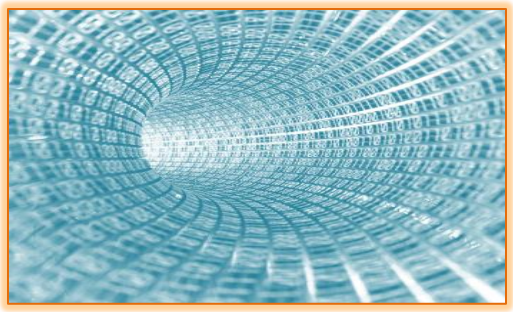


In Search of Impact

Measuring the Impact of Digital Repositories Workshop 2017

Dr. Fran Berman

Hamilton Professor of Computer Science, RPI

Chair, Research Data Alliance / US

Rensselaer

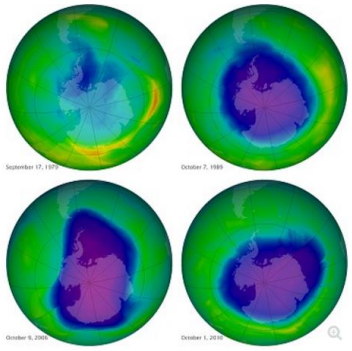
why not change the world? SM

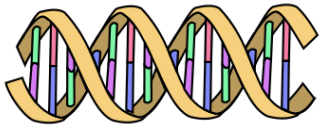
Why do repositories matter to the community?

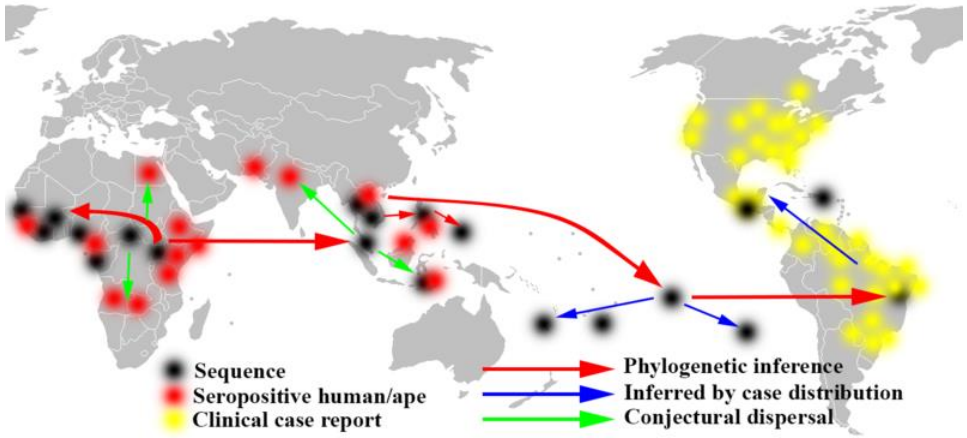
Repositories provide a safe and accessible home for data

Data drives innovation

Innovation drives societal and scientific advancement





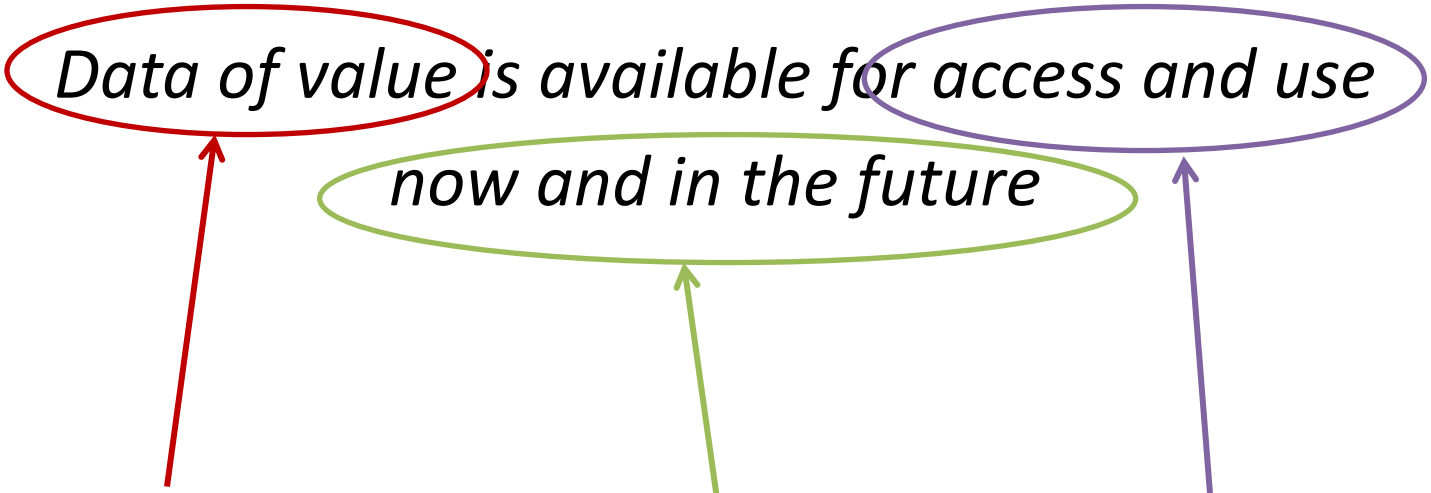


Rensselaer
why not change the world? SM

Fran Berman

Images: Wikimedia commons, NASA, https://blogs.microsoft.com/iot/2015/11/02/microsoft-and-esri-work-together-to-fuel-innovation-through-location-based-services/

Deconstructing impact: What does it mean for a repository to be successful?



What makes data valuable?

What are best practices for data repository stewardship and preservation?

What "data services" do we need?

Deconstructing impact: What does it mean for a repository to be successful?

