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

http://bit.ly/stewardshipgap

Rensselaer

University of Colorado
Boulder

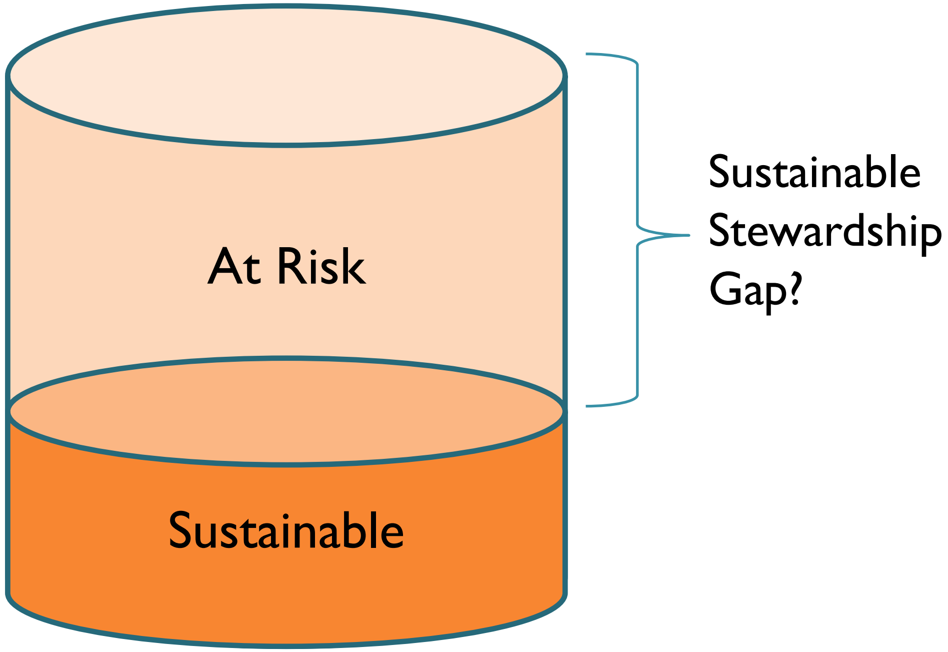
ALFRED P. SLOAN FOUNDATION
1934

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

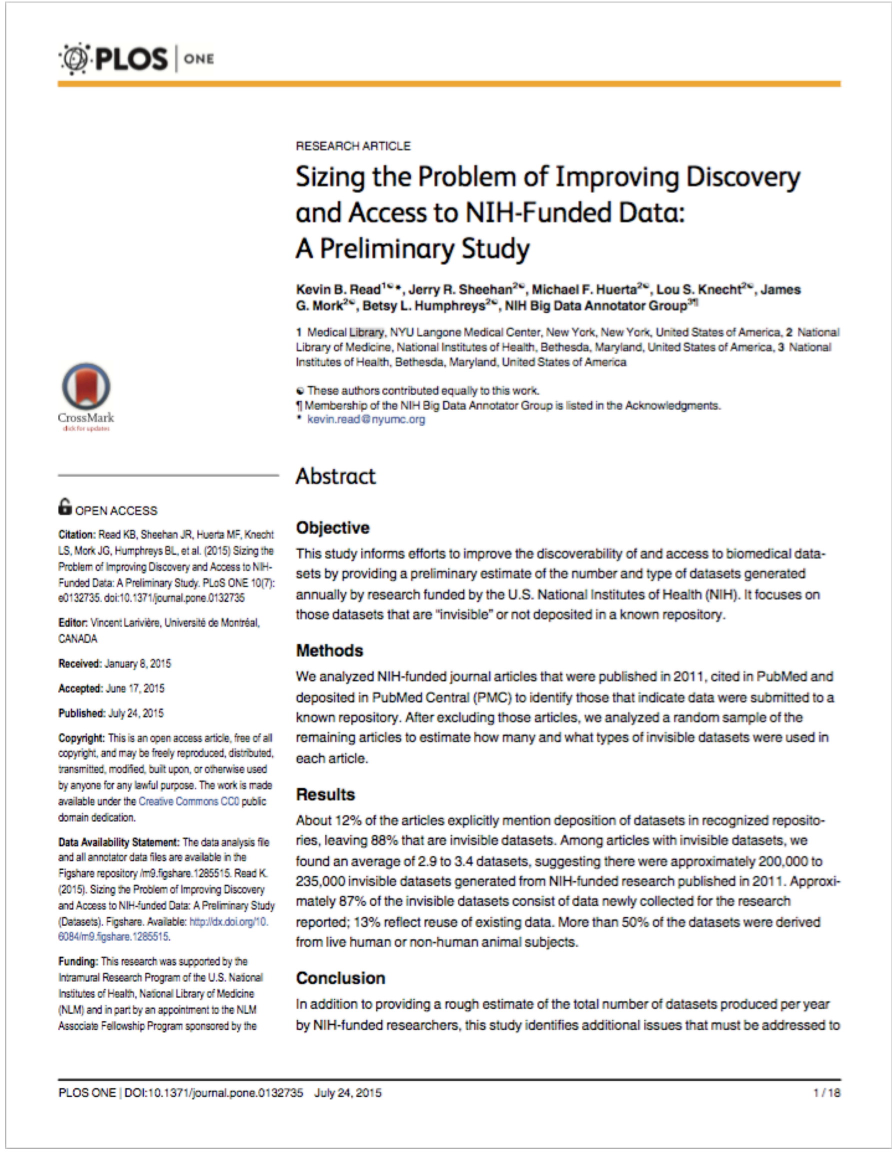
- **Research data → innovation.**
  - Research increasingly expected to be available to the broader research community and general public *now* and *in the future*.
- **Preservation and stewardship of research data often ad hoc with much of it at risk**
  - *How much is sustainable?*
  - *What data is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



