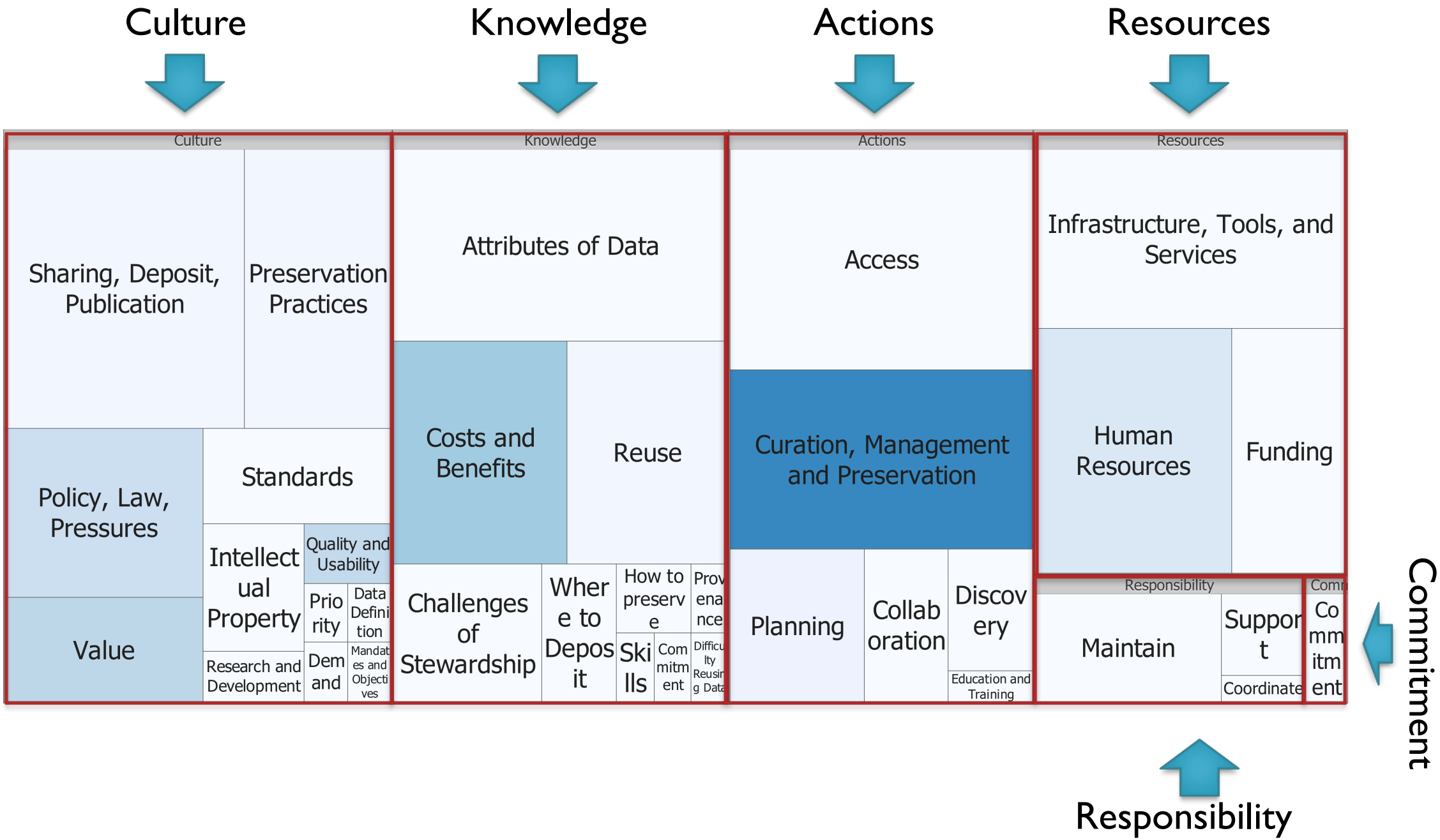
The Stewardship literature is extensive

See our bibliography at: http://bit.ly/1PD9vvO

Seven important themes: **Culture, Knowledge, Resources, Actions, Responsibility, Commitment, and Value** (which is inside Culture but overarching in its importance)

This **tree diagram** takes the literature we've explored and shows the important topics scaled to their prevalence in the literature, divided into six themes




Culture
Knowledge
Actions
Resources
Culture
Sharing, Deposit, Publication
Preservation Practices
Policy, Law, Pressures
Standards
Value
Intellectual Property
Quality and Usability
Priority
Data Definition
Research and Development
Demand
Mandates and Objectives
Knowledge
Attributes of Data
Costs and Benefits
Reuse
Challenges of Stewardship
Where to Deposit
How to preserve
Provenance
Skills
Commitment
Difficulty Reusing Data
Actions
Access
Curation, Management and Preservation
Planning
Collaboration
Discovery
Education and Training
Resources
Infrastructure, Tools, and Services
Human Resources
Funding
Responsibility
Maintain
Support
Coordinate
Commitment
Responsibility
Commitment

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

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?**

What to measure and how?

PHASE 1: PRELIMINARY INVESTIGATION

What to Measure

- Is there a gap?

What to Measure

- Is there a gap?
  - What is the value of data and for how long will they be valuable
  - What is the extent of stewardship commitment on data

What to Measure

- Is there a gap?
  - What is the value of data and for how long will they be valuable
  - What is the extent of stewardship commitment on data
- Who can act to address the gap?

