Prepared for

Scholarly Communications Workshop
University of Pittsburgh

January 2013

What good is a research library inside a research enterprise, or vice-versa?"

Dr. Micah Altman
<http://micahaltman.com>

Director of Research, MIT Libraries
Non-Resident Senior Fellow, Brookings Institution

CC BY SA

Collaborators*

- Jonathan Crabtree, Merce Crosas, Myron Guttman, Gary King, Michael McDonald, Nancy McGovern
- Research Support

Thanks to the Library of Congress, the National Science Foundation, IMLS, the Sloan Foundation, the Joyce Foundation, the Massachusetts Institute of Technology, & Harvard University.

* And co-conspirators

Related Work

Reprints available from: micahaltman.com

- Micah Altman, Michael P Mcdonald (2013) A Half-Century of Virginia Redistricting Battles: Shifting from Rural Malapportionment to Voting Rights to Public Participation. *Richmond Law Review*.
- Micah Altman, Simon Jackman (2011) Nineteen Ways of Looking at Statistical Software, 1-12. In Journal Of Statistical Software 42 (2).
- Micah Altman (2013) Data Citation in The Dataverse Network ®,. In Developing Data Attribution and Citation Practices and Standards: Report from an International Workshop.
- Micah Altman (2012) "Mitigating Threats To Data Quality Throughout the Curation Lifecycle, 1-119. In Curating For Quality.
- Micah Altman, Jonathan Crabtree (2011) Using the SafeArchive System :TRAC-Based Auditing of LOCKSS, 165-170. In Archiving 2011.
- Kevin Novak, Micah Altman, Elana Broch et al. (2011) Communicating Science and Engineering Data in the Information Age. In National Academies Press.
- Micah Altman, Jeff Gill, Michael McDonald (2003) Numerical issues in statistical computing for the social scientist. In John Wiley & Sons.

This Talk

Why now?

What good is a research library in a research enterprise?

What good is a research enterprise in a research library?

Obligatory Disclaimers

Personal Biases: Social/Information Scientist, Software Engineer, Librarian, Archivist

*"It's tough to make predictions, especially about the future!"**

*Attributed to Woody Allen, Yogi Berra, Niels Bohr, Vint Cerf, Winston Churchill, Confucius, Disreali [sic], Freeman Dyson, Cecil B. Demille, Albert Einstein, Enrico Fermi, Edgar R. Fiedler, Bob Fourer, Sam Goldwyn, Allan Lamport, Groucho Marx, Dan Quayle, George Bernard Shaw, Casey Stengel, Will Rogers, M. Taub, Mark Twain, Kerr L. White, etc.

Why now?

"At the risk of stating the obvious, the complex system of relationships and products known as scholarly communication is under considerable pressure."

– Ann J. Wolpert*
Nature 420, 17-18, 2002

* Director, MIT Libraries; Board Chair, MIT Press; my boss

Some General Trends in Scholarship

Lots More Data

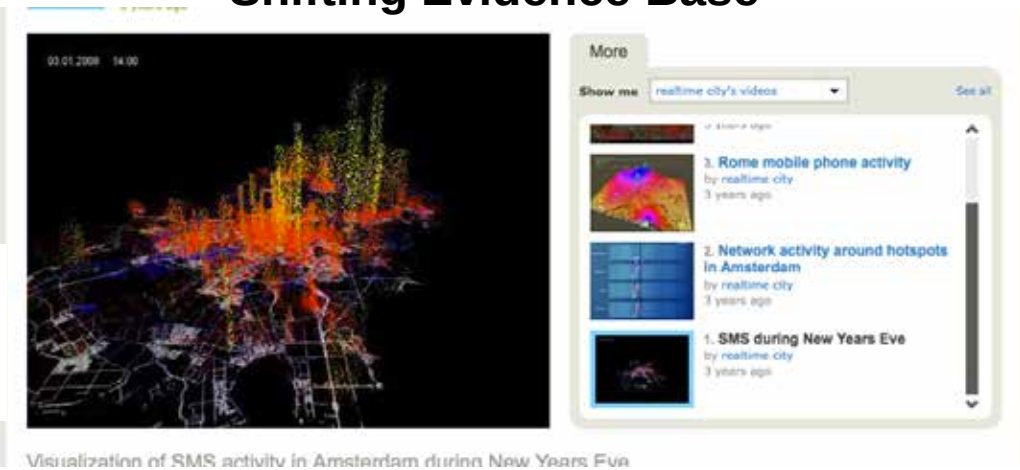
More Open

creative commons

GPL v3 Free Software

Open Knowledge Foundation

Shifting Evidence Base



Visualization of SMS activity in Amsterdam during New Years Eve

High Performance Collaboration (here comes everybody…)

Publish, then Filter

SSRN Social Science Research Network
Tomorrow's Research Today

arXiv.org

PLoS one
accelerating the publication of peer-reviewed science

NBT? … More Everything

Mobile

Forms of publication

Contribution & attribution

Cloud

Open

Publications

Interdisciplinary

Personal data

Mashups

Students

Readers

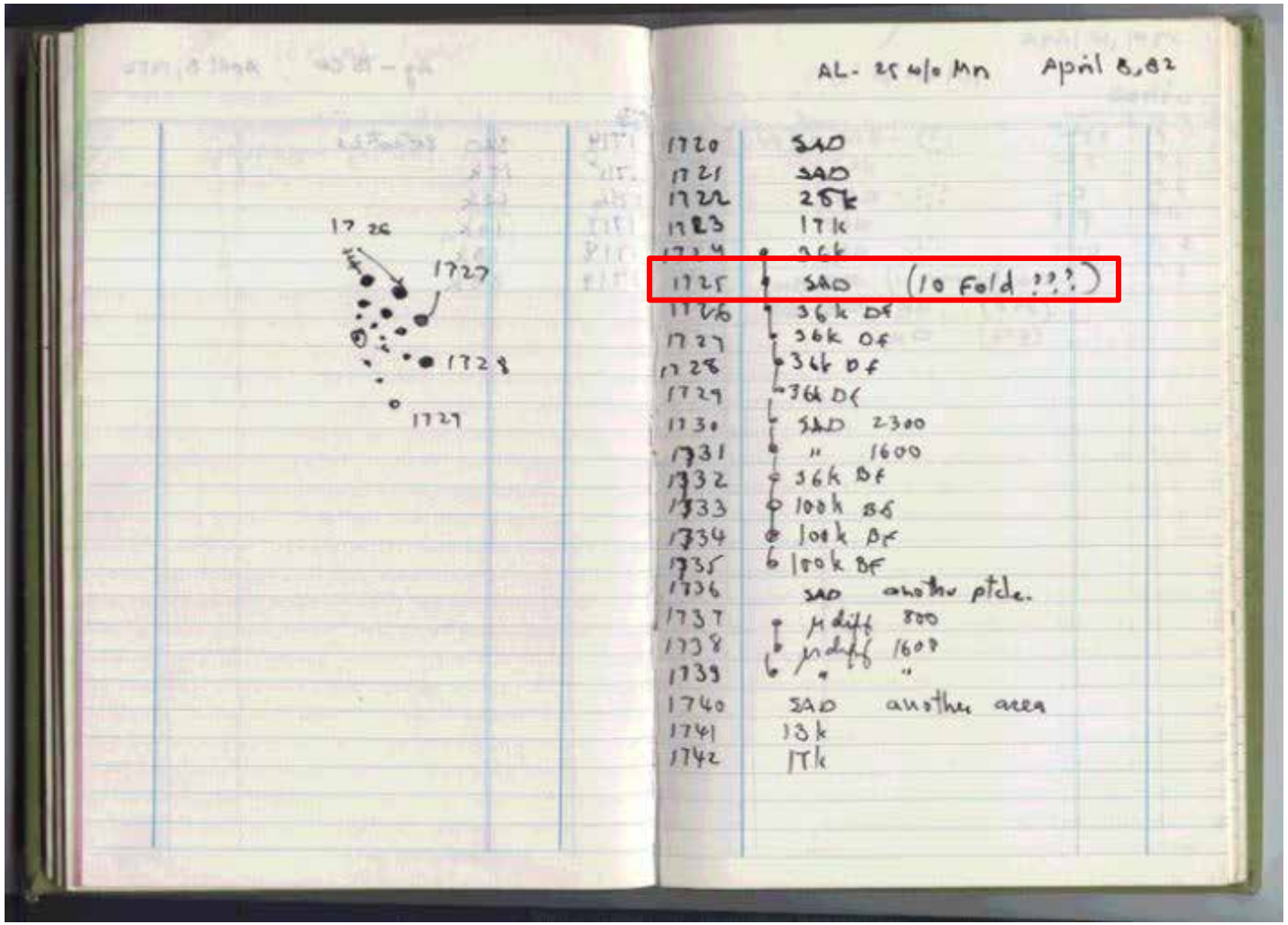
Funders

Everybody

Increasing recognition of problems in current practice…

Unpublished Data Ends up in the "Desk Drawer"

- Null results are less likely to be published
- Outliers are routinely discarded



Daniel Schectman's Lab Notebook Providing Initial Evidence of Quasi Crystals

Increased Retractions, Allegations of Fraud

The New York Times

Fraud Case Seen as a Red Flag for Psychology Research

Dr. Stapel was able to operate for so long, the committee said, in large measure because he was "lord of the data," the only person who saw the experimental evidence that had been gathered (or fabricated). This is a widespread problem in psychology, said Jelte M. Wicherts, a psychologist at the University of Amsterdam. In a recent survey, two-thirds of Dutch research psychologists said they did not make their raw data available for other researchers to see. "This is in violation of ethical rules established in the field," Dr. Wicherts said.

In a survey of more than 2,000 American psychologists scheduled to be published this year, Leslie John of Harvard Business School and two colleagues found that 70 percent had acknowledged, anonymously, to cutting some corners in reporting data. About a third said they had reported an unexpected finding as predicted from the start, and about 1 percent admitted to falsifying data.

Erosion of Evidence Base

- Researchers lack archiving capability
- Incentives for preserving evidence base are weak

Examples

Intentionally Discarded: "Destroyed, in accord with [nonexistent] APA 5-year post-publication rule."

Unintentional Hardware Problems "Some data were collected, but the data file was lost in a technical malfunction."

Acts of Nature The data from the studies were on punched cards that were destroyed in a flood in the department in the early 80s."

Discarded or Lost in a Move "As I retired …. Unfortunately, simply didn't have the room to store these data sets at my house."

Obsolescence "Speech recordings stored on a LISP Machine…, an experimental computer which is long obsolete."

Simply Lost "For all I know, they are on a [University] server, but it has been literally years and years since the research was done, and my files are long gone."

Compliance with Replication Policies is Low

- Compliance is low even in best examples of journals
- Checking compliance manually is tedious

Report on the American Economic Review Data Availability

Table I: Data and code submission results by year of publication

	2006	2007	Mar-08
Articles Published[7]	98	100	22
Articles Subject to Data Policy	61	63	11
Articles Investigated	13	24	2
With Readme File	12	23	1
	(92%)	(96%)	(50%)
With complete submission[8]	7	12	1
	(54%)	(50%)	(50%)
With proprietary data instructions	1	10	0
	(8%)	(42%)	(0%)
Articles Investigated believed replicable without contacting the author(s)	8	22	1
	(62%)	(92%)	(50%)

Got Replicability?

