Organizational challenges to promoting data sharing, stewardship and preservation

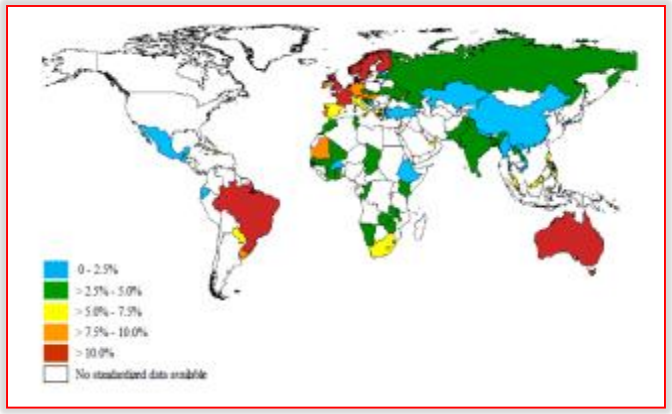
Dr. Francine Berman

2019-2020: Katherine Hampson Bessell Fellow, Radcliffe Institute, Harvard University

Hamilton Distinguished Professor of CS, RPI

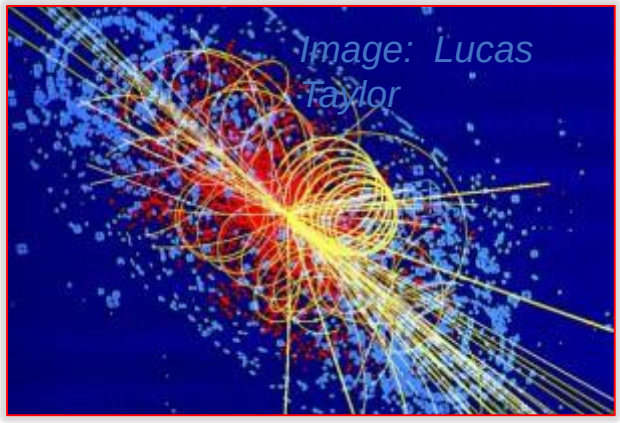
Co-founder, Research Data Alliance

Data drives Discovery



Who is most at risk to contract asthma?

How can we increase wheat yields?



How accurate is the Standard Model of Physics?

How can we best address energy needs and sustain the environment ?

Image: Ceinturion, Wikipedia

Stewardship, preservation, infrastructure and tools needed to efficiently support data sharing and data-driven discovery

- Data is not useful if you can't find it.
- Data is not an asset if you don't know what it means.
- Data needs to be in the right form for analysis.
- Data needs to be preserved for results to be reproducible.

Challenges that must be addressed for effective data stewardship and data sharing

- Availability of tools to advance use and user experience
- Development of an effective curation, storage and preservation environment
- Stable resources / support of organizational infrastructure

Stewardship and Preservation in Commerce, Academia, Government: differences in stakeholder alignment

Digital Data Stakeholders			
	Commercial organizations	**Academic Institutions**	**Government agencies**
Those who **generate** the data	Customers / Company / Others	Researchers / Others	
Those who **benefit** from use of the data	Company	Scientific and Broader Community	
Those who **own or have rights** to the data	Company	Institution / community	
Those who **preserve** the data	Company	Researcher / Repository / ??	
Those who **pay** for infrastructure	Company	Funders / Institution	

Why is stable organizational support for stewardship and preservation such a hard sell for Academic Institutions?

Specific Challenges:

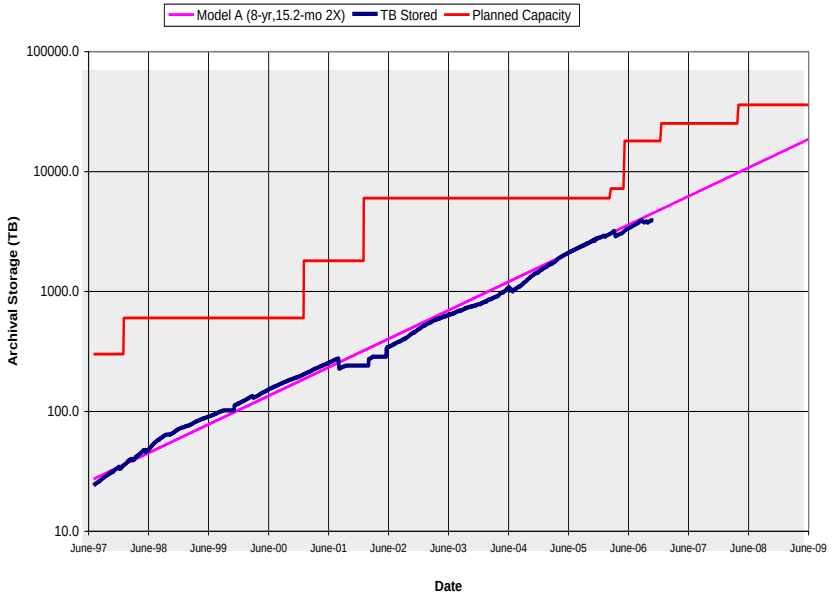
- Fear of commitment
- Lack of newsworthiness
- Misalignment with conventional incentive structures