What to Measure

- Is there a gap?
  - What is the value of data and for how long will they be valuable
  - What is the extent of stewardship commitment on data
- Who can act to address the gap?
- How much data and what kind is at risk?

What to Measure

- Scope of data interest
  - Data resulting from sponsored research or creative work in the US, whether publicly or privately funded (we have focused on research outputs, primarily federally-funded)
- Unit of Analysis: Project
  - A body of work that has a defined scope and resources and a distinct beginning and end (not necessarily a single grant)

How to Measure

- Interviews
- Whom to ask
  - Those responsible for project data
  - Principle Investigators, staff involved in data production and management

What to ask

Project Context	Purpose, domains of science, collaborators, funders, size and characteristics of data (Responsibility, Knowledge)
Commitment	For how much of the data is there 1) a commitment to preserve 2) an intention to preserve 3) no intention to preserve (no intention to delete) 4) the data are temporary (and will be deleted)
Stewardship	Who stewarding data, what is being done to take care of them, concerns about stewardship, prospects when current commitment has ended (Culture, Responsibility, Commitment, Resources, Actions)
Value	Why is the data valuable and for how long, how does the valuation affect stewardship decisions, worthwhile to reassess the value in the future? (Culture, Activities)

PROJECT CONTEXT

Respondents

- 17 Respondents in 16 disciplines from 13 institutions (31 contacts)
- Data Sets Ranged from tiny to 50 TB

Researcher Disciplines

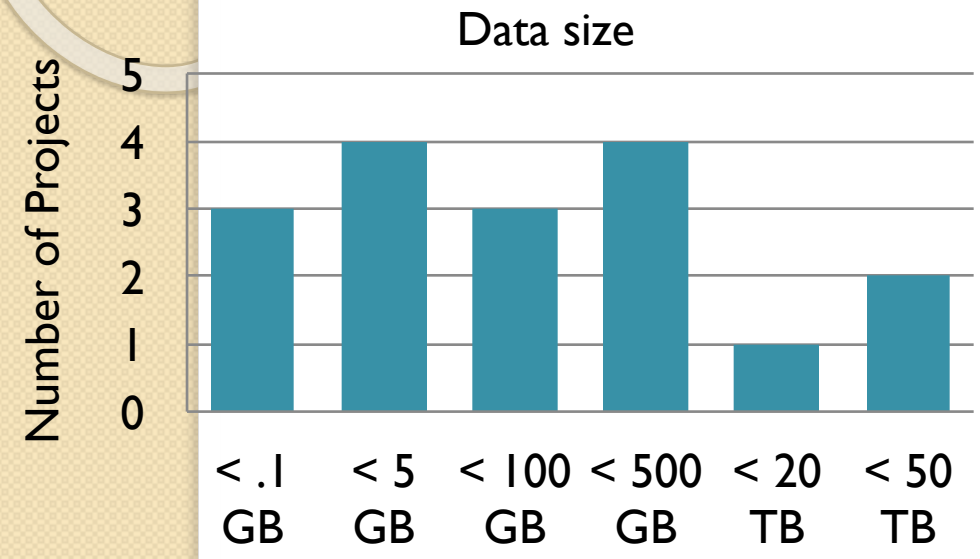
- Geography
- History
- Archaeology
- Economics
- Political science
- Psychology
- Public administration
- Information
- Education
- Environmental studies
- Physical performance & recreation
- Neuroscience
- Astronomy
- Computer sciences
- Physics
- Statistics

Respondents

- 17 Respondents in 16 disciplines from 13 institutions (31 contacts)
- Data Sets Ranged from tiny to 50 TB