Data of value is available for access and use

now and in the future

What makes data valuable?

What are best
practices for da
repository
stewardship an
preservation?

ata
o we

Data and Value

What data is valuable …

to society?

– Official and historically valuable data (Census information, government records, Shoah Collection, etc.)

to the research community

– Data from instruments, studies, projects; data underlying publications and results

to me?

– Financial data, digital family photos; personal records, etc.

Many kinds of valued research data

[http://www.colorado.edu/ibs/cupc/stewardship_gap/]

– Data that is valuable for one's own research
– Data that is in demand by other researchers for replication or reuse
– Data that is mandated to be preserved by policy or regulation
– Data that is expected to be preserved as part of good scholarly practice
– Data that is highly cited
– Data for which value accrues over time
– Data that underlies assessment reports
– Data that is costly to reproduce or cannot be reproduced
– Data that is timely, costly or difficult to create, etc.

Images: personal, http://www.politicususa.com/2016/08/20/bad-twitter-skills-backfiring-donald-trump.html

Value is in the eye of the beholder

Broad spectrum of valuable community data

Research Data Alliance Domain Data-focused Groups:

- Agricultural Data Interest Group (IG)
- Empirical humanities metadata Working Group (WG)
- Fisheries Data Interoperability WG
- International Materials Resource Registries WG
- On-Farm Data Sharing WG
- Rice Data Interoperability WG
- Chemistry Research Data IG
- Geospatial IG
- Global Water Information IG
- Health Data IG
- Linguistics Data IG
- Marine Data Harmonization IG
- Small Unmanned Aircraft Systems' Data IG
- Etc.

SDSS

tair

PDB PROTEIN DATA BANK

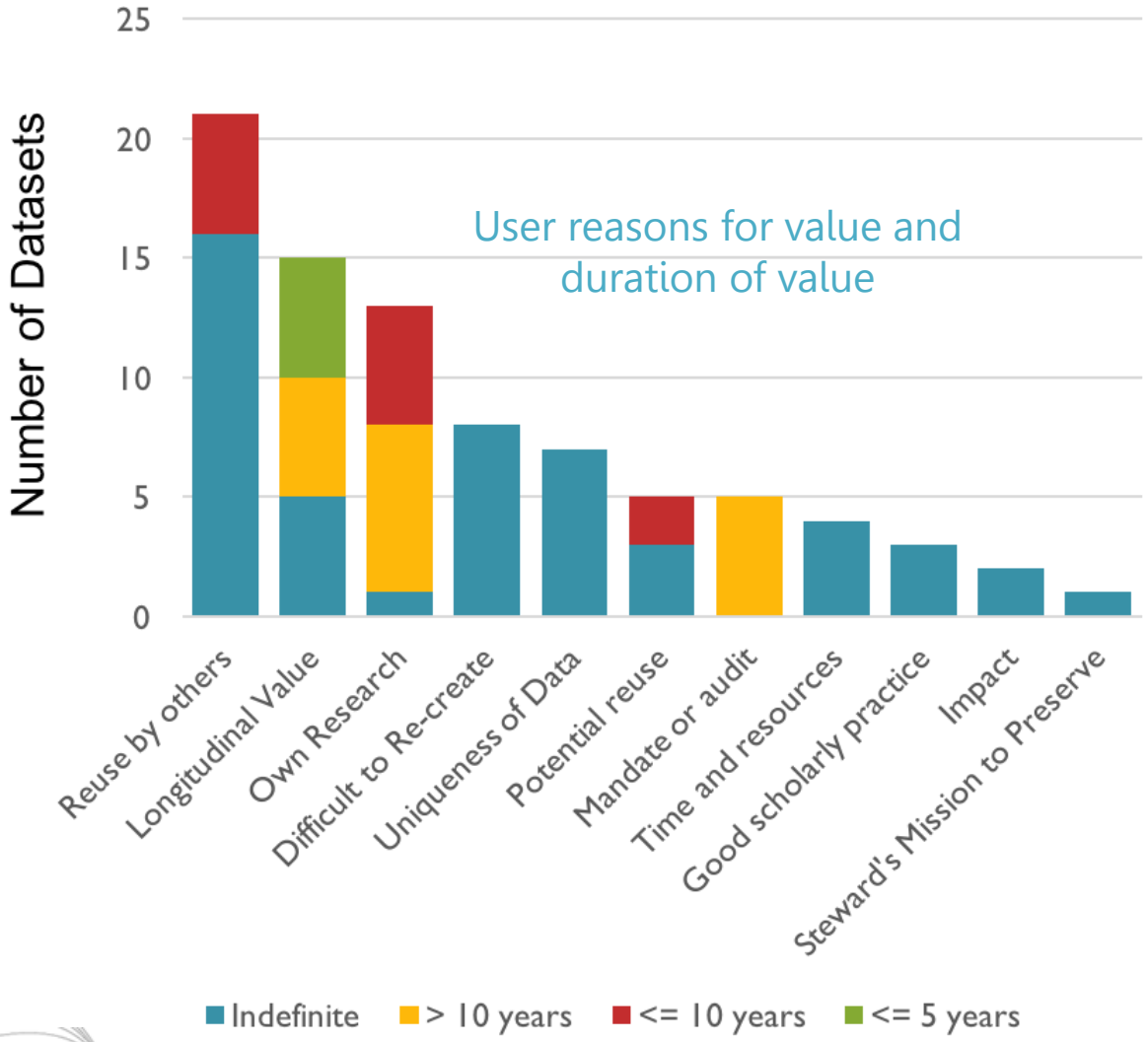
Reference collections

Image: https://www.kickstarter.com/projects/336056946/the-astronomy-legacy-project