(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.**



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

CrossMark click for updates

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

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

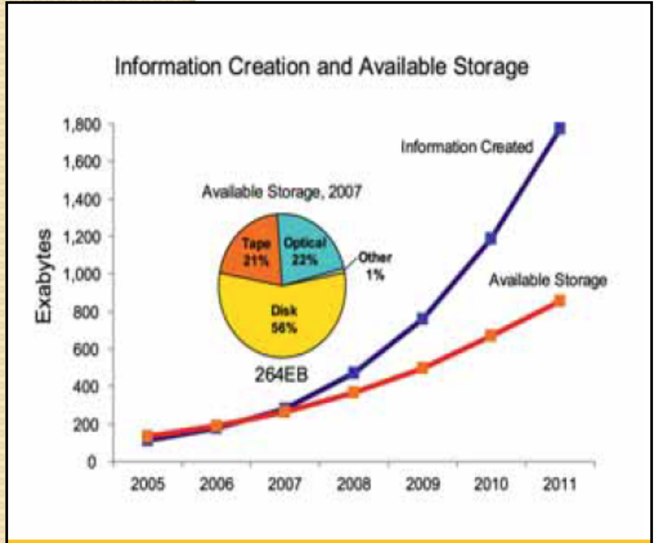
In addition to providing a rough estimate of the total number of datasets produced per year by NIH-funded researchers, this study identifies additional issues that must be addressed to

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.

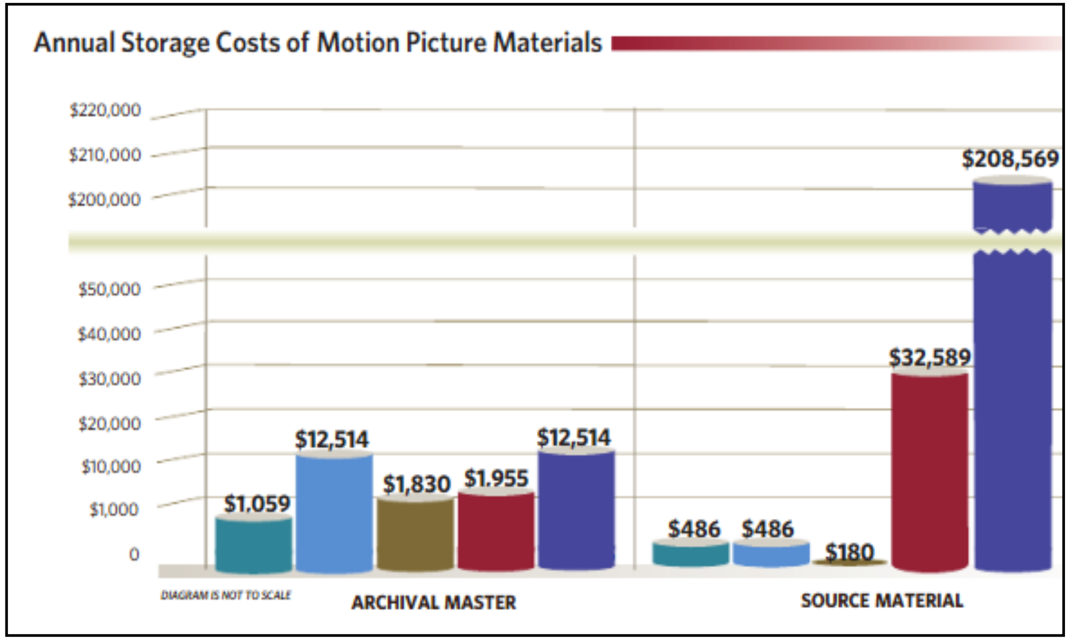
From this perspective, they are not interested in data preservation or even data sharing other than as a necessary precondition to data reuse; they are interested in conformance to their data management and sharing policies because it is the only way they can create the preconditions for data reuse.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