Resulting data represent 32 domains of research

Data Description

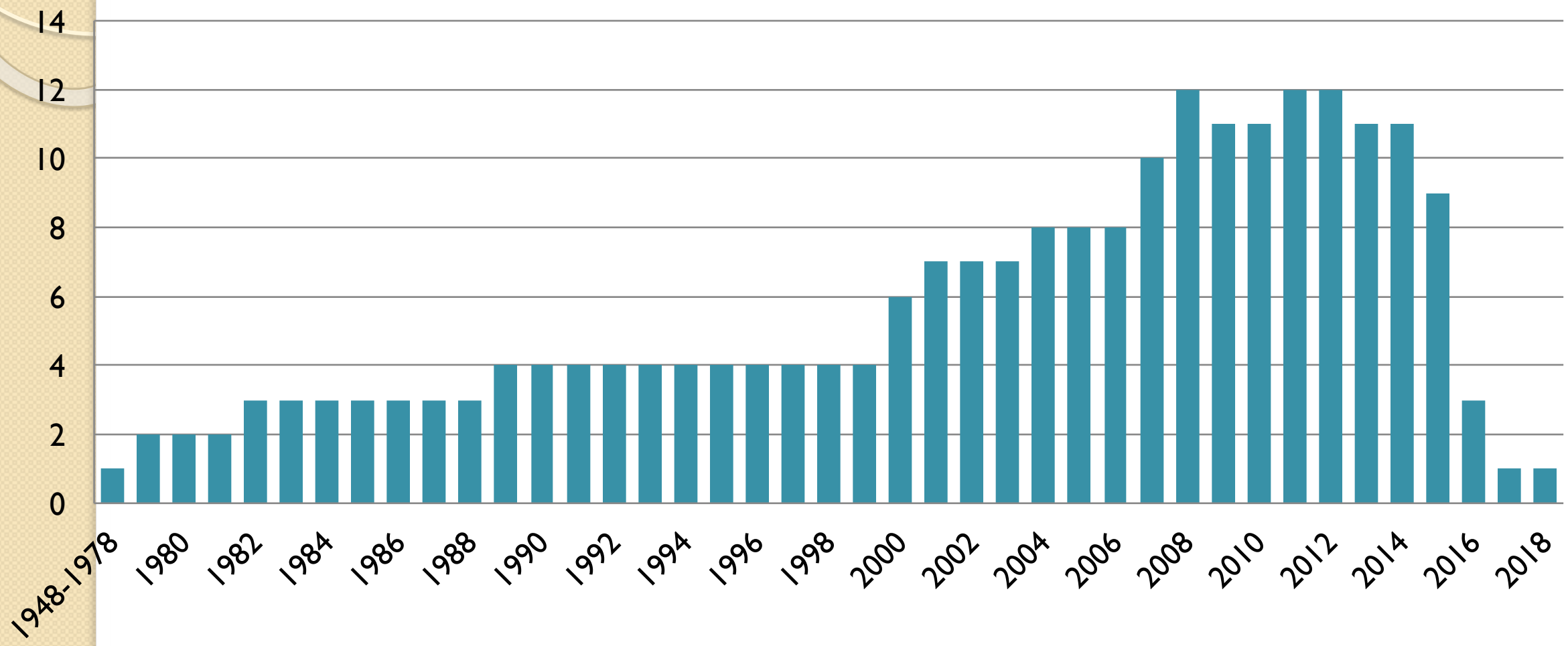


17 projects, 39 datasets

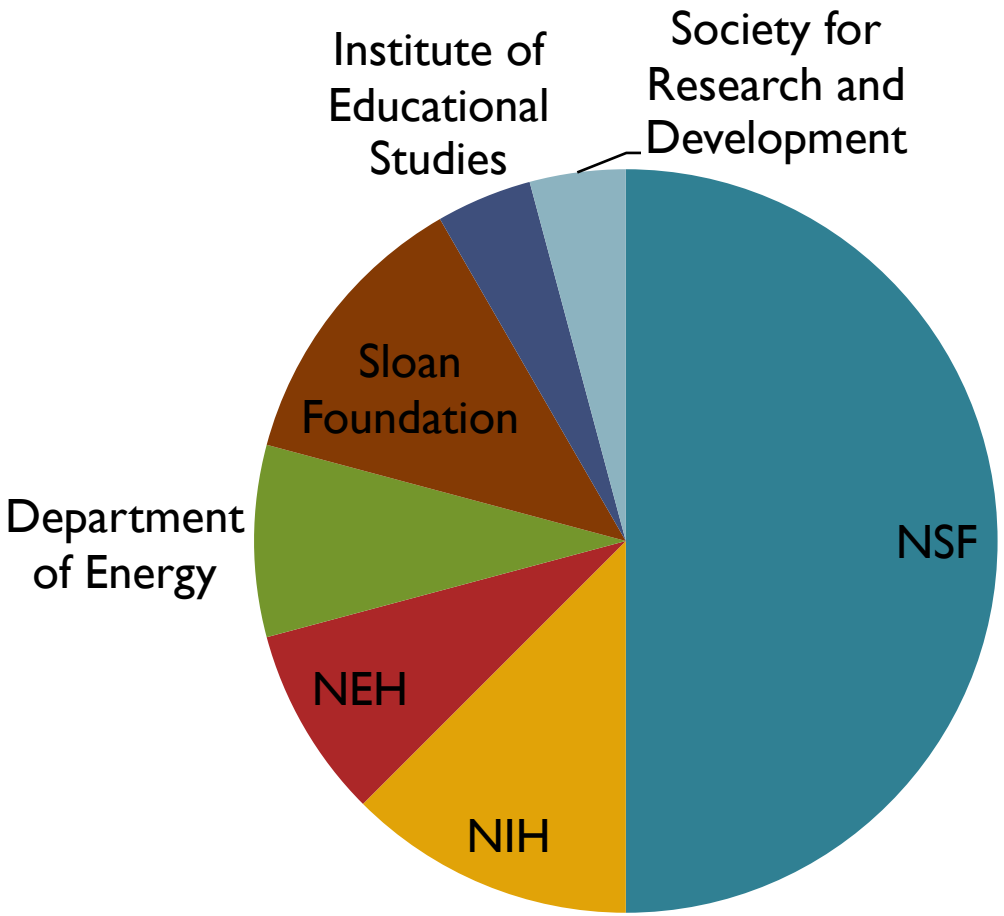
- Video, Audio, Text
- Digital image streams
- Data from interviews, questionnaires, surveys
- Chat files
- Field same of vegetation and soils
- Housing prices
- Simulation models of land use
- Voltage measurements
- Software
- Topic models
- Tag clouds
- Behavioral action logs
- GIS information
- Plant and animal diversity data
- Maps, on-site images
- Database graphs
- Service and configuration data
- Business transaction information