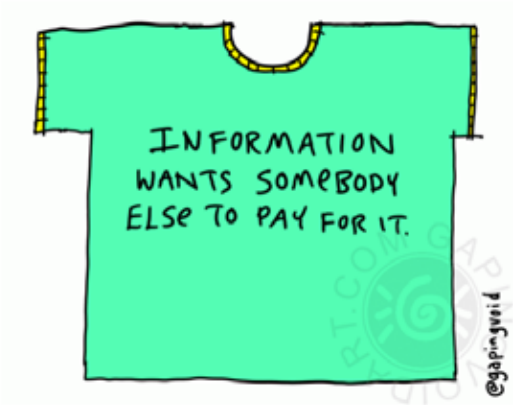
The *Journal of Money, Credit and Banking* Archive

B.D. McCullough.[1]

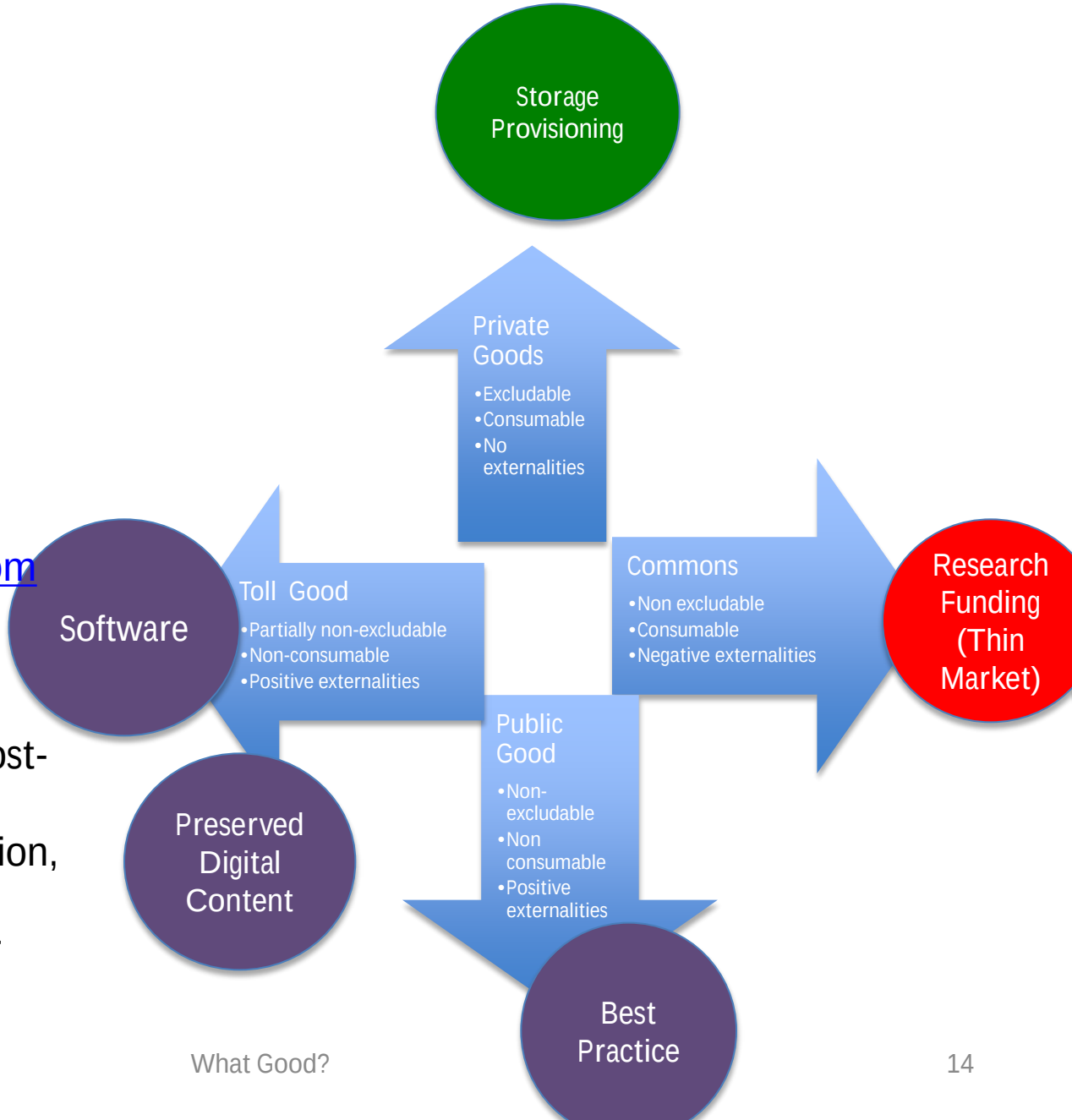
Table 1: Lifetime Compliance

Journal	Empirical articles	Entries	Compliance %
J. App. Econometrics	292	290	99
Fed. St. Louis Review	219	162	74
JMCB	193	66	34
J. Bus. Econ. Statistics	342	121	35

Knowledge is not a Private good



Source: © Hugh Macleod, Gapingvoid Art, gapingvoid.com



- Libraries often operate on “cost-recovery-minus”
- Subsidize knowledge production, long-term access, reuse
- Recognize costs necessary for broader impacts in research budgets

Observations

- Practice of science – researchers, evidence base, and publications are all shifting (often to edges)
- Filtering, replication, integration and reuse are increasing in impact relative to formal publication
- Increasing production and recognition of information assets produced by institution beyond traditional publications
- Planning for access to scholarly record should include planning for long-term access à beyond the life of a single institution
- Important problems in scholarly communications, information science & scholarship increasingly require multi-disciplinary approaches.
- Since knowledge is not a private good à Pure-market approach leads to under-provisioning

What good is a research library in a research enterprise?

Why Now – Library Version

- Physical collections (&size) decreasingly important
- Traditional metrics are decreasingly relevant
- Traditional service demand declining (reference, circulation)
- Rising journal costs
- External competition & disintermediation
- Library staff skills outdated
- Library space targeted

Why Now – Library Version @MIT

Over last 5 years, major internal efforts, including…

- Reorganization along ‘functional’ lines
- Increase in systematic institutional evaluation
- Pro-active/coordinated faculty liaison program
- Implementation of institutional OA mandate

Why Now – Trends @MIT*

Undergraduate Student	2005	270	12.7%
	2008	421	18.6%
	2011	182	9.4%
Graduate Student	2005	154	5.2%
	2008	461	15.2%
	2011	206	7.5%
Faculty	2005	29	9.9%
	2008	55	19.6%
	2011	57	19.6%
Other Research & Academic Staff	2005	164	21.7%
	2008	428	32.3%
	2011	269	24.5%
PostDoc	2005	49	15.0%
	2008	131	27.2%
	2011	129	20.7%
Overall	2005	666	10.3%
	2008	1496	20.3%
	2011	843	12.6%

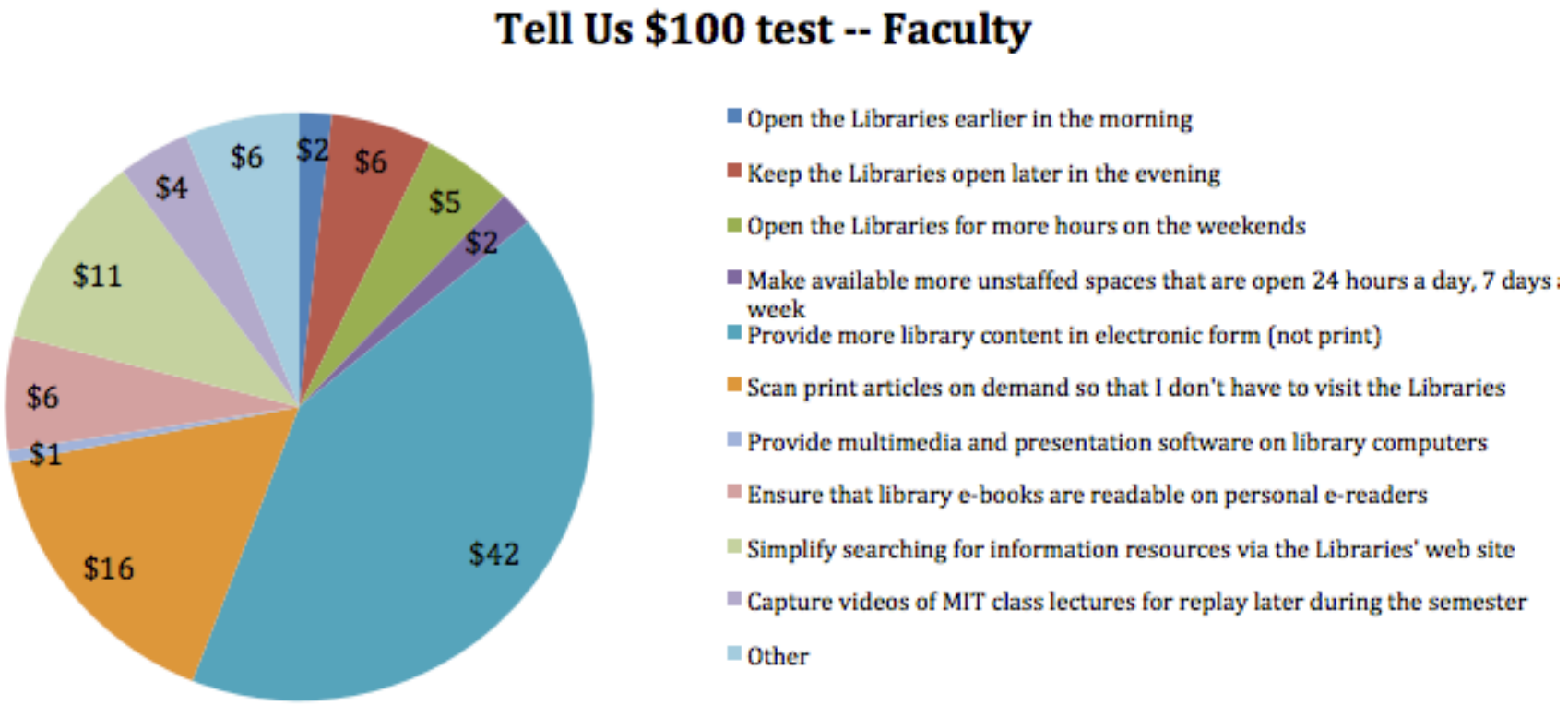
% of Community who Never Sets Foot in a Library Space

Undergraduate Student	2005	82.2%
	2008	84.7%
	2011	78.9%
Graduate Student	2005	89.4%
	2008	88.5%
	2011	80.8%
Faculty	2005	84.9%
	2008	87.3%
	2011	84.8%
Other Research & Academic Staff	2005	85.3%
	2008	81.2%
	2011	77.9%
PostDoc	2005	89.2%
	2008	92.0%
	2011	80.1%
Overall	2005	86.4%
	2008	86.2%
	2011	79.9%

% of Users Satisfied/Very Satisfied

*Source:
http://libguides.mit.edu/content.php?pid=286364&sid=2381371
Warning, typical low response rates among key sub-populations