Source: IDC, 2008

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



COLOR KEY

- ■ ALL FILM
- ■ FILM CAPTURE, 4K MASTER
- ■ DIGITAL CAPTURE TO HDCAM SR TAPE, 1920 X 1080 MASTER
- ■ DIGITAL CAPTURE TO 2K DATA, 2K MASTER
- ■ DIGITAL CAPTURE TO 4K DATA, 4K MASTER

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

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

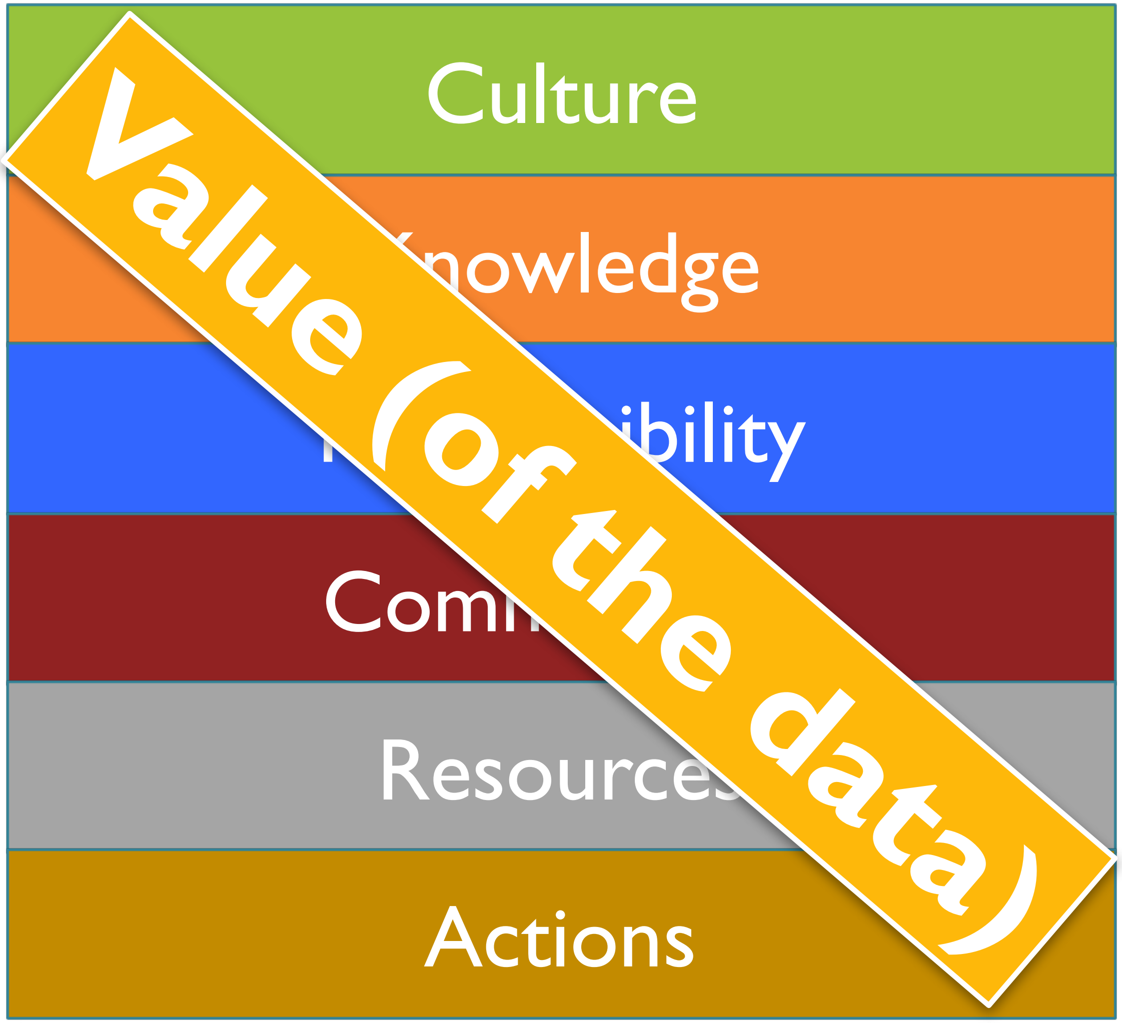
Resources

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

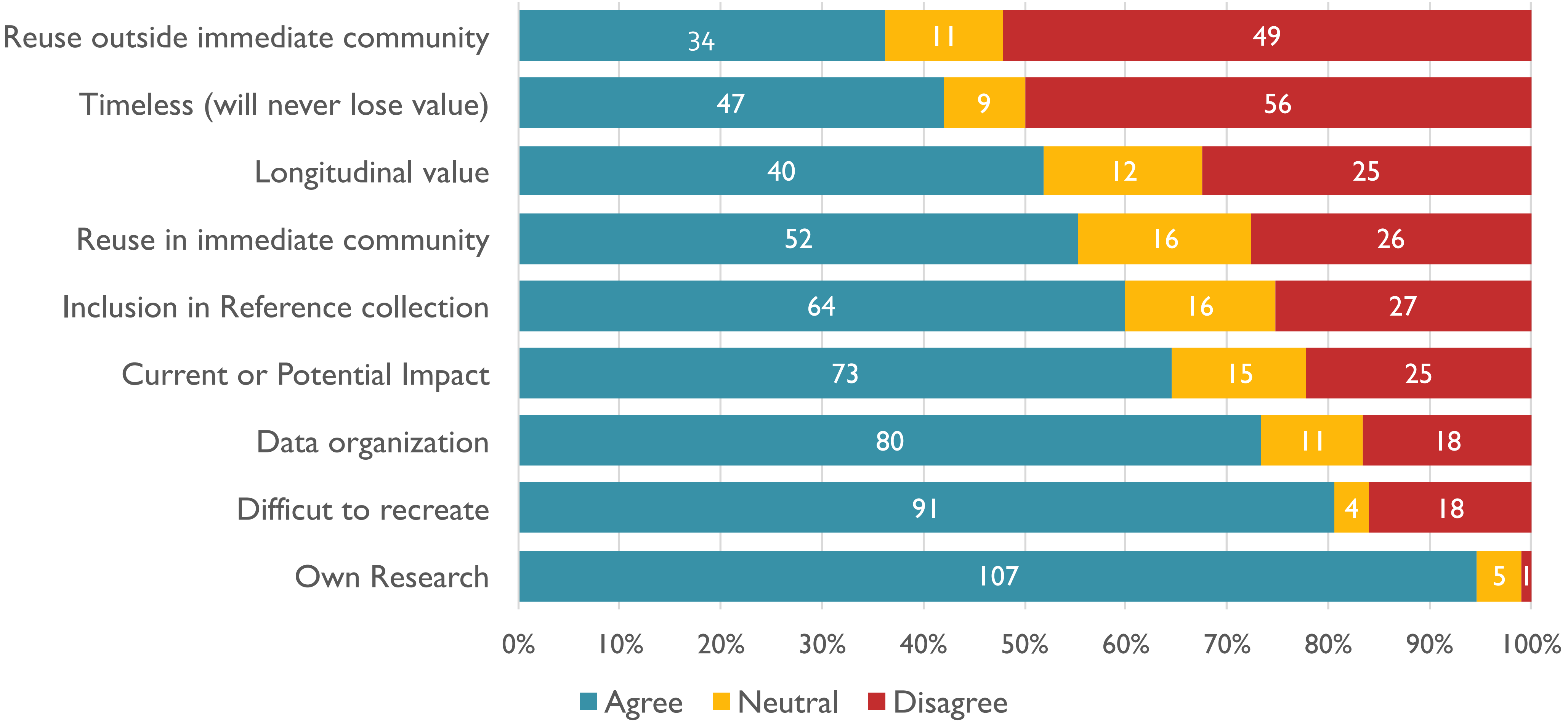
Gap between the people, money, infrastructure, and tools needed to steward data, and what is now available