Project Years

Multi-year projects are represented in each project year



Project Funding

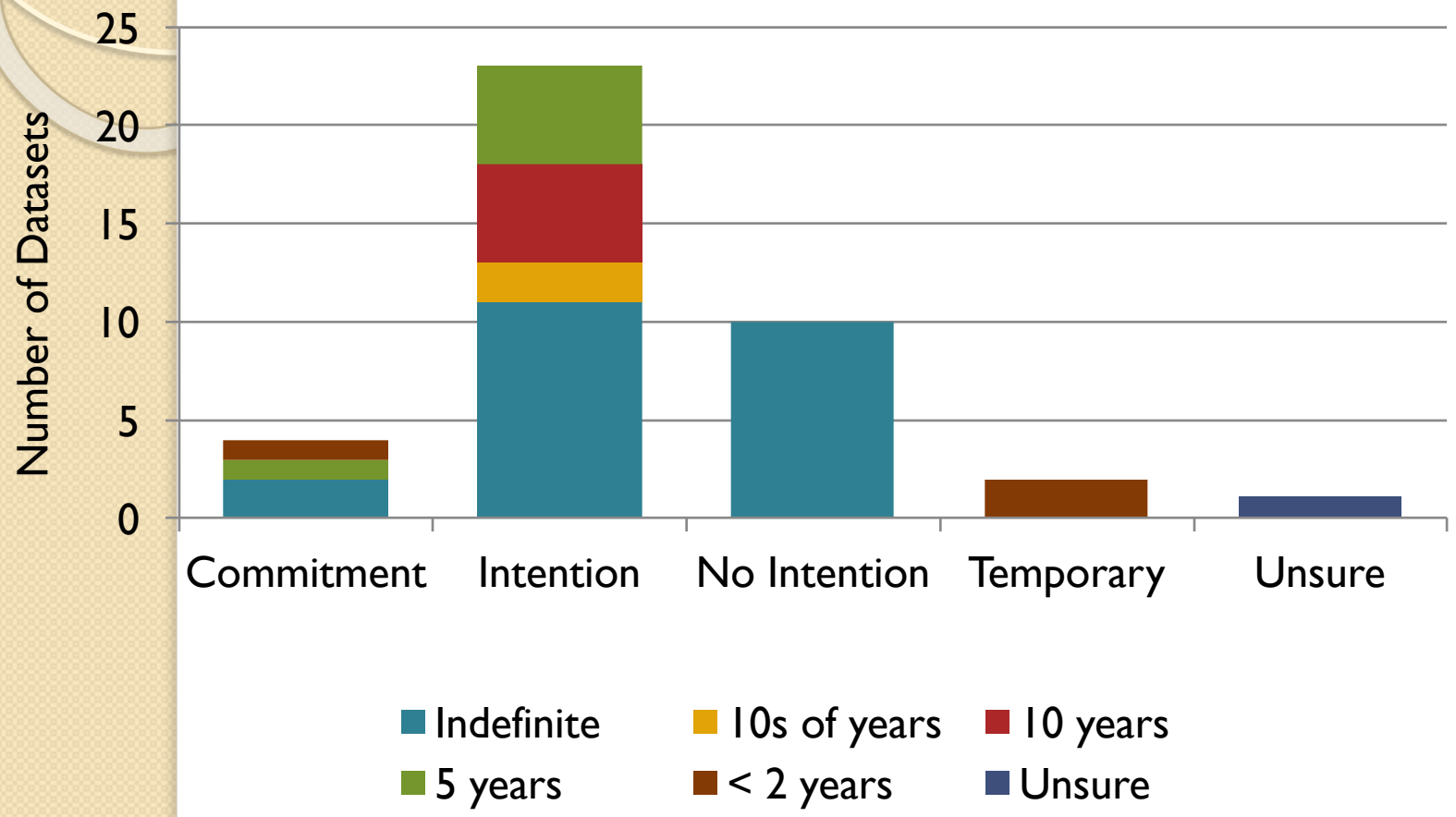


Limitations

- Small number of respondents, but observations are revelatory
- Weak on biological science and medicine
- Our next set of sample cases will add 50 more observations by late spring