USC Shoah Foundation
The Institute for Visual History and Education

Irreplaceable collections

Value over time



Stewardship Gap Study

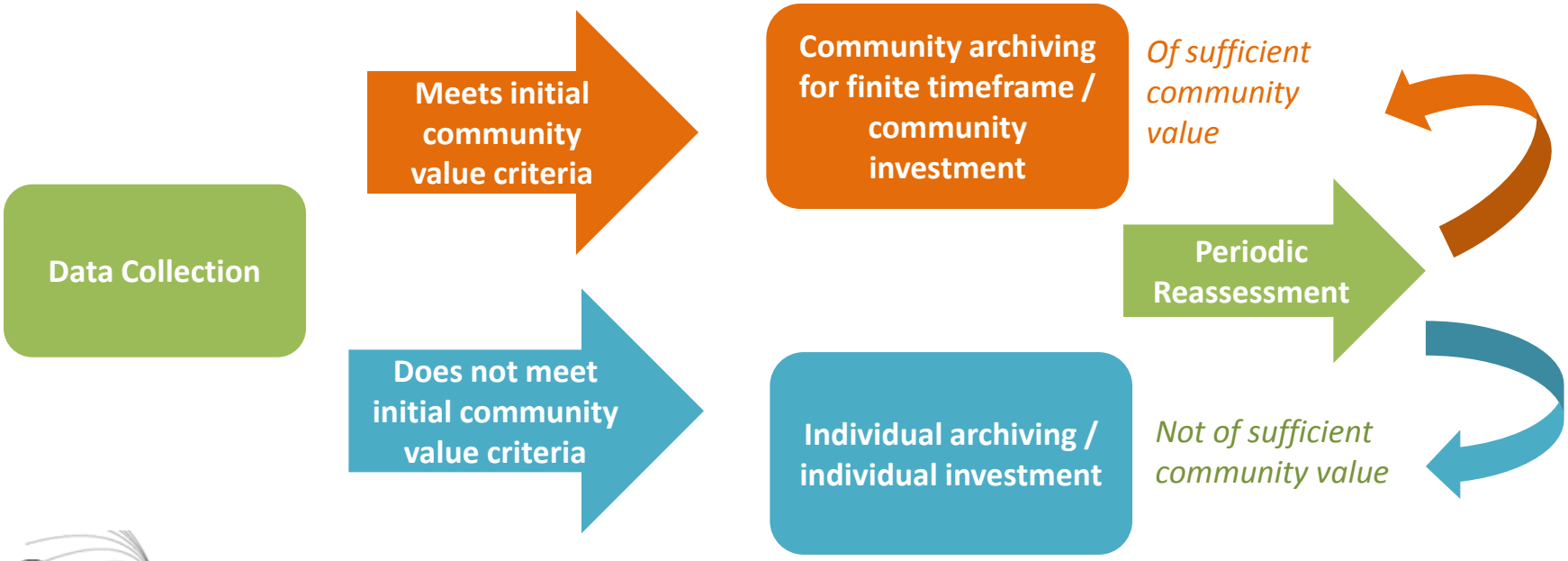
[**Myron Gutmann**, Jeremy York, Fran Berman + Advisors]:

- 46 respondents
- 120 Datasets
- 79 Domain Areas
- Respondent research sponsors: NSF (50+ datasets) NIH (35 datasets), <10: NASA, NEH, Sloan, Bureau of Reclamation, DoE, DoD, CDC, etc.
- More info: http://www.colorado.edu/ibs/cupc/stewardship_gap/

Rensselaer
why not change the world? SM

What value of data is worth what amount of stewardship investment and for how long?

- Value and investment discussion largely decoupled. What mechanisms should we have for pairing value and investment?
- *Finite / customized stewardship investment. Where are the thresholds? What should the criteria be?*



What metrics speak to user value for repository datasets?

- **"Popularity":**
  - Data collections associated with highly cited publications
  - Data collections with many downloads
  - Data collections with many hits, distinct users, return users, etc.
  - Data collections with large user base
- **"Responsibility"**
  - Data that is expected to be retained by stakeholders or community
  - Unique or hard to replace collections
  - Data in dark archives for other sites, etc.
- **"Empowerment"**
  - Data behind key community results and discoveries (as measured by prizes, key publications, etc.)
  - Community reference collections
  - Data on which new results depend

Deconstructing impact: Access and Use

Data of value is available for access and use now and in the future

What "data services" do we need?

Access – Where is the data? / Does the data I need exist?

- Repositories generally make it easy to find their datasets.
- *Going up one level:* how do users find the right repositories (and their datasets)?

- Usual search engines currently inadequate
  - Is there **sufficient metadata** to find the data?
  - **Keyword problem** – which term do I use?
  - **Timeliness** – which dataset or version is the most recent? Used in the publication, study, experiment?

Word cloud on data stewardship literature generated by Jeremy York from the Stewardship Gap Project http://www.colorado.edu/ibs/cupc/stewardship_gap/resources.html

Making data accessible is not good enough

- **Services and additional information critical to make data useful**
  - Data is not an asset if you don't know what it means.
  - Data is not useful if you can't find it.
  - Data needs to be in the right form for analysis.
  - Data needs to be preserved for results to be reproducible.

Repository Services: Users want more than just a big hard drive

Would data storage be helpful to you in your [Research Data Alliance] Interest or Working Group?

Respondent group 1:

Please don't think I am looking a gift horse in the mouth, but my comment would be that I hope it is **permanent storage** and that wherever this storage is available from, it has **sufficient infrastructure** to guarantee backups and [as]sign **persistent identifiers** to it, otherwise in another decade, the next generation of data rescuers will be rescuing the same set of data.

Also, and this sounds mean and ungrateful, but having been through a similar situation in []where a certain cloud company offered free storage, and forgot to mention that they were going to **charge for people to access the data**. … **who is going to manage** and coordinate the storage over the longer term.