Gap between the actions taken to facilitate stewardship of data and the actions needed

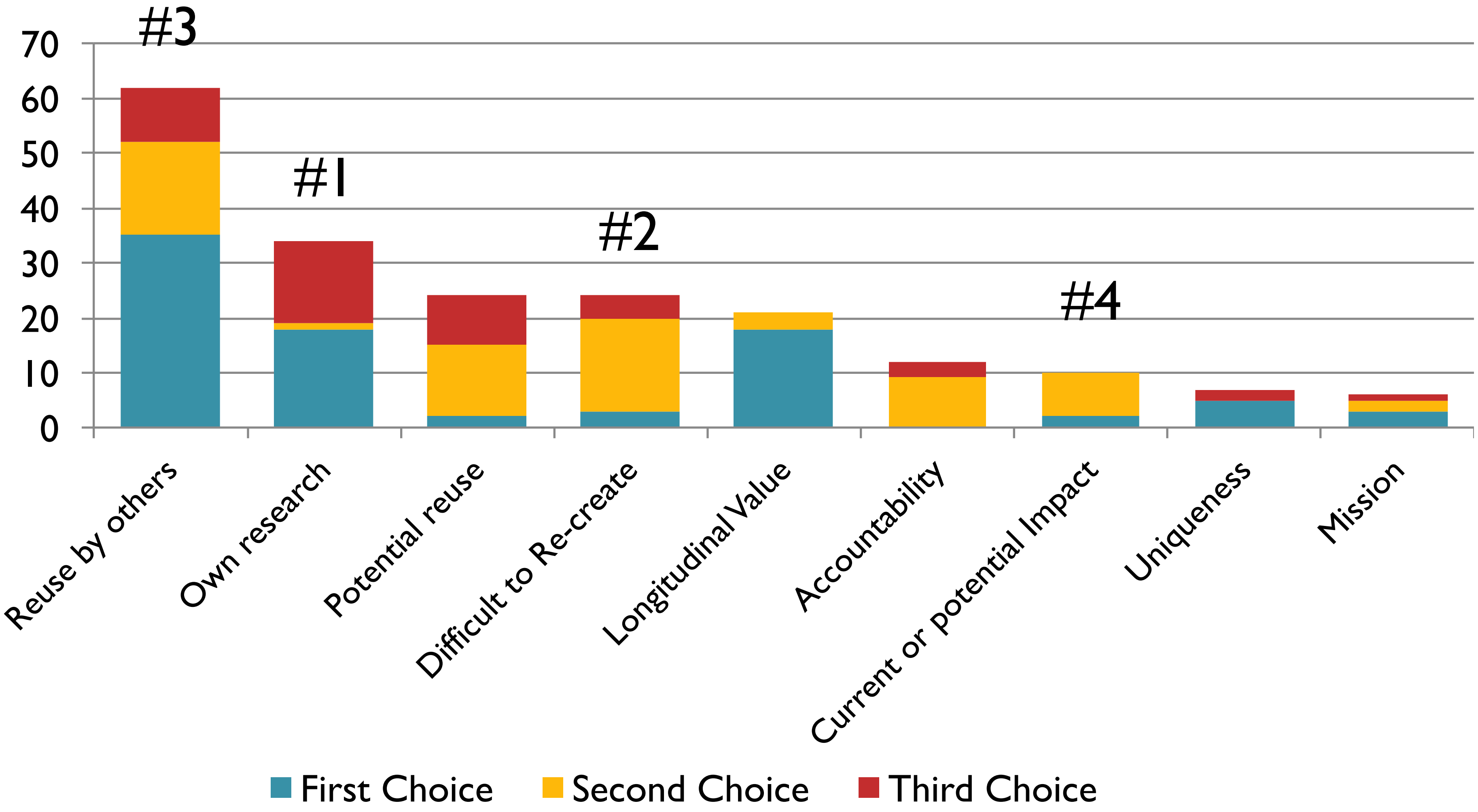
The Critical Importance of Value

- 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?**

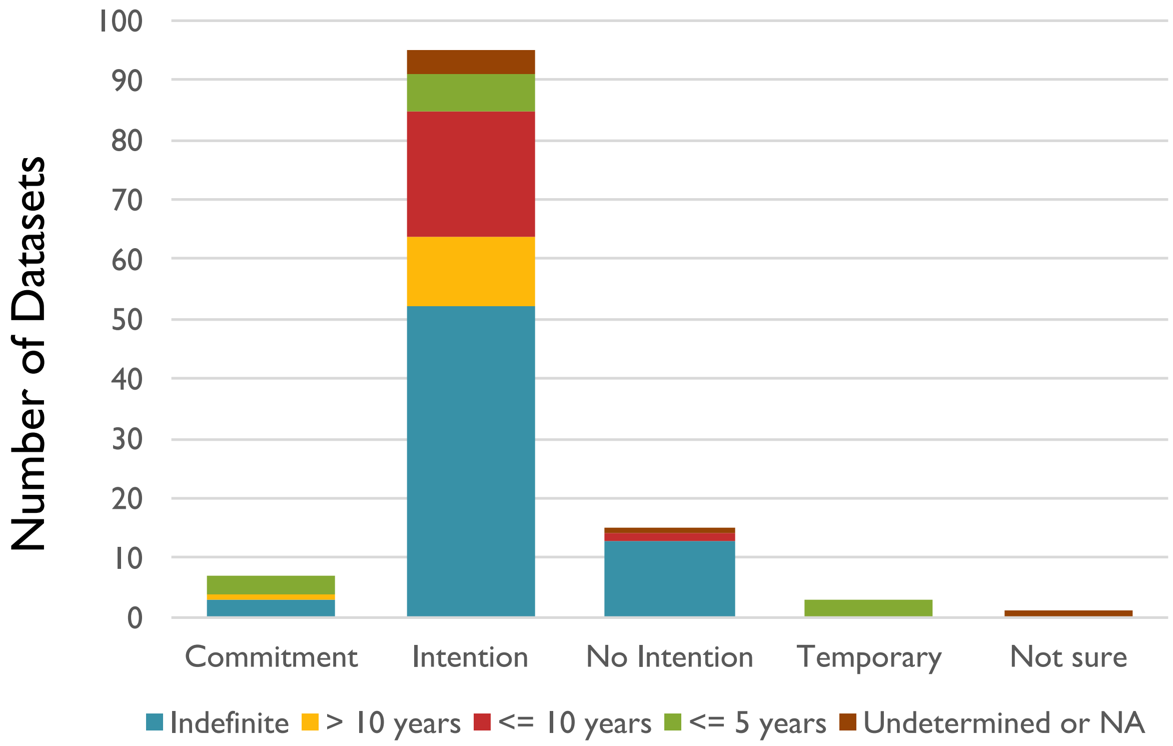
Researcher Agreement with Type of Value



Reasons for Value with Greatest Impact on Preservation Decisions



Type of Commitment and Term of Commitment



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

Term of Value

Term of Commitment or Intention

	Indefinite	> 10 years	<= 10 years	<= 5 years	Undetermined
Indefinite	31	6	9	18	4
> 10 years		10			2
<= 10 years	10	4	7	1	
<= 5 years	3	1	1	7	
Undetermined	3			1	1

But How Much Commitment Is There?