COMMITMENT AND VALUE

Type of Commitment and Term of Commitment



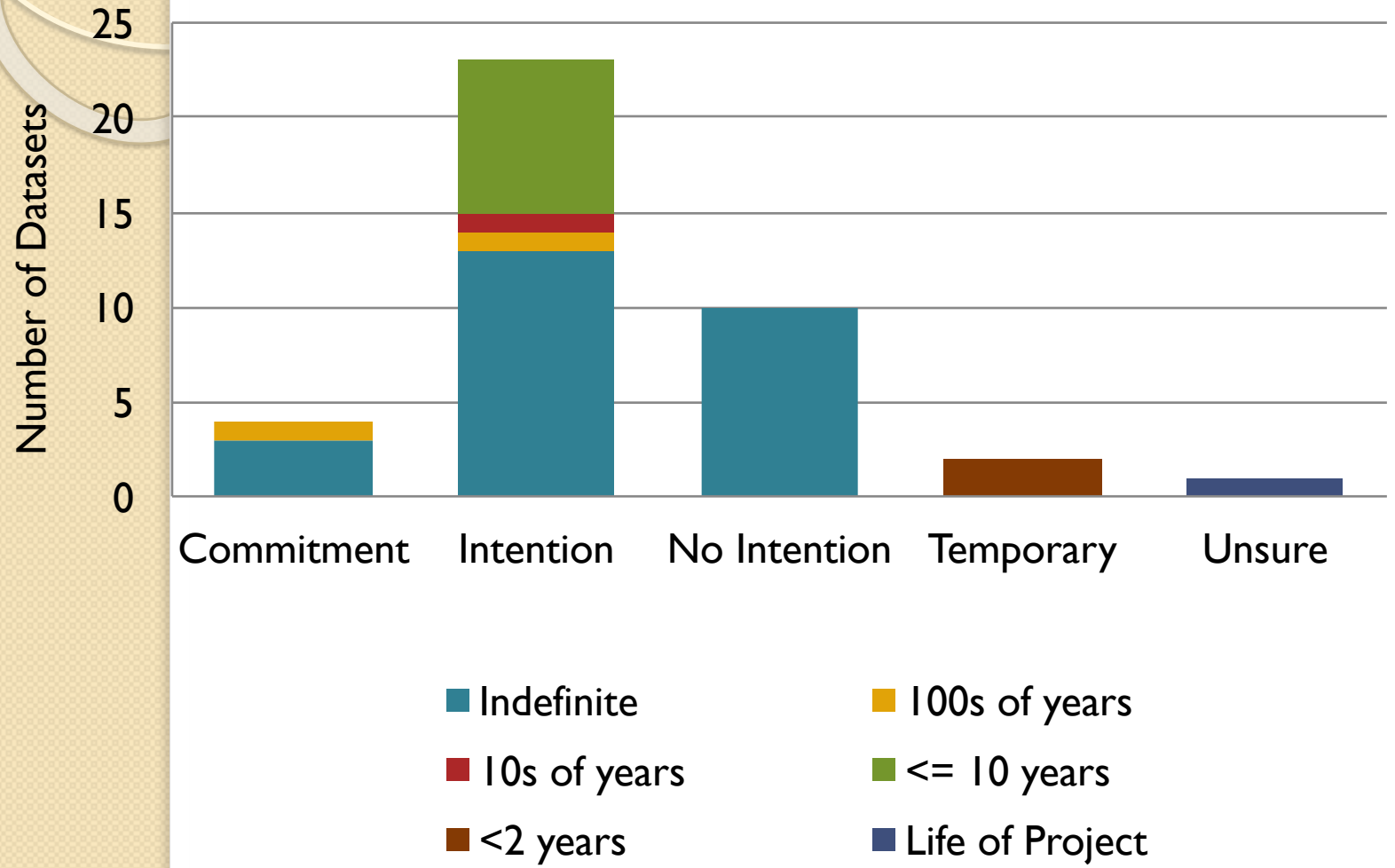
*One project reported two commitment levels on the same data

Researchers want to keep data for a long time, but the desire is not matched by commitment

- 3/5 of datasets have an intention to preserve
- For 3/4 of these, the intention is 10+ years
- 1/10 of 10+ yr datasets have commitment

Do intentions translate into preserved data?