Respondent group 2:

There are certainly projects as well as points on our project roadmap where additional sites would be useful. As far as information that might be helpful to feed back to the partner, I'd wonder the following:

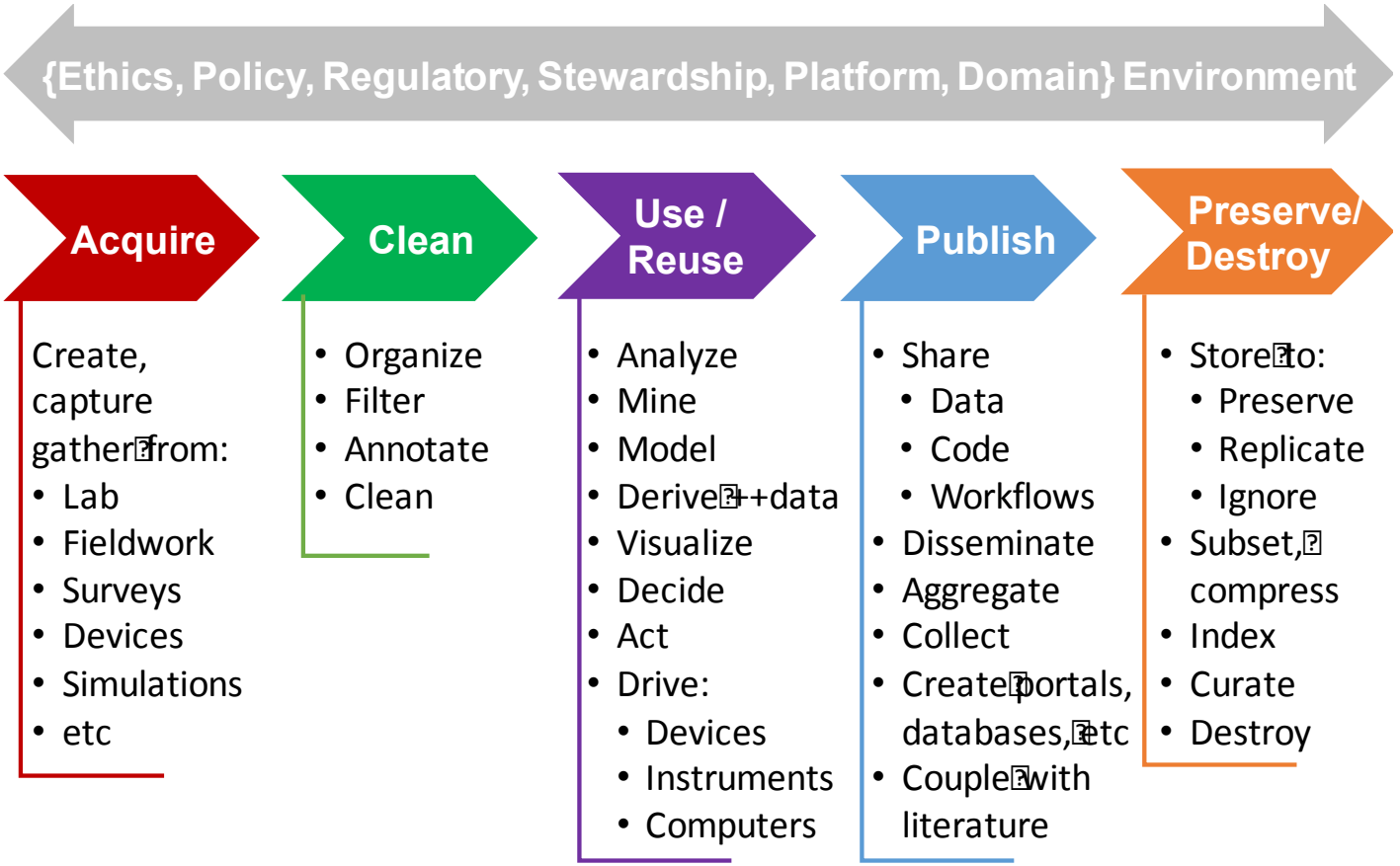
- What **type of storage** (object store vs. NFS-style).
- What **security** regimes are supported? (We only support open data at the moment, but if they can host PHI, HIPAA, that is an interesting data point).
- **Size.** Generally our pilots use 1 TB – 10 TB or 70-250 TB. …
- Are there are base level expectations or a **Service Level Agreement** (SLA) that would be offered along with the storage?
- If the storage becomes **inaccessible**, will and when can it be expected to recover? · Duration: storage is useful starting at 2 years with a sweet spot of 3. Longer (in years) is always better.

Respondent group 3:

If we were just trying to back up just our V0 … data, we would need at least **100TB**. To back up all of [], we would need **1PB.** That number is expected to grow to **4PB** by 2022.

Rensselaer
why not change the world?℠

What Services do Users Want?



Rensselaer
why not change the world? SM

Data Life Cycle from NSF CISE Data Science Subcommittee Report, https://www.nsf.gov/cise/ac-data-science-report/

Many service models specific to community and use cases; no one-size-fits-all

ICPSR Services

- Management and documentation tools
- Organization and data cleaning tools
- Support for privacy, confidentiality, security
- Sampling and workflow tools, etc.

Protein Data Bank Services

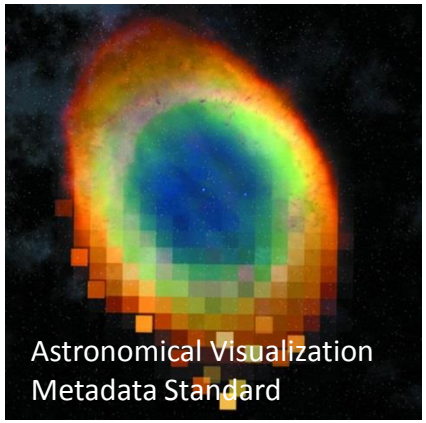
- Ingest tools
- Visualization tools
- Sequence and structure alignment, protein symmetry, analysis tools
- Education and training tools, etc.