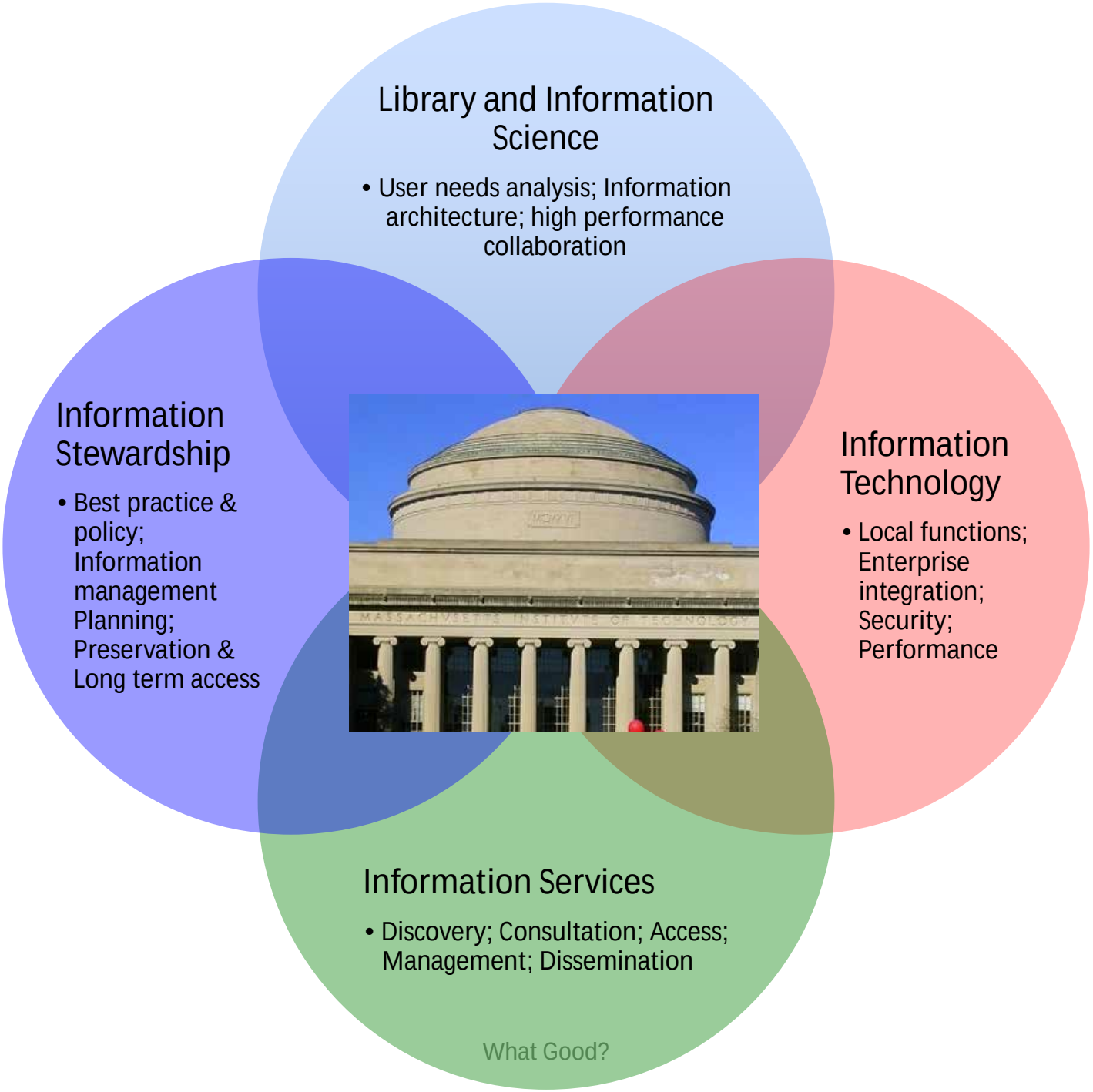
Why Now – Trends @MIT*



Why Now – Trends @MIT*

- About 20–25% of Postdocs, Faculty, and Research staff *never* set foot in the a library-managed space
- Levels of overall satisfaction high, but decreasing
- Highest priority for faculty is access to more digital and scan-on-demand material
- Faculty participation in evaluation surveys low…

Library Services -- It's Complicated!



Observations

- Since knowledge is not a private good à Pure-market approach leads to under-provisioning
- Need coherent economic models, business models, and infrastructure (policy, procedure, technology) for valuing, selecting, managing, disseminating durable information assets

How could libraries contribute to the future of scholarship?

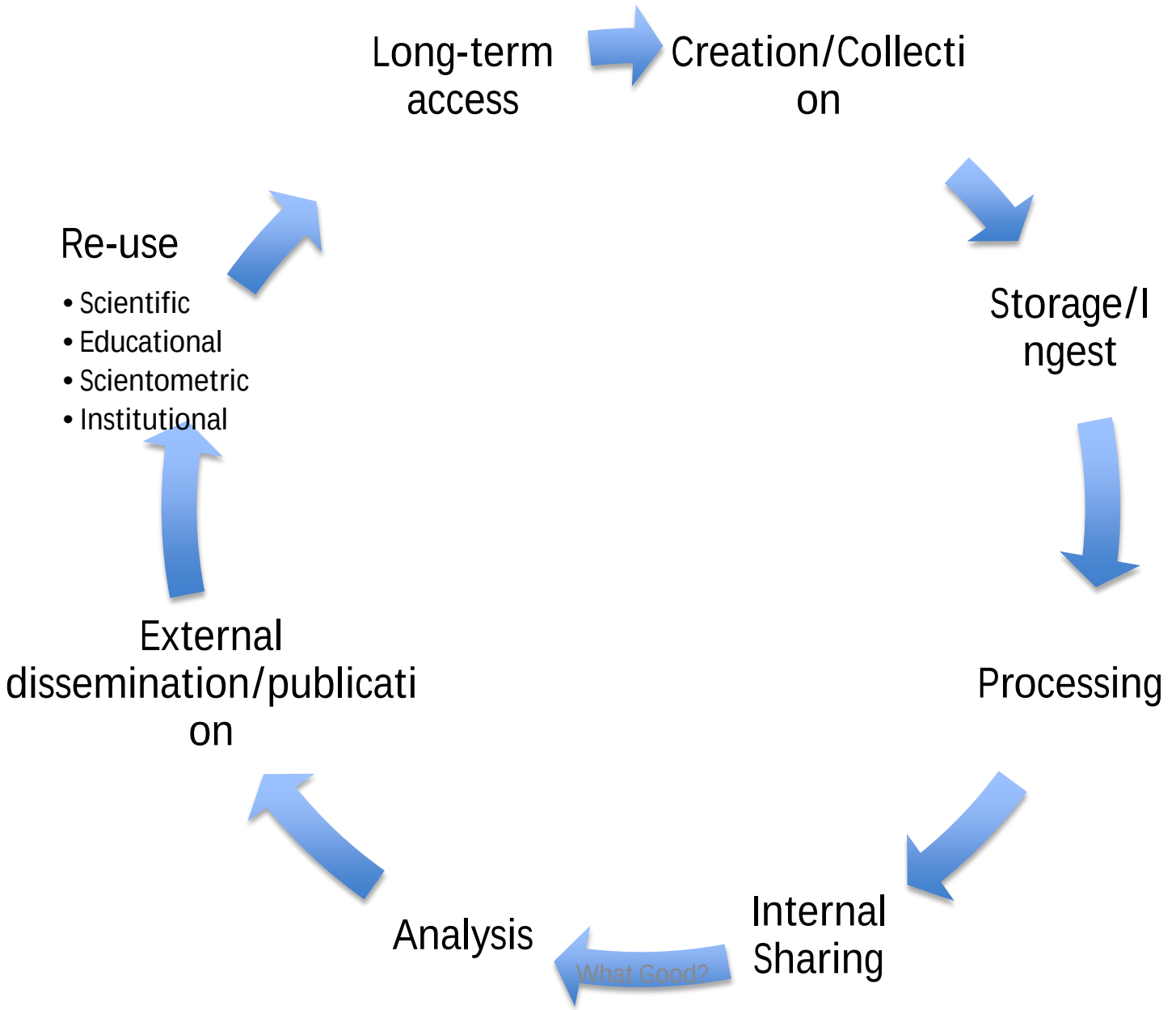
Library Core Competencies

- Information stewardship
  - View information as durable assets
  - Manage information across multiple lifecycle stages
- Information management lifecycle
  - Metadata
  - Information organization & architecture
  - Processes
- Spans disciplines
  - Inter-disciplinary discovery
  - Multi-disciplinary access
- Service
  - Models of user need
  - Culture of service
- Trust
  - Library is trusted as service
  - Library is trusted as honest broker

Library Core Values

- Long term view
  - Universities are long-lived institutions
  - Many actors in universities engaged with scholarly communication/record focused on short term incentives
  - Libraries culture, values, perspective weigh heavily long-term analysis and responsibilities
- Information is for use
  - “Every scholar her data; Every datum her scholar.”
- Service
  - “Save the time of the scholar.”
- Growth
  - “The library is a growing organism.”

Information Lifecycle



Why IDs? Why Now?

Core Requirements for Community Information Infrastructure

- Stakeholder incentives
  - recognition; citation; payment; compliance; services
- Dissemination
  - access to metadata; documentation; data
- Access control
  - authentication; authorization; rights management
- Provenance
  - chain of control; verification of metadata, bits, semantic content
- Persistence
  - bits; semantic content; use
- Legal protection & compliance
  - rights management; consent; record keeping; auditing
- Usability for…
  - discovery; deposit; curation; administration; annotation; collaboration
- Business model
- Trust model

See: King 2007; ICSU 2004; NSB 2005; Schneier 2011

*plus ça change, plus c'est la même folie**

- Budget constraints
- Invisibility of infrastructure
- Organizational biases
- Cognitive biases
- Inter- and intra- organizational trust
- Discount rates and limited time-horizons
- Deadlines
- Challenging in matching skillsets & problems
- Legacy systems & requirements
- Personalities
- Bureaucracy
- Politics

Observations

- Need to develop coherent economic models, business models, and infrastructure (policy, procedure, technology) for valuing, selecting, managing, disseminating durable information assets
- Library core institutional values align well with future needs of research instution
- Need to reframe library culture around core institutional values in context of new patterns of knowledge productions and institutions; and retool processes and infrastructure
- Need to move from pure service to service plus collaboration
- This will not be easy…

What good is a research enterprise in a research library?

Theory

- Future-aware planning, incorporating the best-of-class research findings in information science, data science, and other fields into our policies, planning and practices.
- Identify, gain recognition for, and generalize the innovations that the scholars in the university make to solve their own problems or to advance the information commons.
- Collaborate with researchers in the university to develop innovative approaches to managing research data and research outputs.
- Amplify the impact that university can have on the development of information science, information policy, and scholarly communication through participation the development of standards, policy, and methods related to information science and information management.
- Solve emerging problems in information management that are essential to support new and innovative services.

Personal Examples

"In theory, theory and practice are the same. In practice, they differ."

Future-Aware Library Planning

Example: (Potentially) Everything?