Term of Value

Term of Commitment or Intention

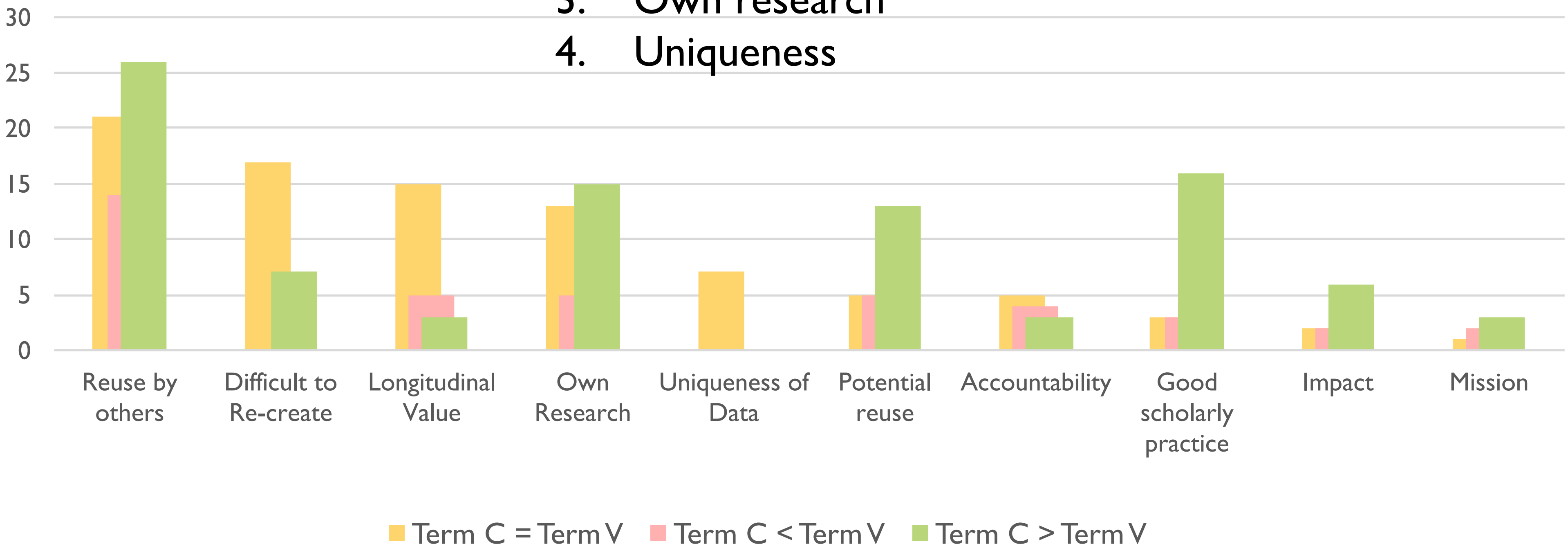
	Indefinite	> 10 years	<= 10 years	<= 5 years	Undetermined
Indefinite	2			1	
> 10 years		1			
<= 10 years				1	
<= 5 years	2	1			
Undetermined				Unsure	

Type of Value with Greatest Impact on Preservation Decisions

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

Where **Term CI = Term V**, the most common types of value are

1. Difficult to re-create
2. Longitudinal
3. Own research
4. Uniqueness



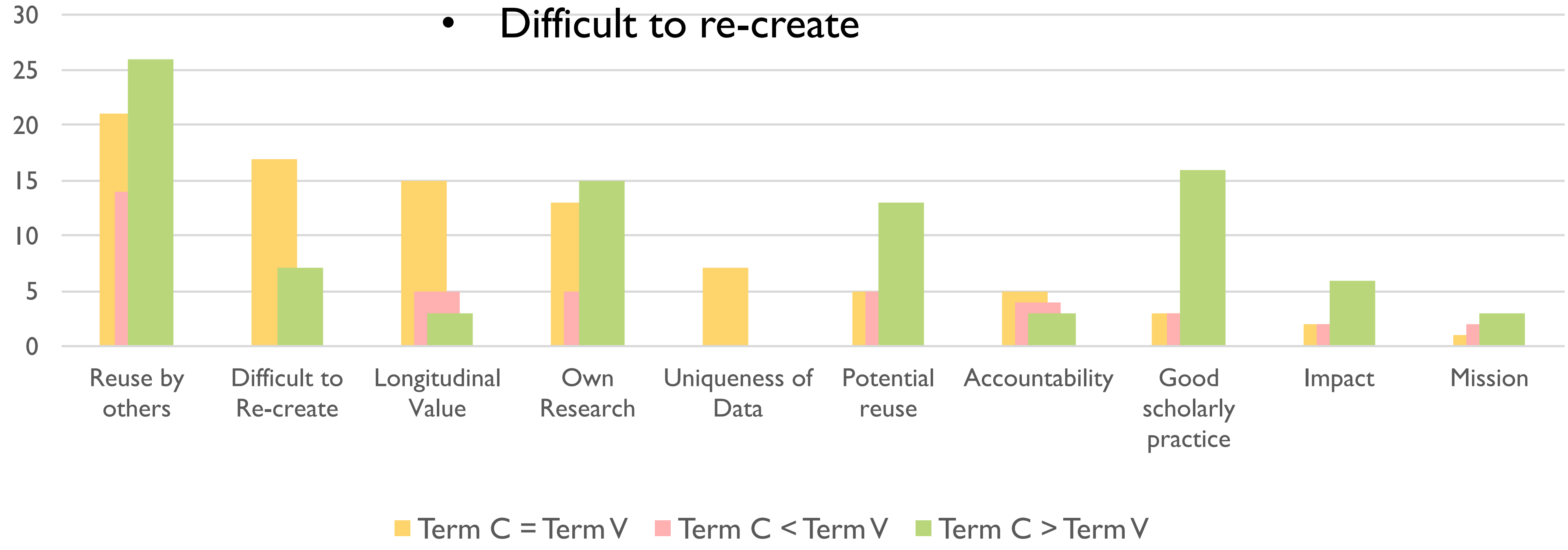
Type of Value with Greatest Impact on Preservation Decisions