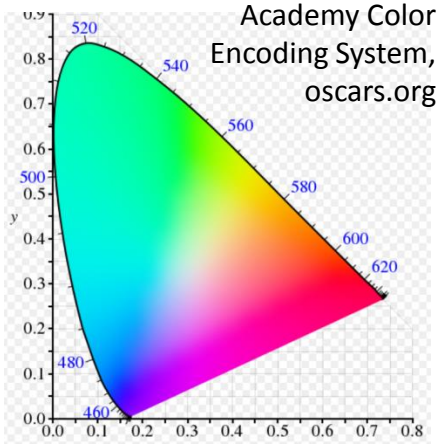
RDA Wheat Interoperability Working Group Recommendations

- Interactive "cookbook" with recommendations and guidelines on data format and standards
- Common wheat-related vocabularies to be made accessible in a human and machine-readable bio-portal
- Prototype interoperability framework for specific use cases, etc.

Services: Standardization and Best Practice

- **Standardization:** Data use cases vary with respect to community consensus and maturity.
  - **Standards needed:** Services much more effective when useful standards have been created
  - **At the right time:** Experimentation with different approaches often needed to develop useful standards
- When is good practice ready for standardization?
- What role should repositories (and funders, publishers, professional societies, domain communities …) play with respect to standardization of existing and differing community practices?


Astronomical Visualization Metadata Standard



Academy Color Encoding System, oscars.org

RDA RESEARCH DATA ALLIANCE

Findable Accessible Interoperable Reusable

asis&t Association for Information Science and Technology

ISO

NIST

Rensselaer why not change the world? SM

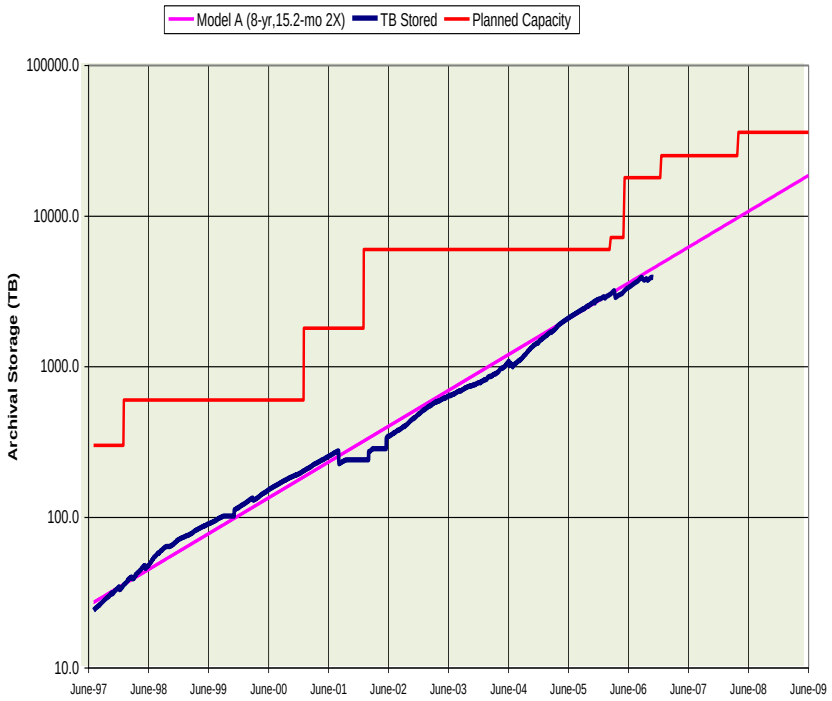
Fran Berman

Images: Nasa.gov, wikimedia

Services Cost

Data Central
datacentral.sdsc.edu

- **Data Central at SDSC**: Support for storage, access and use of community data collections (circa 2000's). [Natasha Balac, lead]
- Services, expertise and resources offered to support:
  - Database hosting and long-term storage
  - Data management and schema design
  - Data analysis, mining, and visualization
  - Portal creation and collection and publication
  - Consulting, training, strategic collaboration
- ***Free to users.*** Cost of services and storage ultimately prohibitive …

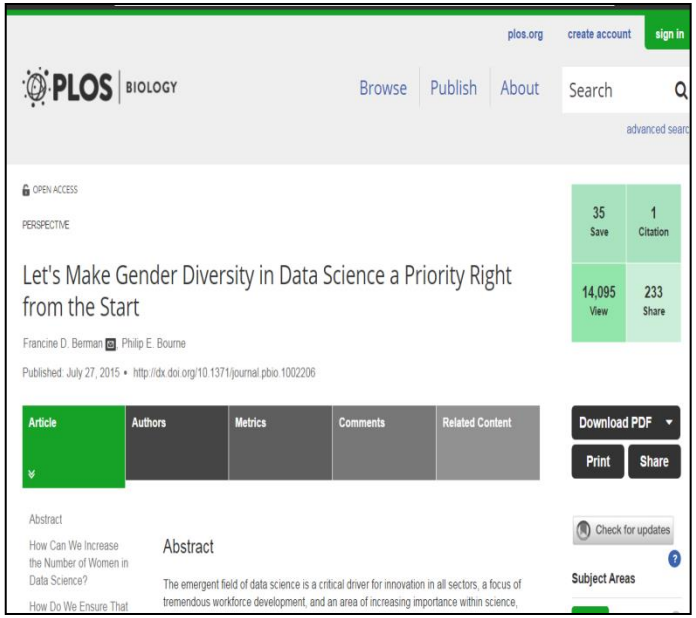


SDSC Data Storage Growth '97-'09 (PBs)

Rensselaer
why not change the world? SM

Measuring Access and Use. What does success look like?