Fear of (Economic) Commitment

Data Infrastructure costs include

- Maintenance and upkeep
- Software tools and packages
- Utilities (power, cooling)
- Space
- Networking
- Security and failover systems
- People (expertise, help, infrastructure management, development)
- Training, documentation
- Monitoring, auditing
- Reporting costs
- Costs of compliance with regulation, policy, etc. …

Resources and Resource Refresh



San Diego Supercomputer Center Data Storage Growth '97-'09

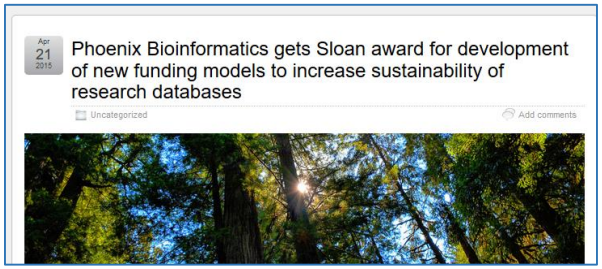
- *Most valuable data replicated*
- *As research collections increase, storage capacity must stay ahead of demand*

Information courtesy of Richard Moore, SDSC

Economic Sustainability: The Arabidopsis Information Resource (TAIR)

- TAIR is a community resource and **on-line model organism database** of genetic and molecular biology data for Arabidopsis thaliana.
- TAIR integrates information about genes, gene products, natural variants, mutant allelles, plant phenotypes, research literature

TAIR supported by NSF between 1999 and 2013



Phoenix BIOINFORMATICS

"... TAIR director Eva Huala told an international meeting ... that introducing a subscription system would destroy, not save, TAIR."

Phoenix Bioinformatics was founded in 2013 by the staff of TAIR, a curated database for plant genome information. After TAIR lost grant funding we pioneered a new sustainable funding model that provides support for TAIR. **Our nonprofit mission is to help other projects achieve sustainable support using the tools and expertise we developed for TAIR.**

(from Phoenix Bioinformatics website)

Do we have to keep everything forever?

What data is of value?

- Data that will be **re-used** by others
- Data that can be used for **longitudinal analysis**
- Data that is **difficult or impossible to reproduce**
- Data **associated with research publications**
- Data that is **required to be retained** by policy or regulation
- Data that is **used by a large community**
- Data whose **collection is expensive** in terms of time and / or resources
- Data **needed to ensure reproducibility** of results, etc.