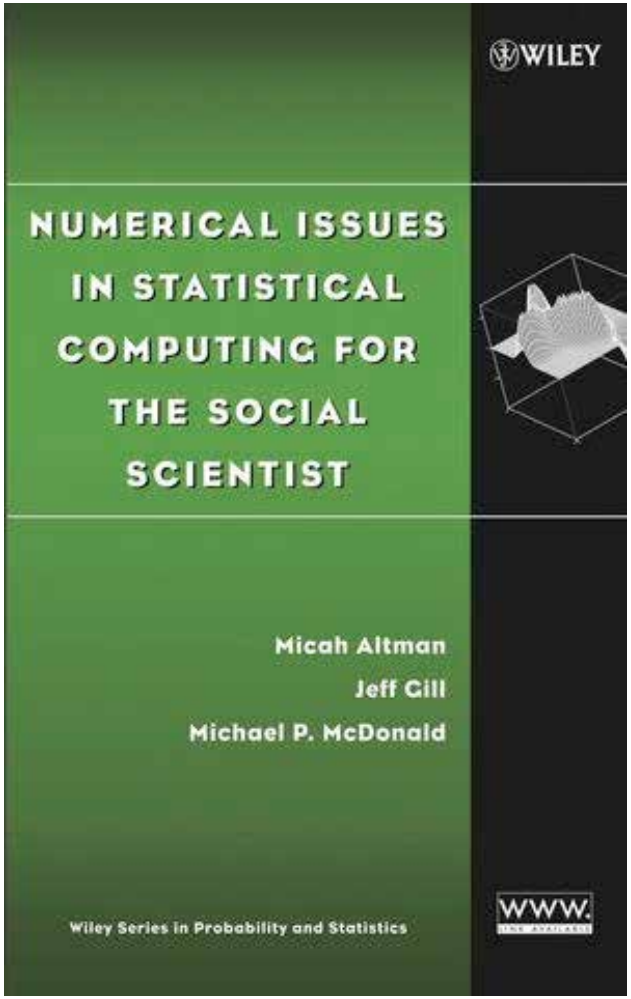
- Web scale discovery
- Social search
- Recommendation systems
- Discovery Personalization
- Bibliographic information visualization
- Research data management
- Long-term digital preservation cost models
- Selection policies
- MOOC content
- Information Annotation
- Library analytics
- Long term storage reliability
- ...

Generalize Local Innovations

Example: Semantic Fingerprints

Solving a Different Problem: Reliability of Statistical Computation

- Original goal:
  analyze robustness of social science statistical models to computational implementation
- Proximate goal:
  replicate high-profile published studies
- Discovered goal:
  verify semantic interpretation of data by statistical software



Universal Numeric Fingerprint
- Now incorporated in the Dataverse Network

Data Set File

Metadata and Format Information

1	4	4	21	···	121
1	2	2	91	···	212
1	9	2	72	···	104
0	2	2	2	···	321
1	6	2	12	···	204
1	9	4	52	···	311
0	3	2	23	···	92
0	2	5	91	···	212
0	5	8	91	···	91
1	9	1	72	···	104
⋮	⋮	⋮	⋮	⋱	⋮
1	2	2	91	···	212

1. Extract variable name, description and summary statistics
2. Convert data set to a preservation format, independent of the software package
3. Apply a cryptographic algorithm to canonical format
4. Get alphanumeric string based on semantic contents of the digital object:
   - uniquely summarizes the contents,
   - but does not convey its information

UNF:5:EKgHvTNfkkS86dNzABIhNw==

Change content → Changes UNF

Change format → Doesn't change UNF

Some Possible Perils of Redistricting

- "In summary, elimination of gerrymandering would seem to require the establishment of an automatic and impersonal procedure for carrying out a redistricting. It appears to be not at all difficult to devise rules for doing this which will produce results not markedly inferior to those which would be arrived at by a genuinely disinterested commission." - [Vickrey 1961]

Crowd-Source it!

DistrictBuilder

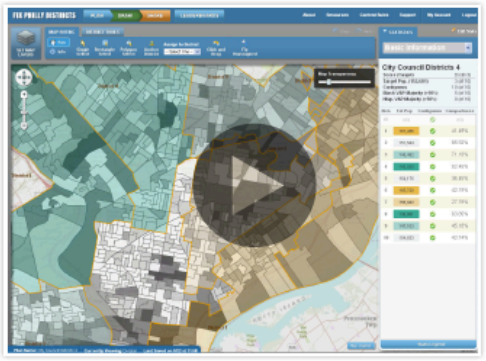
DistrictBuilder is web-based, open source software for collaborative redistricting.

Customizable open source software built by experts

DistrictBuilder was developed in collaboration with leading redistricting experts at the Public Mapping Project. Whether you work for a state or local government agency, an advocacy group or you are a legislator, DistrictBuilder has been designed to provide you with affordable, easy-to-use, customizable redistricting tools that make the redistricting process more open and collaborative across agencies and with the public.

Main Features

- Create and edit district plans
- Use template plans to get started faster
- Import and merge plans from other systems
- Display demographics, election and other data
- Integrate with GoogleMaps, Esri ArcGIS Online, OpenStreetMap or Bing maps
- Show additional reference map layers, like school districts and administrative boundaries
- Automatically calculate contiguity, compactness and population statistics as you create your plan
- Customize demographic, geographic and election data statistics on-the-fly as you build your plan
- Find unassigned areas
- Draw communities of interest and evaluate a plan against them
- Evaluate how closely your proposed plan matches legal requirements
- Save and share your plans via a URL link
- Support public competitions, scoring and leaderboards



Open source license means it's both affordable and part of a community

Unlike proprietary solutions, DistrictBuilder is open source so you don't have to pay a license fee. If your team has the expertise, you can download the source code and build your own redistricting application. Or our dedicated DistrictBuilder implementation team can work hand-in-hand with you to create the perfect application for your needs. And if we built a new feature for your application, it is immediately contributed back to the community of users. The more you add, the more we give back.

System Requirements

The software is built using several open source technologies including:

- Django
- Celery
- PostgreSQL
- GeoServer
- jQuery
- PostGIS

The software can be hosted in several environments:

GET STARTED NOW
Download on GitHub

Services

INSTRUCTIONS AND DOCUMENTATION

General information »
User Manual »
Admin Manual »
Software Documentation »
Software Notice »
Loading an Amazon EC2 AMI Instance »

SUPPORT

Report bugs on GitHub »

IMPLEMENTATION

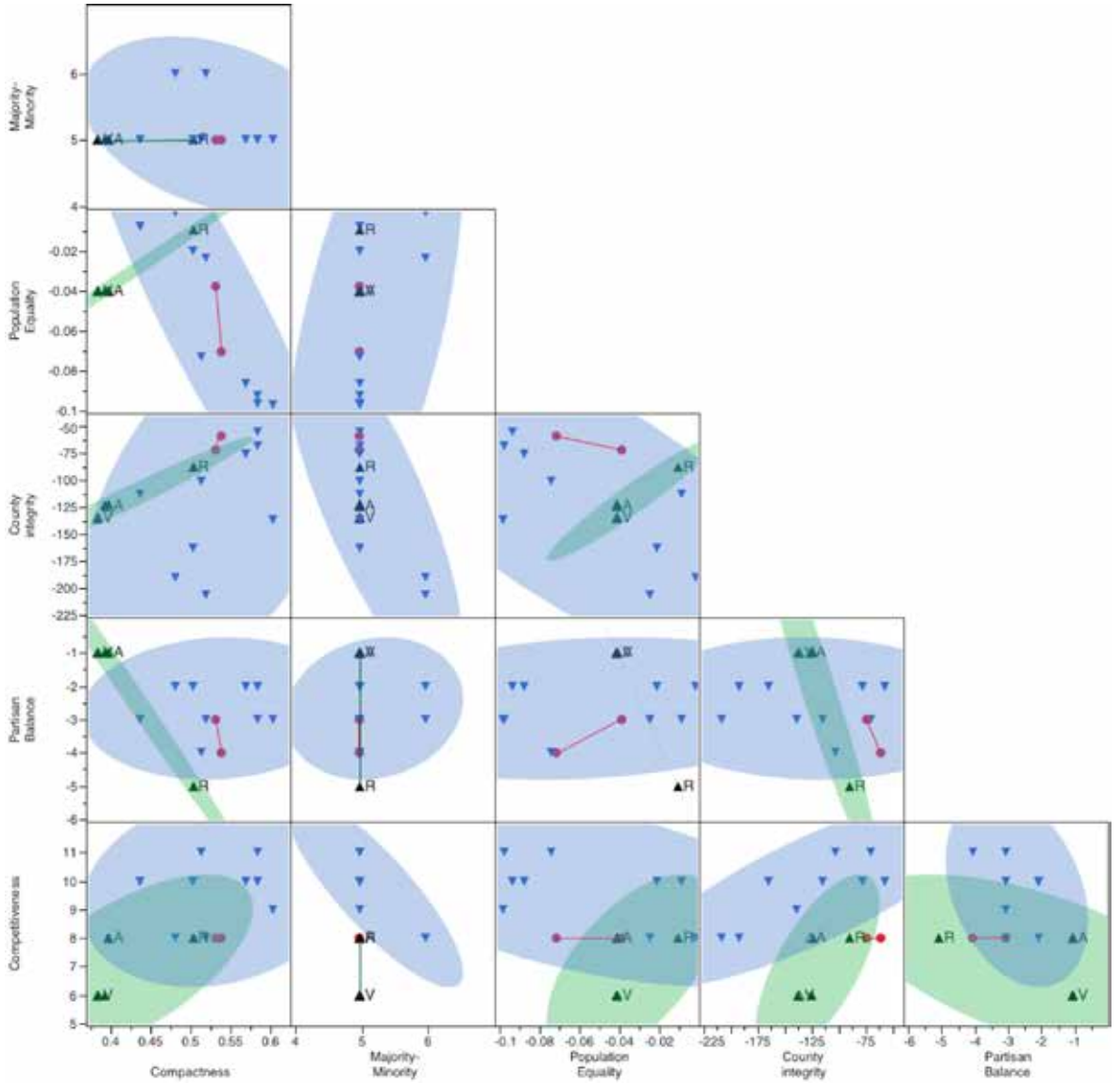
Open source code base can be daunting if you do not have the technical expertise on your team to build the application. Implementation usually requires:

- Database development
- Software configuration
- Graphic design
- Competition setup
- Software customization
- Training
- Hosting

Our software partner, Azavea,

Can Students Draw Better Election Maps than Professional Politicians?

- Yes,
- at least in Virginia,...
- Now analyzing data from many other states...



Collaborate with University Researchers Around Information Management

Example: Privacy Tools for Sharing Research Data

HARVARD
School of Engineering
and Applied Sciences

PRIVACY TOOLS FOR SHARING RESEARCH DATA

A NATIONAL SCIENCE FOUNDATION SECURE AND TRUSTWORTHY CYBERSPACE PROJECT

Home

Principal Investigators

Project Description

Press Release

Positions

With NSF grant, researchers will enhance technologies and policies to protect personal data used in research studies

This project is a broad, multidisciplinary effort to help enable the collection, analysis, and sharing of personal data for research in social science and other fields while providing privacy for individual subjects.