- Some measures of access and use:
  - Publications
  - Citations
  - Downloads
- Is data still valuable even if no-one is currently using it?
- What should the community's role be in determining repository data collections?
- **How should organizations determine investment in stewardship / preservation vs. investment in services?**



Rensselaer
why not change the world? SM

Deconstructing impact: Stewardship and Preservation

Data of value is available for access and use

now and in the future

What are best practices for data repository stewardship and preservation?

Stewardship and Preservation challenges are real

The Atlantic, 12/13
https://www.theatlantic.com/national/archive/2013/12/scientific-data-lost-forever/356422/

Most Scientific Research Data From the 1990s Is Lost Forever

A new study has found that as much as 80 percent of the raw scientific data collected by researchers in the early 1990s is gone forever, mostly because no one knows where to find it.

DANIELLE WIENER-BRONNER | DEC 23, 2013 | NATIONAL

Share Tweet ...

TEXT SIZE

This article is from the archive of our partner 'wire

A new study has found that as much as 80 percent of the raw scientific data

GOOGLE TO HOST TERABYTES OF OPEN-SOURCE SCIENCE DATA

Wired, 1/08
https://www.wired.com/2008/01/google-to-provi/

Sources at Google have disclosed that the humble domain, http://research.google.com, will soon provide a home for terabytes of open-source scientific datasets. The storage will be free to scientists and access to the data will be free for all. The project, knov to the scientific commur Googleplex last August,

GOOGLE SHUTTERS ITS SCIENCE DATA SERVICE

Wired, 12/08,
https://www.wired.com/2008/12/googlescienceda

Rensselaer
why not change the world?℠

Fran Berman

What are best practices for data stewardship?

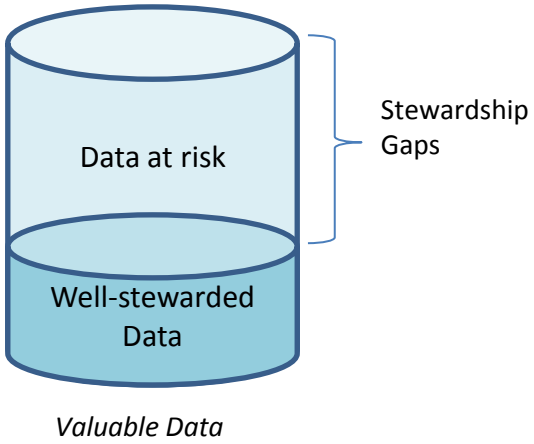
- **Organizational perspective**: What does it mean for a repository to be a good steward?
- What do **users** think good stewardship is?
  - Little data loss / no loss of *their* data
  - Repository reliability / sustainable economics
  - Ease-of-access, support / services for data use
  - Ability to link data to relevant publications
  - Repository respected by the community, etc.

TRAC

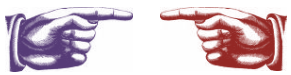
Trustworthy Repositories Audit & Certification

Often various "gaps" between existing stewardship and best practice

- **Resource / Infrastructure / Economic Gaps:**
  - Insufficient funding
  - Insufficient tools for management, use, discovery, preservation
  - Insufficient facilities, utilities, etc.
  - Insufficient staff



- **Cultural / Social Gaps:**
  - Lack of sufficient institutional and/or individual commitments
  - Differing expectations of researchers, stewards, and stakeholders
  - Gap between local practice and good / best practice



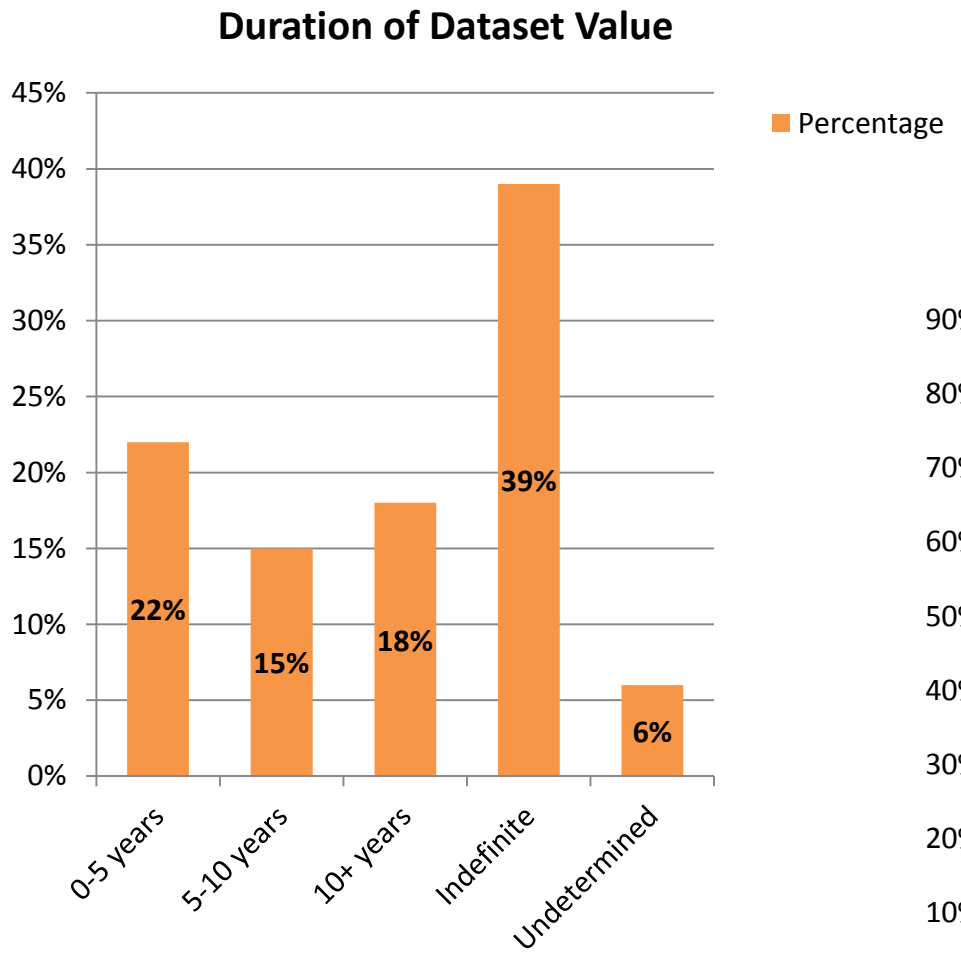
- **Political Gaps:**
  - Lack of policy / practice promoting access, stewardship, sharing
  - Lack of stakeholder support