Where **Term CI > 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



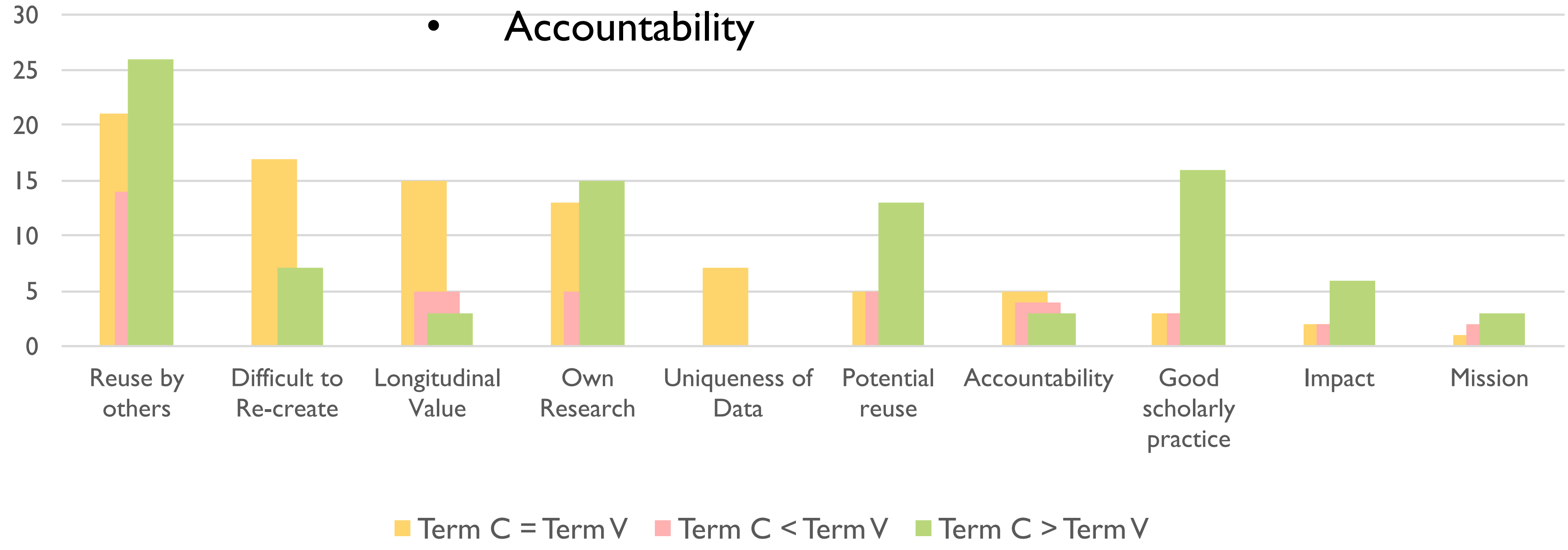
Type of Value with Greatest Impact on Preservation Decisions

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

	Term CI = Term V	Term CI > Term V	Term CI < Term V
Reuse	x	x	x
Difficult to re-create	1	4	
Longitudinal	2	x	1
Own research	3	3	2
Uniqueness	4		
Potential reuse	x	2	3
Accountability	x	x	4
Good scholarly practice	x	1	x
Impact	x	x	x
Mission	x	x	x

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?