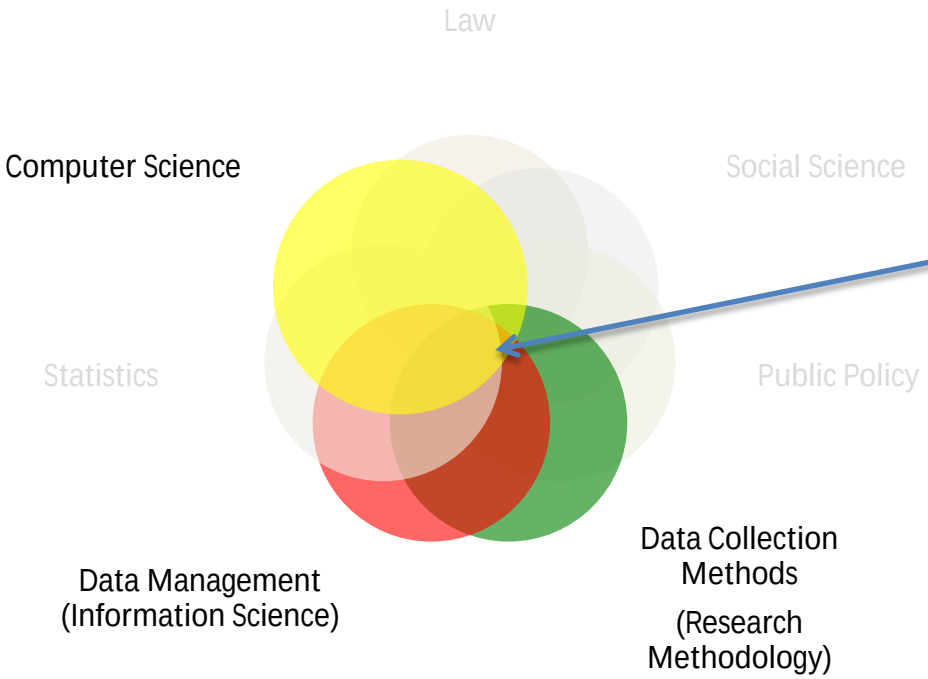
It is a collaborative effort between the Center for Research on Computation and Society, the Institute for Quantitative Social Science, the Berkman Center for Internet and Society, and the Data Privacy Lab.

It received seed funding from Google, and will now be supported primarily as a Frontier project in the NSF Secure and Trustworthy Cyberspace Program.

Send this — Print this

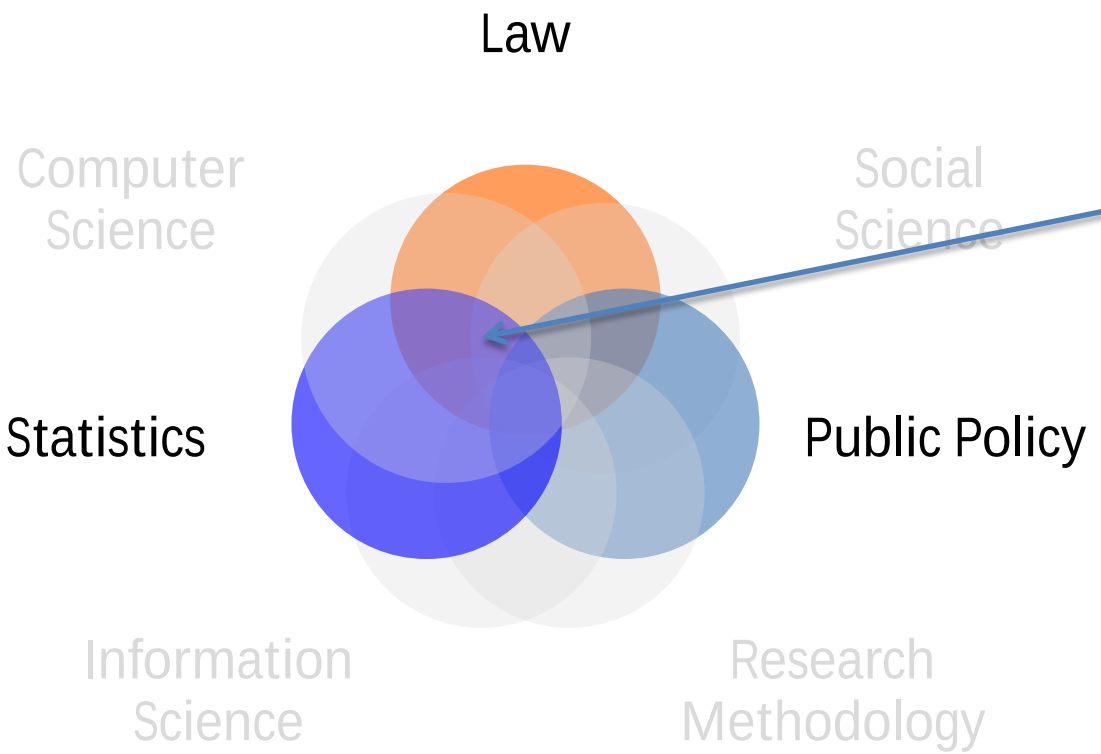
Harvard School of Engineering and Applied Sciences | Copyright © by the President and Fellows of Harvard College

Research at the Intersection of Research Methods, Computer Science, Information Science



- Privacy-aware data-management systems
- Methods for confidential data collection and management

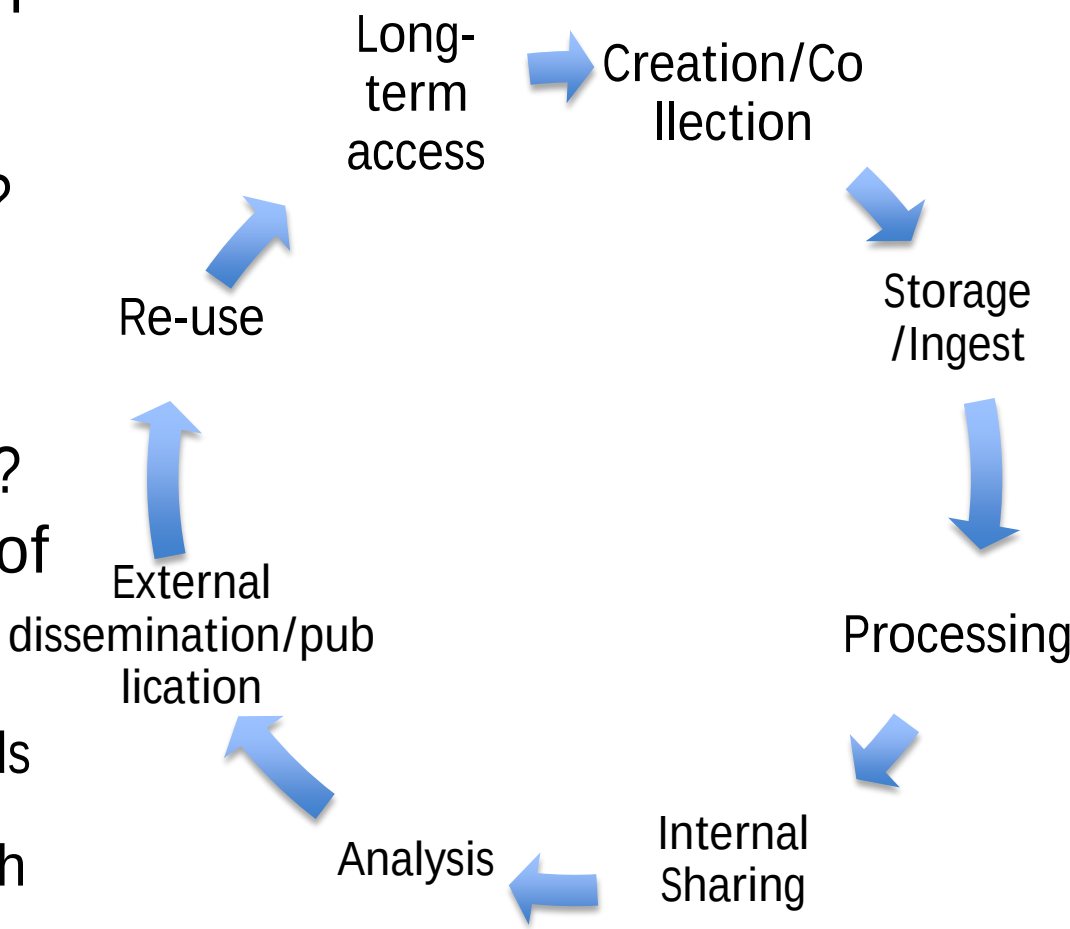
Research at the Information Science, Research Methodology, Policy



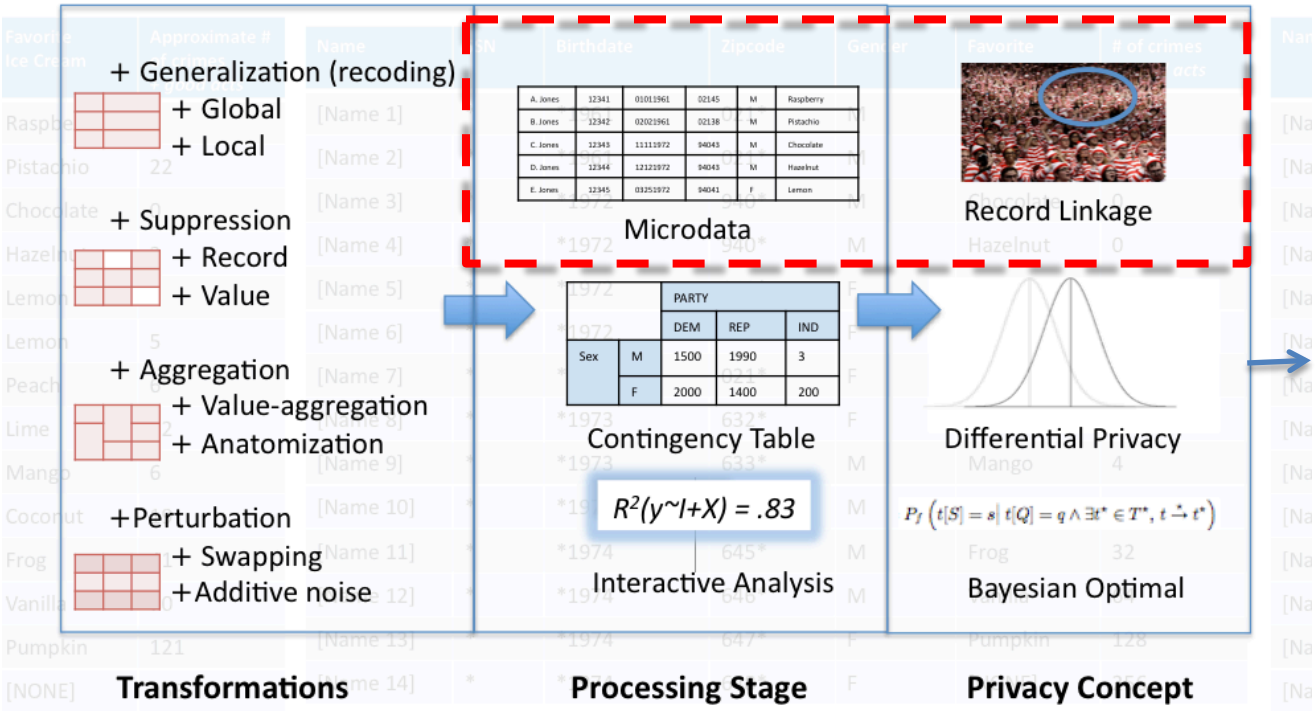
- Creative-Commons-like modular license plugins for privacy
- Standard privacy terms of service; consent terms
- Model legislation – for modern privacy concepts

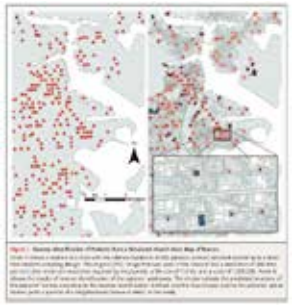
Framework – Information Life Cycle

- Which laws apply to each stage of lifecycle…
- Are legal requirements consistent across stages?
- How to align legal instruments: consent forms, SLA, DUA's to ensure legal consistency?
- Harmonizing protection of privacy in research methods:
  - - Data collection methods for "sensitive data" collection consistent with privacy concept at other stages?