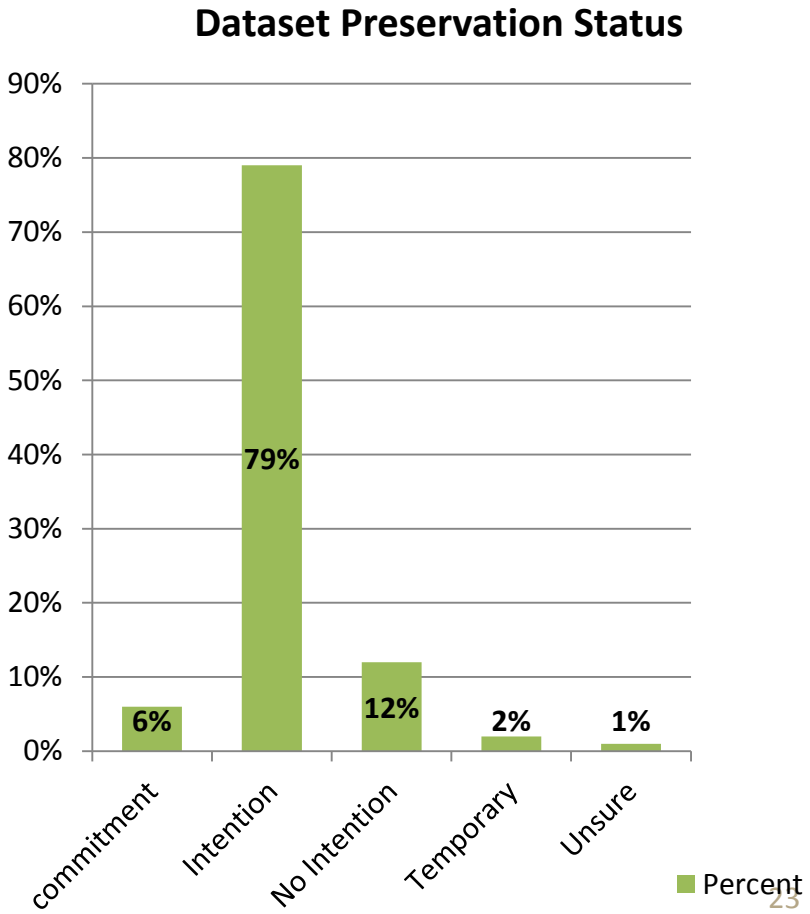
Preservation Challenges: User Intent vs. Stakeholder Commitment

Stewardship Gap Study

[**Myron Gutmann**, Jeremy York, Fran Berman + Advisors]:

- 46 respondents
- 120 Datasets
- 79 Domain Areas
- Respondent research sponsors: NSF (50+ datasets) NIH (35 datasets), <10: NASA, NEH, Sloan, Bureau of Reclamation, DoE, DoD, CDC, etc.
- More info: http://www.colorado.edu/ibs/cupc/stewardship_gap/





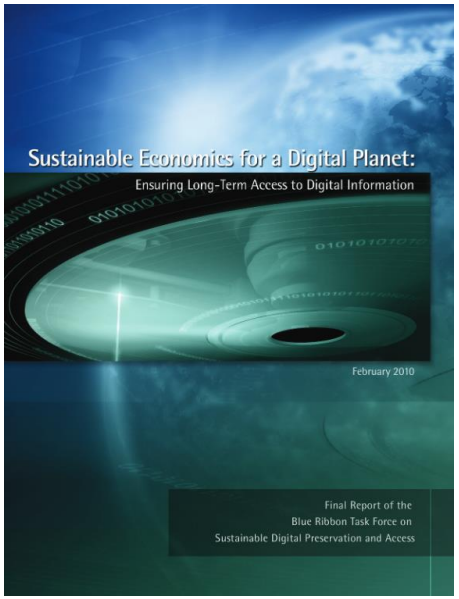
Rensselaer
why not change the world? SM

Fran Berman

Preservation and sustainability challenge: Stakeholder misalignment

Digital Data Stakeholders:

- **Those who generate** the data
- **Those who benefit** from use of the data
- **Those who select** what to preserve
- **Those who own** or have rights to the data
- **Those who preserve** the data
- **Those who pay** for infrastructure



The greater the alignment between key stakeholder groups, the better the prospects for good stewardship and sustainable preservation

[Blue Ribbon Taskforce for Sustainable Digital Preservation and Access, brtf.sdsc.edu]

Why is data stewardship and preservation such a hard sell?