PDB PROTEIN DATA BANK

SDSS

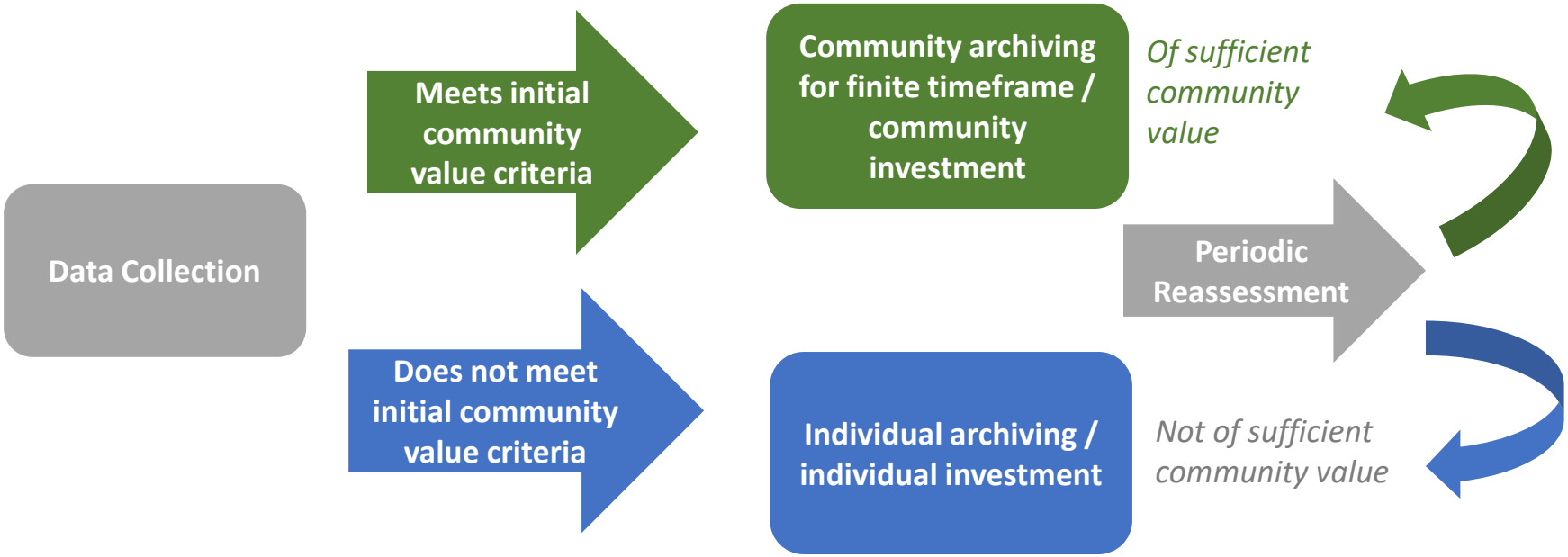
tair

USC Shoah Foundation
The Institute for Visual History and Education

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?*



Fran Berman

The Newsworthiness Problem: Moonshots vs. Disasters

COMPUTERS

USA gains ground in supercomputer world rankings

James Holloway | June 21st, 2019

The Department recorded perfor

Crisp photos of moon landing are missing

Spectacular images of day were stored, forgotten -- and lost

Marc Kaufman, Washington Post
Sunday, February 4, 2007

BBC News Sport Weather Capital Future

NEWS SCIENCE & ENVIRONMENT

Home UK Africa Asia Europe Latin America Mid-East US & Canada Business Health Sci/E

8 October 2013 Last updated at 14:01 GMT

Share

Higgs boson scientists win Nobel prize in physics

COMMENTS (643)

By James Morgan
Science reporter, BBC News

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Au

News & Comment > News > 2014 > March > Article

NATURE | NEWS

Scientists losing data at a rapid rate

Decline can mean 80% of data are unavailable after 20 years.

Elizabeth Gibney & Richard Van Noorden

Science News *from research organizations*

Breakthrough process to create cancer-killing drugs

New technique could produce treatments for wide variety of diseases

Date: June 4, 2019
Source: Dartmouth College
Summary: A new strategy for drug development can be used to produce targeted therapies against a variety of diseases.

Share:

RELATED TOPICS

Health & Medicine
› Pharmacology

FULL STORY

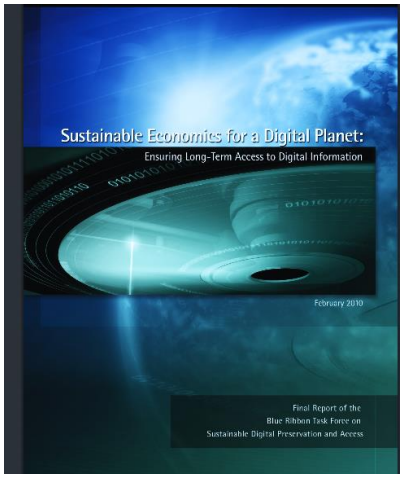
A research team at Dartmouth College has developed a new strategy for drug discovery and

Nearly 12 Million Quest Diagnostics Patients May Have Had Info Exposed In Data Breach

FOLLOW US

OUR | NEWSLETTER

Misalignment of stewardship and preservation support with the usual incentive structure



- **Hard to get recognition and funding**
  - Hard for **researchers** to get funding for tool development. Developing infrastructure generally does not advance "research reputation"
  - Hard to get both R&D funding and institutional funding for **data practitioners**
- **Academic research a relatively small "market"** compared to commercial infrastructure and users of open source SW
  - Research infrastructure often custom to most effectively support discovery

Image by Steve Jurvetson

Not everyone can be a "free rider" …

Organizational Value proposition and metrics of success

- **Development of organizational value proposition that competes with pressing priorities difficult**
  - "Promotes science" and "doesn't break" may not cut it
- **Related problem: What is success and how to measure it?**
  - How to determine the lost opportunity costs of not having adequate data infrastructure?
  - How to determine effectiveness and usefulness of existing infrastructure?
  - How to gauge the impact of infrastructure?
  - Who are the stakeholders (users, beneficiaries, managers) of infrastructure?

  etc.