Model 1 – Input->Output





Name	SSN	Birthdate	Zipcode
* Jones	*	* 1961	021*
* Jones	*	* 1961	021*
* Jones	*	* 1972	9404*
* Jones	*	* 1972	9404*
* Jones	*	* 1972	9404*
* Jones	*	*	021*
* Jones	*	*	021*
* Smith	*	* 1973	63*
* Smith	*	* 1973	63*
* Smith	*	* 1973	63*
* Smith	*	* 1974	64*
* Smith	*	* 1974	64*
* Smith	*	04041974	64*
* Smith	*	04041974	64*

Published Outputs

Some Privacy Concepts Not Well Captured in Law

- ~~Deterministic record linkage~~
- Probabilistic record linkage (reidentification probability)
- K-anonymity
- K-anonymity + heterogeneity
  - Learning theory– distributional privacy [Blum, et. al 2008]
- Threat & vulnerability analysis
- Differential privacy
- Bayesian privacy

Some Potential Research Outputs

- Analysis of privacy concepts in laws
  - For identification
  - Anonymity
  - Discrimination
- Model language for
  - Legislation
  - Regulation
  - License plugins
- Systems/policy analysis
  - Incentives generated by privacy concepts
  - Incentives aligned with privacy concepts
  - Model privacy from game theoretic/social choice & policy analysis point of view
- Data sharing infrastructure needed for managing confidentiality effectively:
  - Applying interactive privacy automatically
  - Implementing limited data use agreements
  - Managing access & logging – virtual enclave
  - Providing chokepoint for human auditing of results
  - Providing systems auditing, vulnerability & threat assessment
  - Ideally:
    - Research design information automatically fed into disclosure control parameterization
    - Consent documentation automatically integrated with disclosure policies, enforced by system

Influence Emerging Information Policies, Practices, and Standards

Examples: ANPRM, ORCID, Data Citation

IRB Treatment of Confidential Information

ANPRM for Revision to Common Rule

Information Related to Advanced Notice of Proposed Rulemaking (ANPRM) for Revisions to the Common Rule

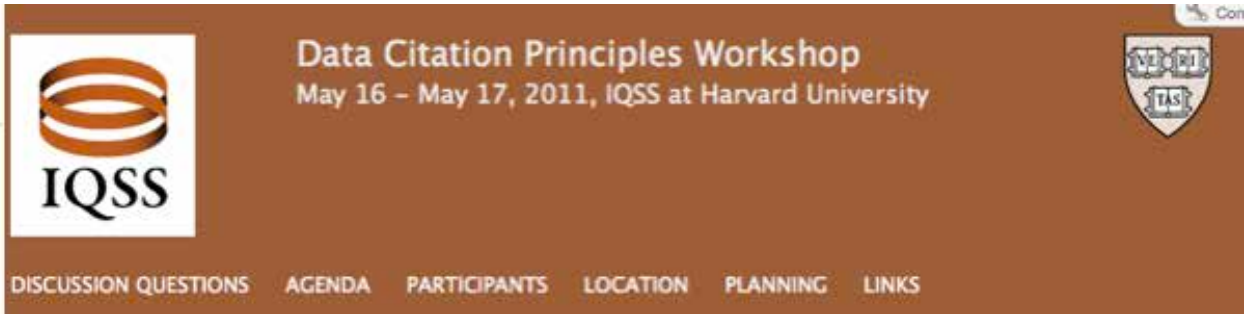
- Read the ANPRM online. (PDF 286KB)
  - Read the Federal Register Notice Extending the Public Comment Period (PDF 140KB)
- Access the July 22, 2011 Press Release describing the ANPRM.
- Frequently Asked Questions regarding the ANPRM.
- How to browse comments that have been submitted regarding the ANPRM.
- Read a table comparing existing regulation with changes in the ANPRM. (PDF KB)

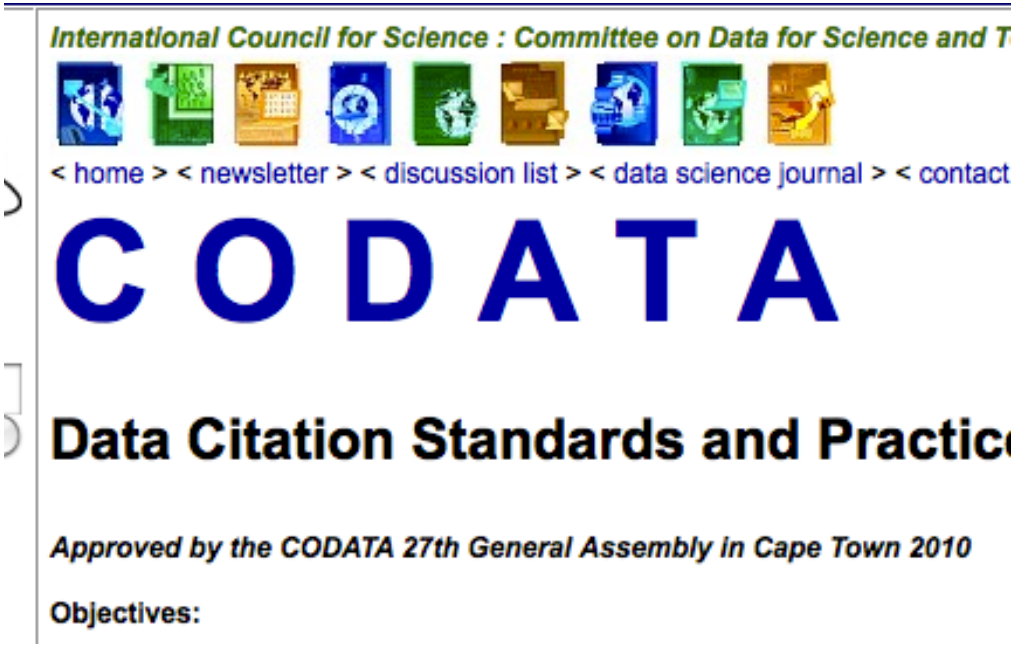
NOTE: The ANPRM comment period has closed .

Help Spread The Word

Grab the badge and URL below for use on your website during the public comment period.

www.hhs.gov/ohrp/humansubjects


IQSS
Data Citation Principles Workshop
May 16 – May 17, 2011, IQSS at Harvard University
DISCUSSION QUESTIONS
AGENDA
PARTICIPANTS
LOCATION
PLANNING
LINKS


International Council for Science : Committee on Data for Science and T
< home > < newsletter > < discussion list > < data science journal > < contact
CODATA
Data Citation Standards and Practic
Approved by the CODATA 27th General Assembly in Cape Town 2010
Objectives:

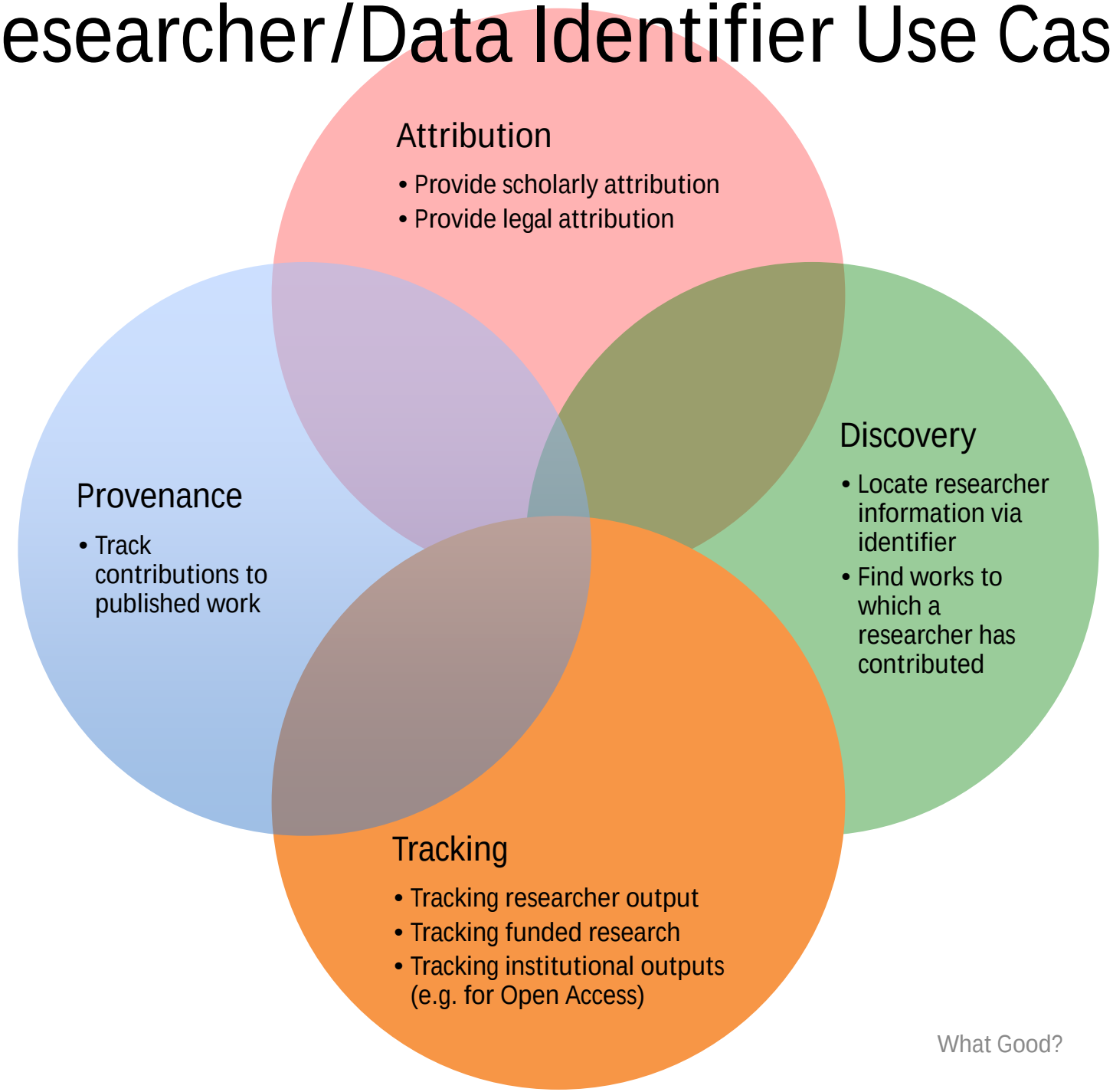
ORCID

ORCID aims to solve the author/contributor name ambiguity problem in scholarly communications by creating a central registry of unique identifiers for individual researchers and an open and transparent linking mechanism between ORCID and other current author ID schemes. These identifiers, and the relationships among them, can be linked to the researcher's output to enhance the scientific discovery process and to improve the efficiency of research funding and collaboration within the research community.

orcid.org

Researcher/Data Identifier Use Cases

Why IDs? Why Now?