Type of Commitment and Term of Value

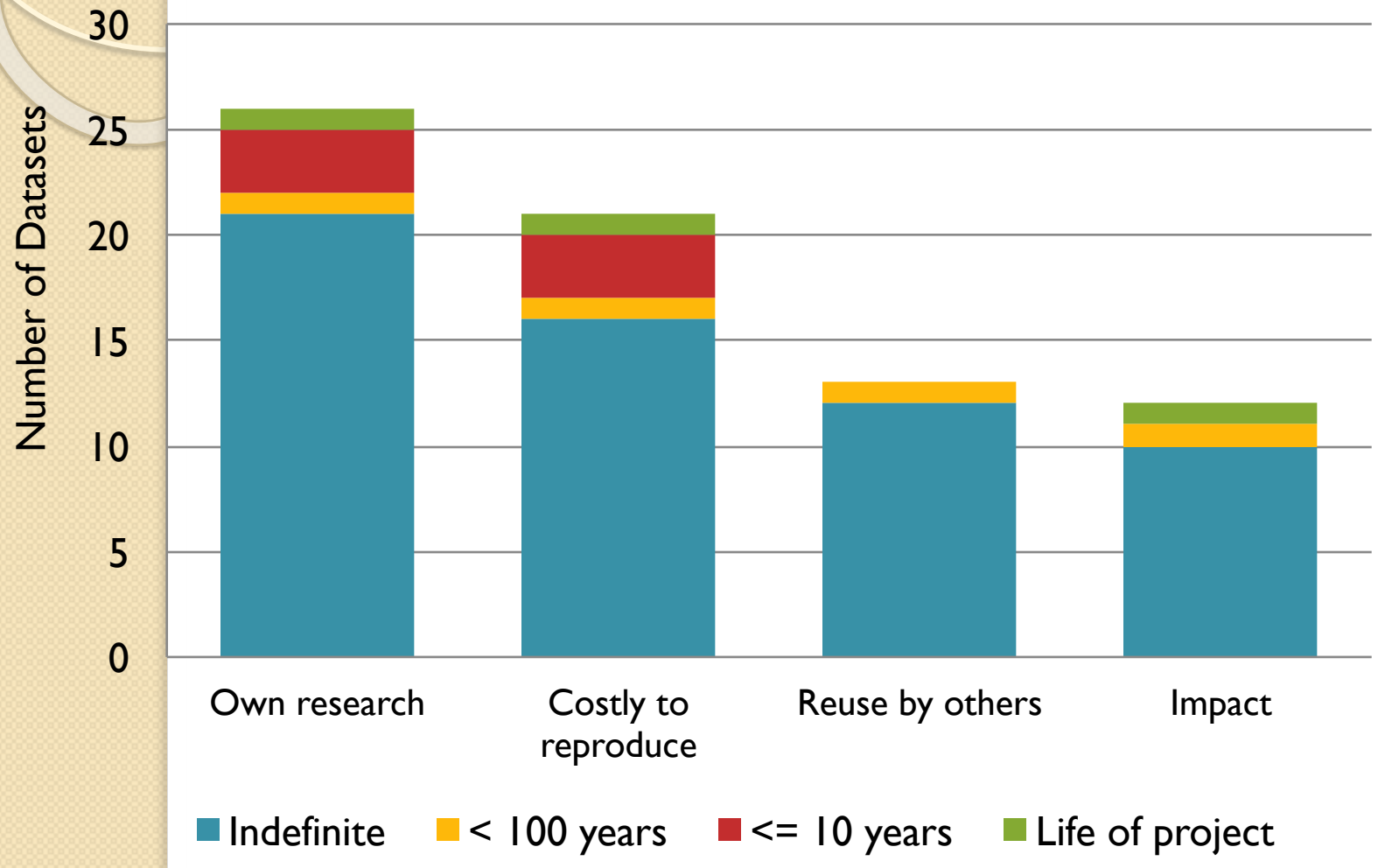


Researchers believe their data have long-term value.

For datasets with >10 years of value:

- 2 out of 34 have a matching commitment
- ~1/3 have no explicit intention to preserve

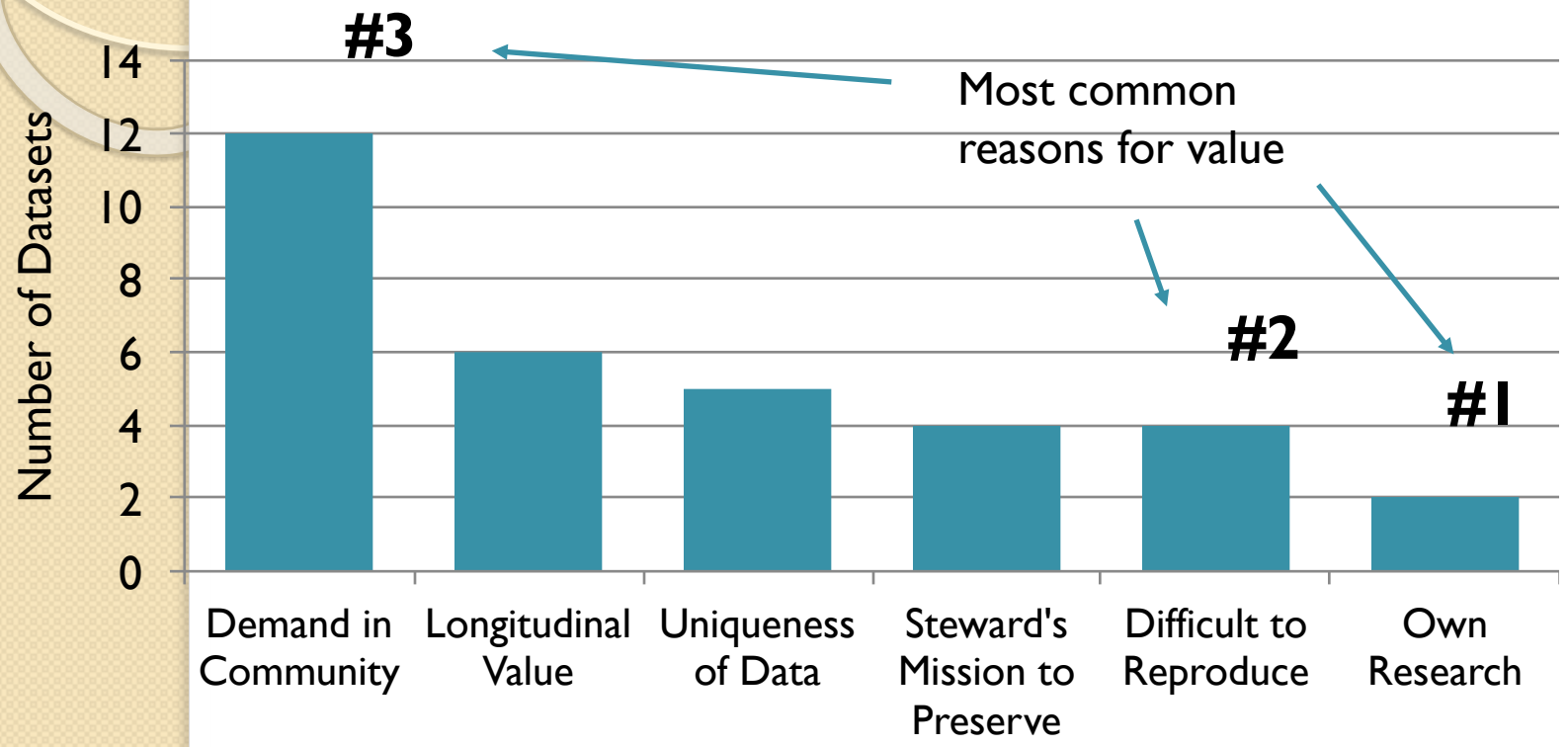
Type of Value, and Term of Value



Most common reasons for data value:

- Their own research use
- Data costly to reproduce
- Reuse by others
- Demonstrated or potential impact