- **Newsworthiness:** Hard to "market" compared to more urgent/short-term competing priorities
- Quantifying **opportunity cost** a challenge
- **No gaps:** Business model must be sustainable and address infrastructure refresh and evolution

	Stewardship and Preservation Infrastructure	Supercomputers
Metrics of Success	High reliability; Minimal data loss and damage	High Performance; good ranking on the Top500 list; application impact
Next Generation Systems	Smooth migration for data critical: Datasets must migrate to new media without loss of data or disruption to users	Growth in capability/capacity key: Compatibility of systems not required although there should be application transition paths
Funding Model	No gaps. Funding must be available for continuous support of data collections	Serial "one time" funding for each new HPC resource possible

Rensselaer
why not change the world? SM

Preservation in the News

- *Where is the data going?*
- *How will we find it?*
- *How will it be sustained?*

NY Times, 12/16,
https://www.nytimes.com/2016/12/01/nyregion/harvesting-government-history-one-web-page-at-a-time.html

Harvesting Government History, One Web Page at a Time

About New York
By JIM DWYER DEC. 1, 2016

Poring over federal records, reports and research at the New York Academy of Medicine on Thursday. "It's exciting stuff," a volunteer said. "It looks dull." Chang W. Lee/The New York Times

By noon on Thursday, Davis Erin Anderson had copied the addresses of a few dozen websites and online PDFs that listed signs of climate change by state and region.

About New York
Twice a week, a chronicle of New York and New Yorkers.

He Sent de Blasio 2 Dozen Letters. He Got Zero Replies.

First Came Giuliani's Input on the Immigration Order. Now There's the Court Test.

Drawing a Line From Alternative Theories to Untruths

In an Age of Cybercrime, Low-Tech Thieves Target Mailboxes

Honey, Soap and Towels: Signs of Welcome a Refugee Family Won't See

See More »

Sustainable Energy

Climate Data Preservation Efforts Mount as Trump Takes Office

Universities host hackathons to save environmental information amid fears the Trump administration will scrub data that undercuts its views.

by James Temple January 20, 2017

NASA researchers believe the West Antarctic Ice Sheet may be in a state of irreversible decline, contributing to rising sea levels.

MIT Technology Review, 1/17,
https://www.technologyreview.com/s/603402/climate-data-preservation-efforts-mount-as-trump-takes-office/

Scientists across the US are scrambling to save government research in 'Data Rescue' events

Dana Varinsky
Feb. 11, 2017, 11:32 AM 3,850

Business Insider, 2/17
http://www.businessinsider.com/data-rescue-government-data-preservation-efforts-2017-2/

- **Groups are downloading and archiving government data out of fear the Trump administration might delete it.**
- **Some sites have already undergone changes.**

Rensselaer
why not change the world? SM

What repository metrics speak to usage and users?

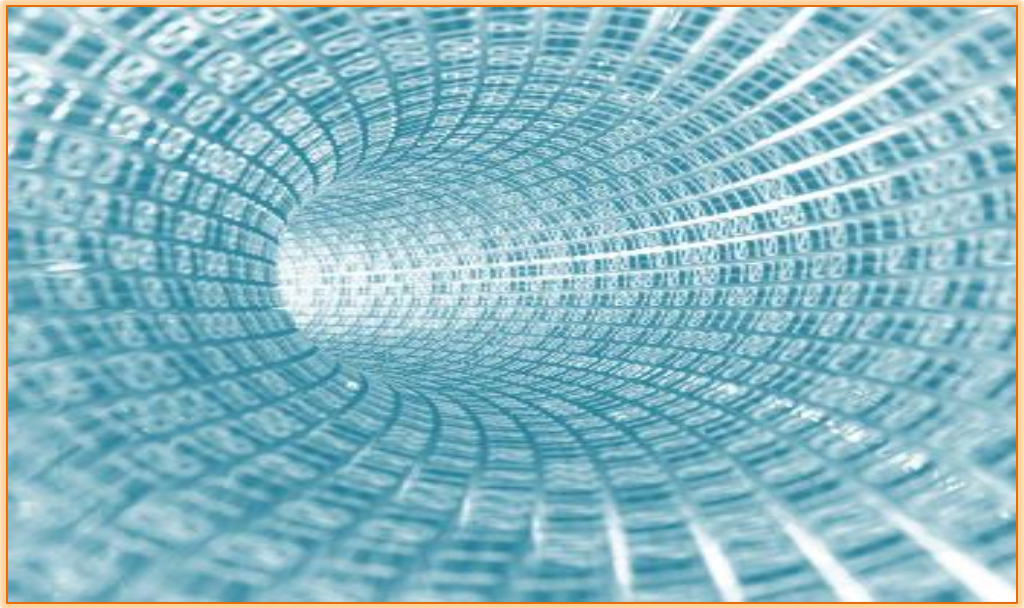
- **Usefulness / usage**
  - Number of collections, number of users, number of return users, number of web hits
- **"Empowerment"**
  - Use of repository data for new results and discoveries as measured by publications, prizes, citations, etc.
- **Responsibility / Community value**
  - Availability of reference collections, unique or hard to replace collections, replication collections to mitigate risk of data loss
- **Adequate "ilities": Reliability / Predictability / Sustainability / Affordability / Discoverability**
  - No data loss, data is easy to find and use, data will be there when you need it, data access and use fees are not a roadblock to effective use

Leveling up: How do we create an ecosystem that supports effective data stewardship?

The Goal: ***Data of value is available for access and use now and in the future***

- **Stakeholder "Bio-diversity":** How can the preservation community create ***cross-sector partnerships*** that protect valued data and **mitigate risks** of data damage and loss?
- **Realistic resourcing:** How do we develop / sustain the resources needed for data stewardship, preservation, use, and access as ***enabling infrastructure*** rather than new innovation
- **Culture change:** How do we create the **technical infrastructure and social structures** for now and the future that will help ensure that we get the most from our data?

Thank You



Rensselaer

why not change the world? SM