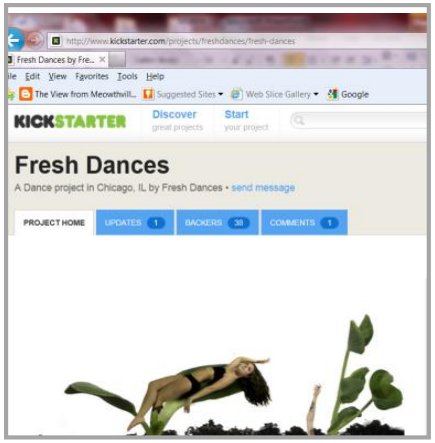
Towards a solution to the lack of adequate stewardship and preservation infrastructure

- **Move funding category from "research"** (exploratory, one-off) **to "infrastructure"** (needed for normal business operations, stable, long-term)
  - What is the right ***research : infrastructure*** ratio that maximizes competitive advantage?
- **Create a sustainability plan with a responsible underlying business model** (e.g. TAIR transition from NSF to Sloan to community)
  - Revisit what is "valuable" on a regular basis
  - Provide users transparency and early warnings about changes in the environment

There is no free lunch but you still need to eat …

Multiple business models used for data infrastructure

Donation, philanthropy

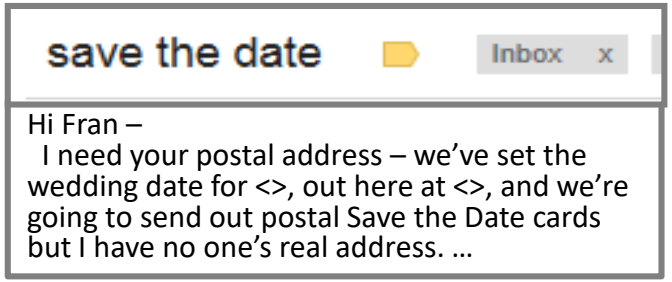


Crowd-sourcing

Subscription

Federal grants and contracts

Pay per service



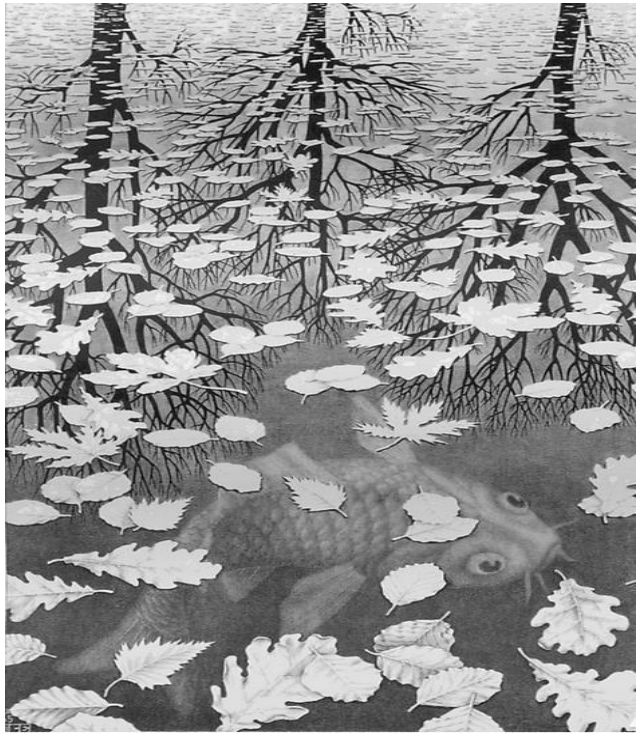
Fran's gmail

Accompanying advertisement

Advertisement

Thank You

bermaf@rpi.edu



"Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Networking and Information Technology Research and Development Program."

The Networking and Information Technology Research and Development (NITRD) Program

Mailing Address: NCO/NITRD, 2415 Eisenhower Avenue, Alexandria, VA 22314

Physical Address: 490 L'Enfant Plaza SW, Suite 8001, Washington, DC 20024, USA Tel: 202-459-9674, Fax: 202-459-9673, Email: nco@nitrd.gov, Website: https://www.nitrd.gov