Reasons for Value with Greatest Impact on Preservation Commitments

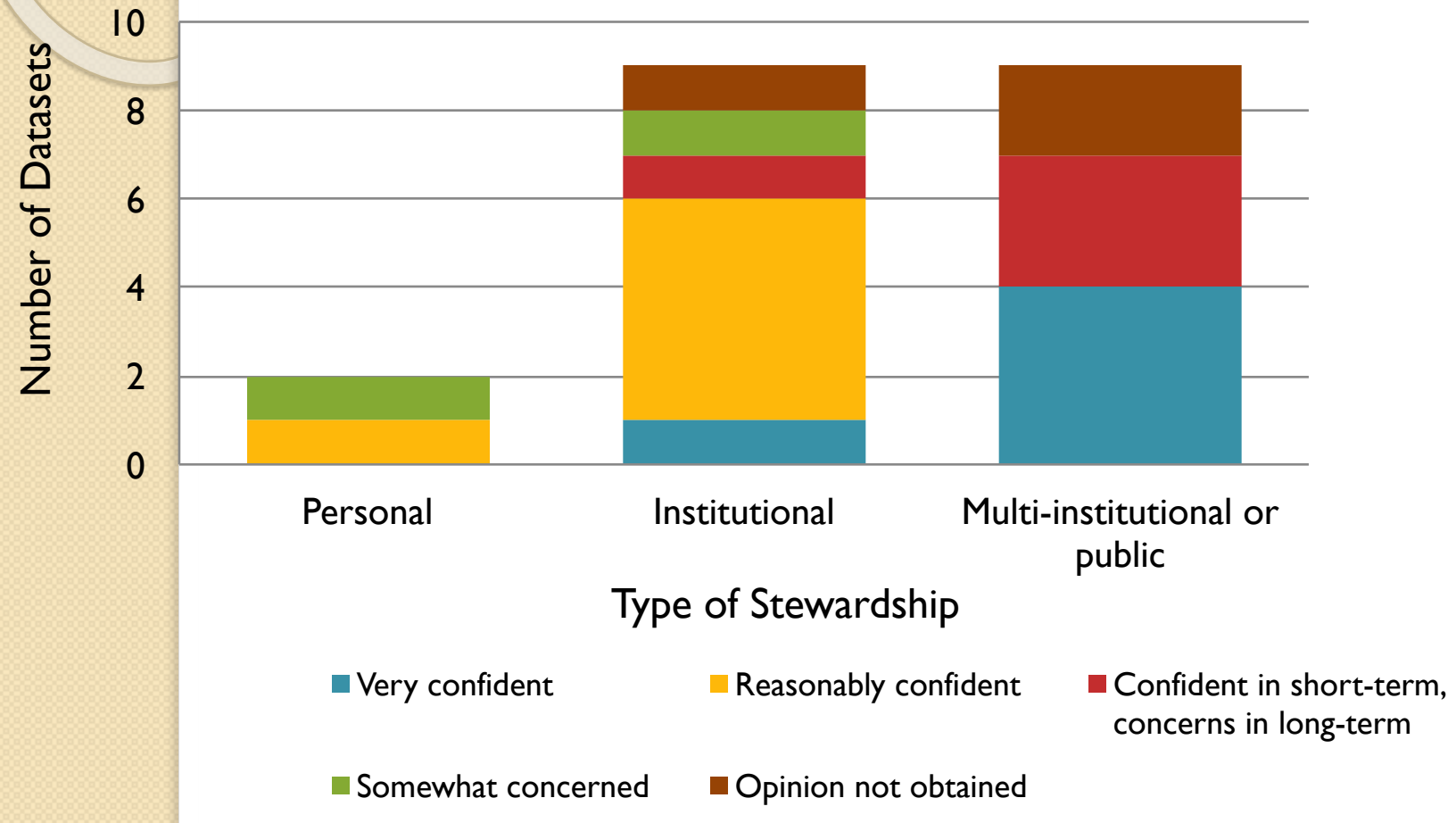


There is a mismatch between the value researchers believe their data to have and the value researchers believe drives preservation commitments

Some types of value had the greatest impact on preservation decisions:

- Community demand
- Unique data
- Data preservation mission
- Data hard to reproduce
- Value for the researcher's own work