Solve emerging problems to support new services

Example: SafeArchive: Collaborative Preservation Auditing

The Problem

"Preservation was once an obscure backroom operation of interest chiefly to conservators and archivists: it is now widely recognized as one of the most important elements of a functional and enduring cyberinfrastructure."
– *[Unsworth et al., 2006]*

"

- Institutions hold digital assets they wish to preserve, many unique
- Many of these assets are not replicated at all
- Even when institutions keep multiple backups offsite,
  many single points of failure remain,
  because replicas are managed by single institution

Potential Nexuses for Long-Term Access Failure

- Technical
  - Media failure: storage conditions, media characteristics
  - Format obsolescence
  - Preservation infrastructure software failure
  - Storage infrastructure software failure
  - Storage infrastructure hardware failure
- External Threats to Institutions
  - Third party attacks
  - Institutional funding
  - Change in legal regimes
- Quis custodiet ipsos custodes?
  - Unintentional curatorial modification
  - Loss of institutional knowledge & skills
  - Intentional curatorial de-accessioning
  - Change in institutional mission

Source: Reich & Rosenthal 2005

Enhancing Reliability through Trust Engineering

- Incentives:
  - Rewards, penalties
  - Incentive-compatible mechanisms
- Modeling and analysis:
  - Statistical quality control & **reliability estimation**, **threat-modeling and vulnerability assessment**
- Portfolio Theory:
  - **Diversification (financial, legal, technical, institutional …)**
  - Hedging
- Over-engineering approaches:
  - Safety margin, **redundancy**
- Informational approaches:
  - **Transparency (release of information permitting direct evaluation of compliance)**; common knowledge,
  - Crypto: signatures, fingerprints, non-repudiation
- Social engineering
  - Recognized practices; shared norms
  - Social evidence
  - Reduce provocations
  - Remove excuses
- Regulatory approaches
  - Disclosure; Review; Certification; **Audits**
  - Regulations & penalties
- Security engineering
  - Increase effort for attacker: harden target (reduce vulnerability); increase technical/procedural controls; , remove/conceal targets
  - Increase risk to attacker: surveillance, detection, likelihood of response
  - Reduce reward: deny benefits, disrupt markets, identify property

Audit [aw-dit]:

An independent evaluation of records and activities to assess a system of controls

Fixity mitigates risk *only if used for auditing*.

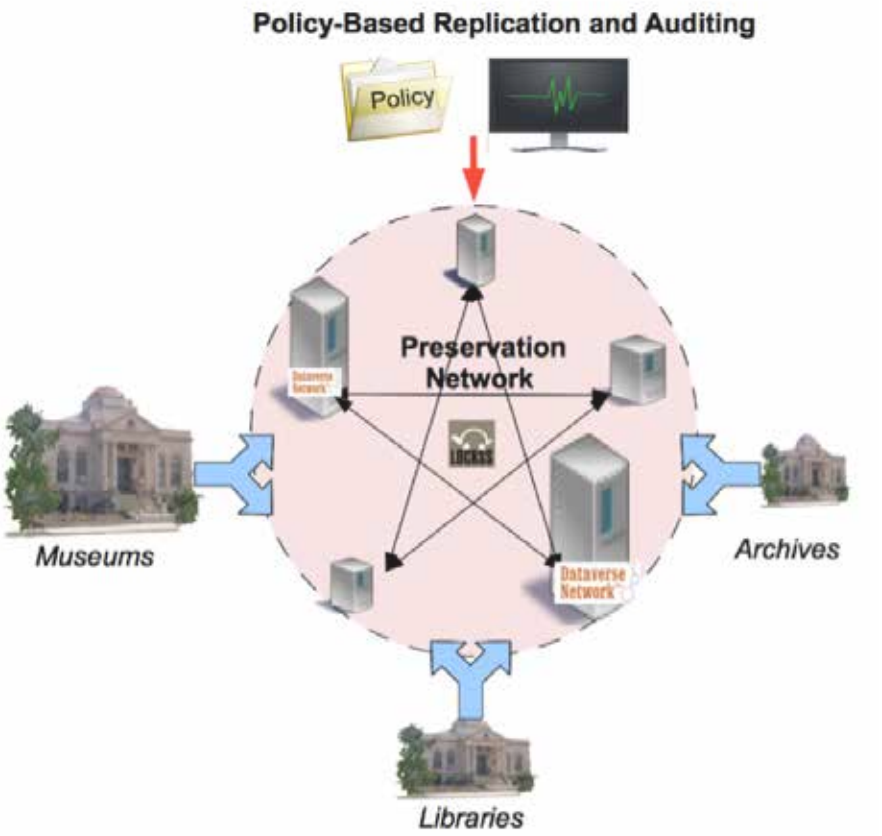
Summary of Current Automated Preservation Auditing Strategies

LOCKSS	*Automated; **decentralized** (peer-2-peer); **tamper-resistant** auditing & repair; for **collection integrity.***
iRODS	*Automated centralized/**federated** auditing for **collection integrity; micro-policies.***
DuraCloud	*Automated; centralized auditing; for **file integrity.** (Manual repair by DuraSpace staff available as commercial service if using multiple cloud providers.)*
Digital Preservation Mechanism	*In development…*
SafeArchive	*Automated; **independent**; **multi-centered**; auditing, repair and **provisioning; of existing LOCKSS storage networks;** for collection integrity**, for high-level policy (e.g. TRAC) compliance.***

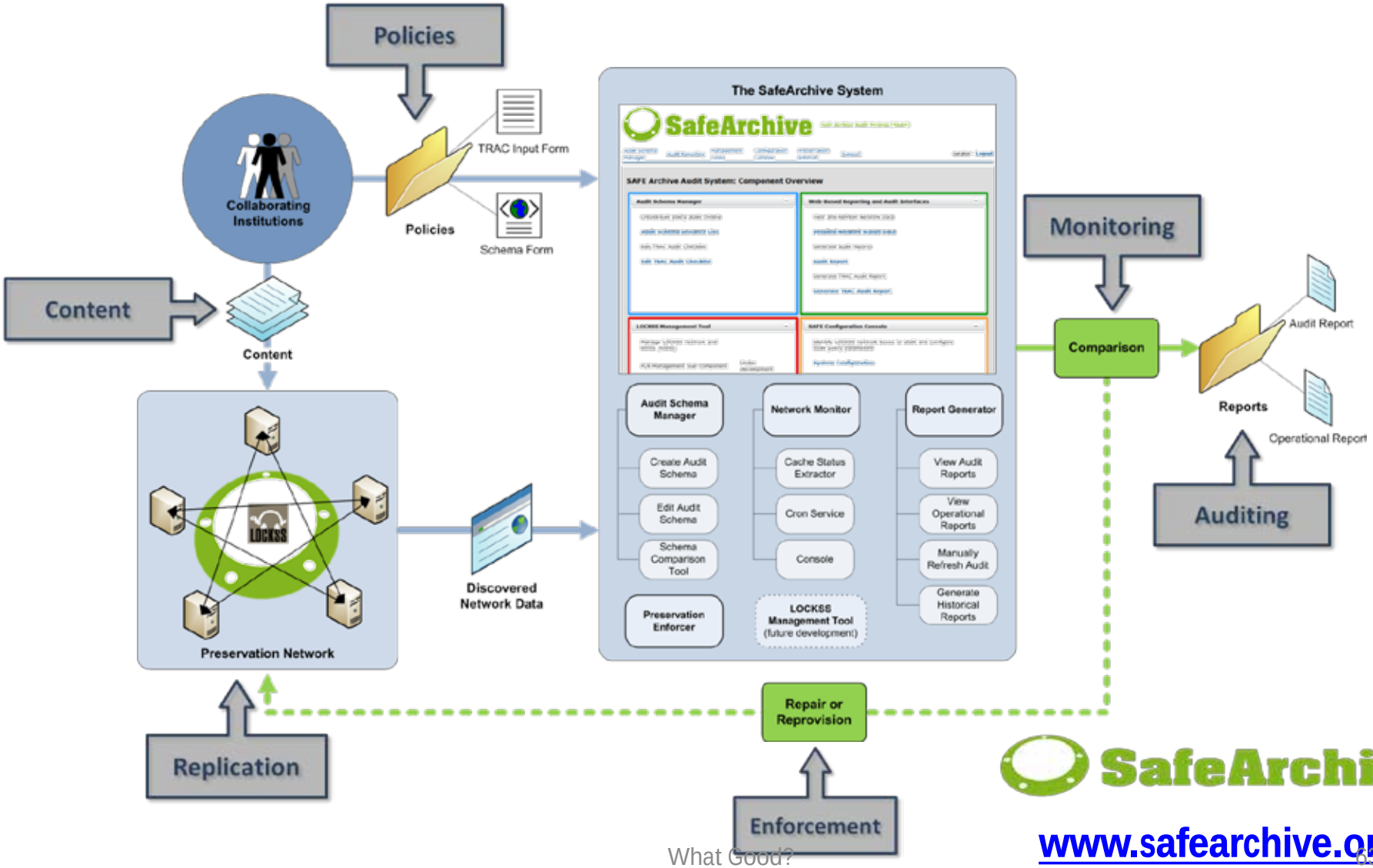
SafeArchive:
TRAC-Based Auditing & Management of Distributed Digital Preservation

Facilitating collaborative replication and preservation with technology…

- *Collaborators* declare explicit non-uniform resource commitments
- *Policy* records commitments, storage network properties
- *Storage layer* provides replication, integrity, freshness, versioning
- *SafeArchive software* provides monitoring, auditing, and provisioning
- *Content* is harvested through HTTP (LOCKSS) or OAI-PMH
- *Integration of* LOCKSS, The Dataverse Network, TRAC



Implementation

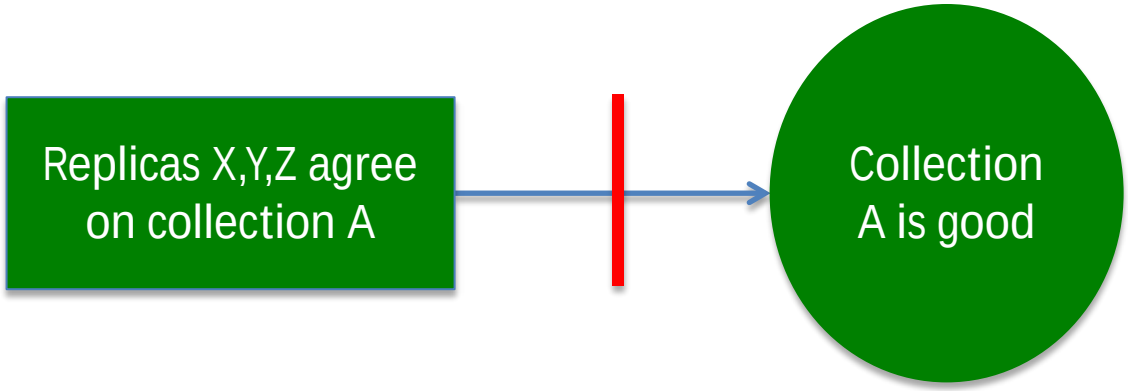


Lesson 1:
Replication agreement does **not** prove collection integrity

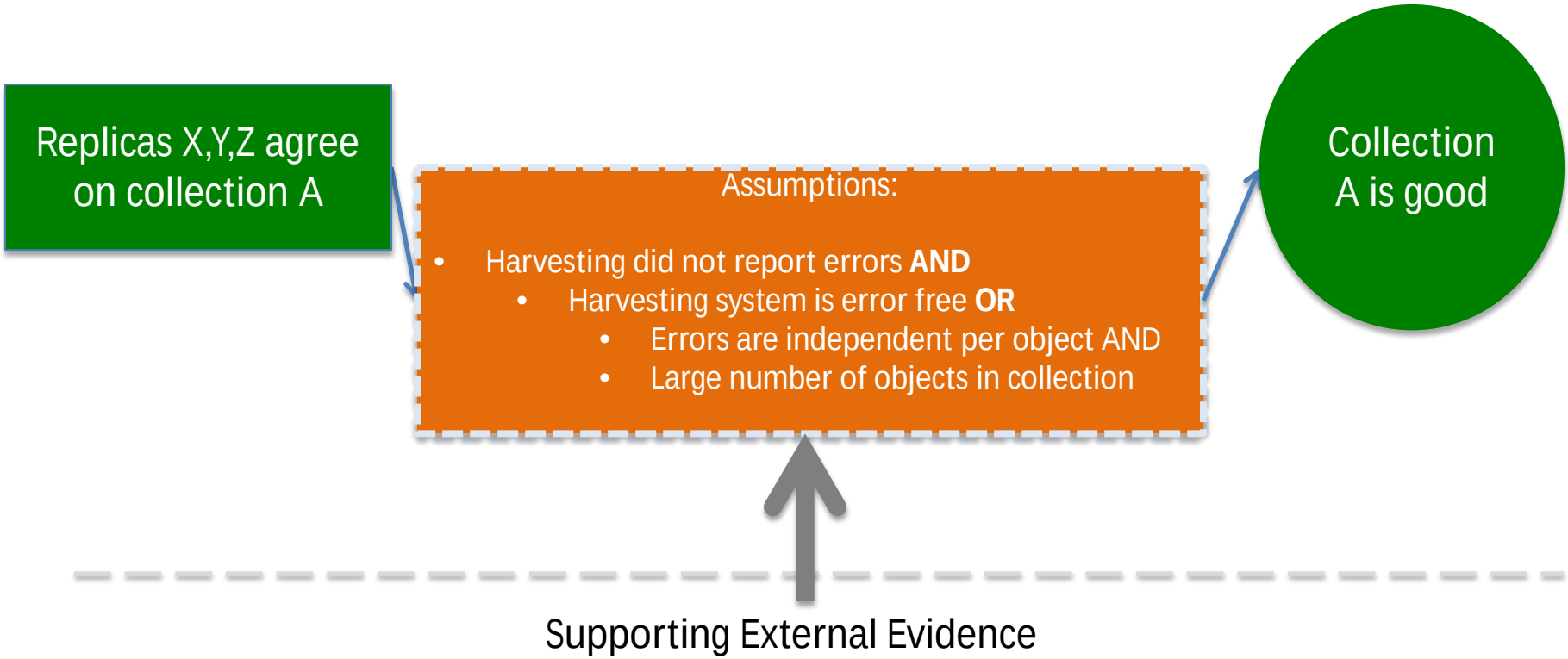
What you see

Replicas X,Y,Z agree on collection A

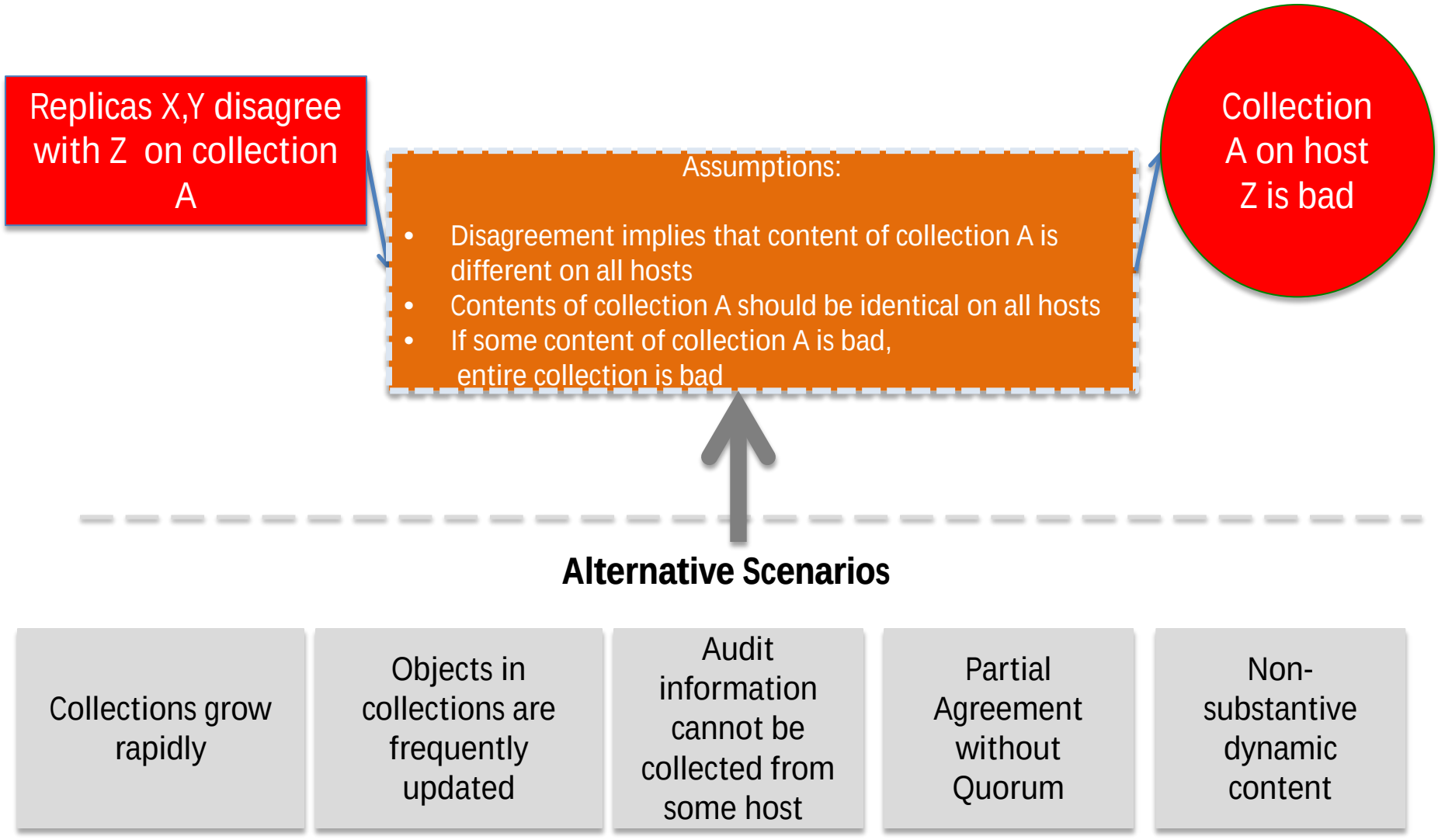
What you are **tempted** to conclude:



What can you infer from replication agreement?



What can you infer from replication failure?



What else could be wrong?

Round 1 hypothesis

Disagreement is real, but doesn't matter in long run

1.1 Temporary differences. Collections *temporarily* out or sync
(either missing objects or different object versions) – will resolve over time
(E.g. if harvest frequency << source update frequency, but harvest times across boxes vary significantly)

1.2 Permanent point-in-time collection differences that are artefact of synchronization.
(E.g. if one replica always has version n-1, at time of poll)

Hypothesis 2**: Disagreement is real, but nonsubstantive.**

2.1.Non-Substantive collection differences (arising from dynamic elements in collection that have no bearing on the substantive content)

2.1.1 *Individual URLS/files that are dynamic and non substantive* (e.g., logo images, plugins, Twitter feeds, etc.) cause content changes (this is common in the GLN).

2.2.2 dynamic content embedded in substantive content (e.g. a customized per-client header page embedded in the pdf for a journal article)

2.2. Audit summary over-simplifies à loses information

2.2.1 Technical failure of poll can occur when still sub-quora "islands" of agreement, sufficient for policy

Hypothesis 3**: Disagreement is real, matters**

Substantive collection differences

3.1 *Some objects are corrupt* (e.g. from corruption in storage, or during transmission/harvesting)

3.2 Substantive *objects persistently missing from some replicas*
(e.g. because of permissions issue @ provider; technical failures during harvest; plugin problems)

3.3 *Versions of objects permanently missing*
(Note that later "agreement" may signify that **a later version** was verified)

Phases: 1) setup; 2) self-audit; 3) test compliance ; 4) audit PLNs

Council of Prairie and Pacific University Libraries

- 9 Institutions
- Dozens of collections – 10000s documents, images, ETDS
- *Goal* - 'Multiple' verified replicas

Digital Federal Depository Library Program "USDocs"

- Dozens of institutions
- 580+ collections – millions of documents
- *Goal*: "Many" replicas, "many" regions

Data-PASS

- 6 institution
- 5 collections – 100,000s of databases and documentation files
- *Goal* - 3 replicas, in at least 3 regions

Lessons learned

Trust, but continuously verify
--
Things go wrong without obvious symptoms

Distributed Digital Preservation Works
--
With appropriate monitoring and tuning

Replication agreement is not collection agreement
--
Use additional information to verify collections and policy

Transparency Aids Preservation
--
Visible information on operation and collections is essential

Distributed Preservation Requires Distributed Auditing
--
Standard analytics are not enough

Findings

- **Most collections were successfully replicated, however …**
- **Networks unable to demonstrate policy compliance during audit**

Observations

- In a rapidly changing environment, every upgrade may involve a research problem
- To collaborate with researchers sometimes requires the capacity to do one's own research
- Information policies, practices, and standards are changing at multiple levels, information science research program can influence development outside disciplinary boundaries

Bibliography (Selected)

- University Leadership Council, 2011, Redefining the Academic Library: Managing the Migration to Digital Information Services
- W. Lougee, 2002. Diffuse Libraries: Emergent Roles for the Research Library in the Digital Age
- C. Hess & E. Ostrom 2007, Understanding Knowledge as a Commons
- King, Gary. 2007. An Introduction to the Dataverse Network as an Infrastructure for Data Sharing. Sociological Methods and Research 36: 173–199NSB
- International Council For Science (ICSU) 2004. ICSU Report of the CSPR Assessment Panel on Scientific Data and Information. Report.
- B. Schneier, 2012. *Liars and Outliers*, John Wiley & Sons
- David S.H. Rosenthal, Thomas S. Robertson, Tom Lipkis, Vicky Reich, Seth Morabito. "Requirements for Digital Preservation Systems: A Bottom-Up Approach", D-Lib Magazine, vol. 11, no. 11, November 2005.
- National Science Board (NSB), 2005, Long-Lived Digital Data Collections: Enabling Research and Education in the 21rst Century, NSF. (NSB-05-40).

Questions?

E-mail: Micah_altman@alumni.brown.edu

Web: micahaltman.com

Twitter: @drmaltman