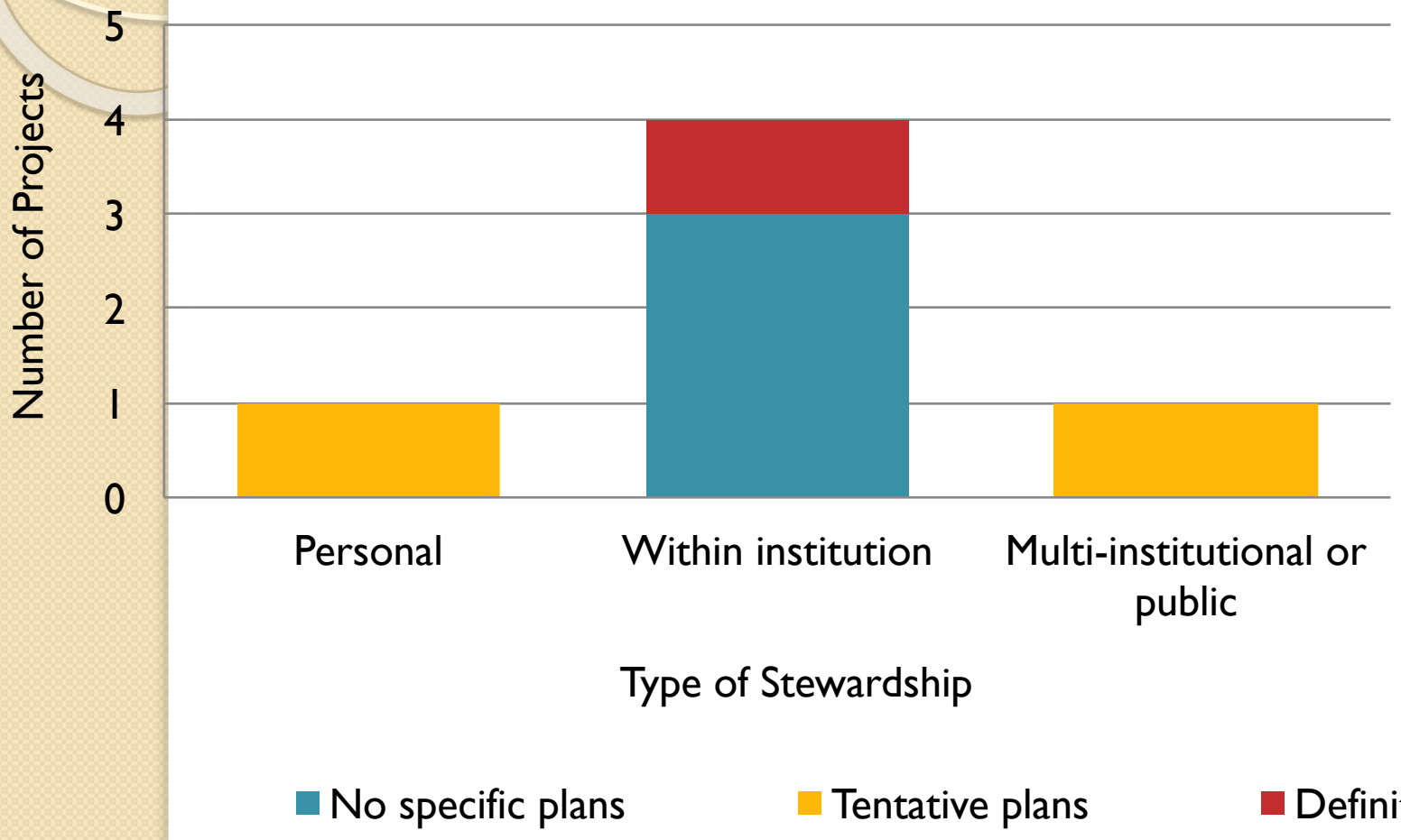
Confidence in Stewardship



In 13 out of 20 stewardship locations researchers felt very (5) or reasonably (8) confident in the ability of the data steward to fulfill the preservation commitment on the data

How well-founded is this confidence?

Prospects for stewardship when the existing commitment/intention is over



Few researchers had specific plans for stewardship; many assumed that their institution would take on that role.

Progress on Objectives (1)

1. To get a good sense of the "sponsored research data universe" by identifying a ***sampling frame and strategic case studies*** that provide an accurate and meaningful view of research data stewardship on a broader scale.

   → Working on in Phase 2

2. To assess the stewardship gap by developing a robust ***evaluation instrument***, flexible to multiple levels on which research data is created and maintained, and capable of **providing useful information** for data stewards, research administrators, and other stakeholders to underlie strategic decision-making about research data stewardship.

   → Developed in Phase 1 and refined for Phase 2

Progress on Objectives (2)

3. To produce a set of ***actionable recommendations*** and summary ***reports*** that can help guide strategic decisions about the stewardship gap, research data stewardship landscape, and needed efforts to ensure sustainable long-term access to valuable sponsored research data.

→ Pending

Next Steps

- 50 more interviews with a more structured sample in the next couple of months
- Added questions about
  - Are data collected to share or to test a specific hypothesis?
  - Use of secondary data (previously implicit)
  - Was the primary goal of transferring responsibility to share with others or to preserve data?
  - Expectations about stewardship of project data
- Make a decision about a future, more comprehensive study

What have we learned so far?

- There's a lot of diversity in research data stewardship, which makes our task challenging but exciting
- One of the challenges is a need to improve knowledge translation about data between researchers, data scientists, and data stewards
- Researchers want to have their data well stewarded, but don't always get the commitments that would ensure long-term stewardship

From Gaps to Policy: Possible Examples

Commitment

If researchers don't always get the commitments that would ensure long-term stewardship, find ways to give them and stewardship organizations incentives to do so

From Gaps to Policy: Possible Examples

Knowledge

Data management plans have a lot to teach us, but they need to be more informative and more readily available. Find ways to improve DMPs and make them useful for data science research

From Gaps to Policy: Possible Examples

Value

Researchers distinguish degrees and durations of data value for different purposes. Provide policy structures to use information about value to inform stewardship

Topics for Discussion

- What do we need to do to make this relevant with you? What additional information do we need for findings from our project to have policy implications
- What have we missed and what else should we be thinking about?
- How do the limits of our methodology (a small number of detailed interviews) affect our results and future work?

Stewardship Gap Bibliography: http://bit.ly/1PD9vvO

Tag cloud of bibliography topics Generated 1/10/2016 from https://www.jasondavies.com/

Jeremy.York@colorado.edu
Myron.Gutmann@colorado.edu