Author(s): Paul Conway, Ph.D., 2010

Citation Key

for more information see: http://open.umich.edu/wiki/CitationPolicy

Use + Share + Adapt

{ Content the copyright holder, author, or law permits you to use, share and adapt. }

PD-GOV **Public Domain – Government**: Works that are produced by the U.S. Government. (17 USC § 105)

PD-EXP **Public Domain – Expired**: Works that are no longer protected due to an expired copyright term.

PD-SELF **Public Domain – Self Dedicated**: Works that a copyright holder has dedicated to the public domain.

(cc) ZERO **Creative Commons – Zero Waiver**

(cc) BY **Creative Commons – Attribution License**

(cc) BY-SA **Creative Commons – Attribution Share Alike License**

(cc) BY-NC **Creative Commons – Attribution Noncommercial License**

(cc) BY-NC-SA **Creative Commons – Attribution Noncommercial Share Alike License**

GNU-FDL **GNU – Free Documentation License**

Make Your Own Assessment

{ Content Open.Michigan believes can be used, shared, and adapted because it is ineligible for copyright. }

PD-INEL **Public Domain – Ineligible**: Works that are ineligible for copyright protection in the U.S. (17 USC § 102(b)) *laws in your jurisdiction may differ

{ Content Open.Michigan has used under a Fair Use determination. }

© FAIR USE **Fair Use**: Use of works that is determined to be Fair consistent with the U.S. Copyright Act. (17 USC § 107) *laws in your jurisdiction may differ

Our determination **DOES NOT** mean that all uses of this 3rd-party content are Fair Uses and we **DO NOT** guarantee that your use of the content is Fair.

To use this content you should **do your own independent analysis** to determine whether or not your use will be Fair.

SI 640 DIGITAL LIBRARIES AND ARCHIVES

2010 Week 6: Large-Scale Repositories

THEMES FOR THIS WEEK

- Institutional Repositories
  - Past, present, future
- Asset Management
- Preservation Digital Libraries
- HathiTrust Case Study
- Discussion

• Crowe, The Case for Institutional Repositories (2002).

ORIGINS OF INSTITUTIONAL REPOSITORIES

1. Institutional R

2. Asset Mgmt

3. Preservation DL

4. HathiTrust

5. Discussion

- Serials pricing crisis in libraries
- Open access journals as a solution
  - Politics, process, intellectual property
  - Creative commons, economic models, etc.
- Advances in storage and management technologies
  - Open source software (tools and a model)
  - New roles for libraries in tech management

PRICING TRENDS IN ARL LIBRARIES

- ARL Statistics, 2003-04

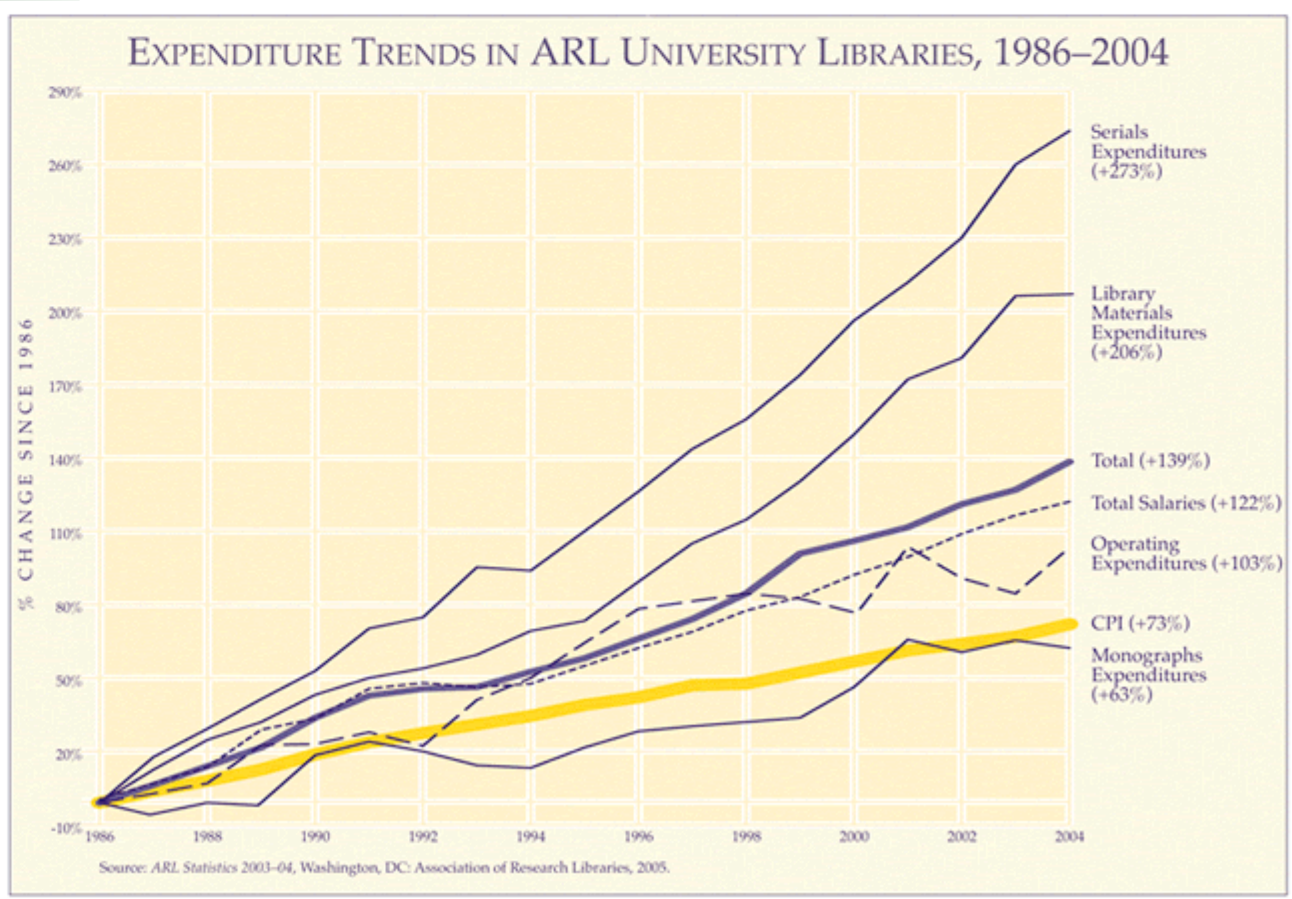
1. Institutional R

2. Asset Mgmt

3. Preservation DL

4. HathiTrust

5. Discussion

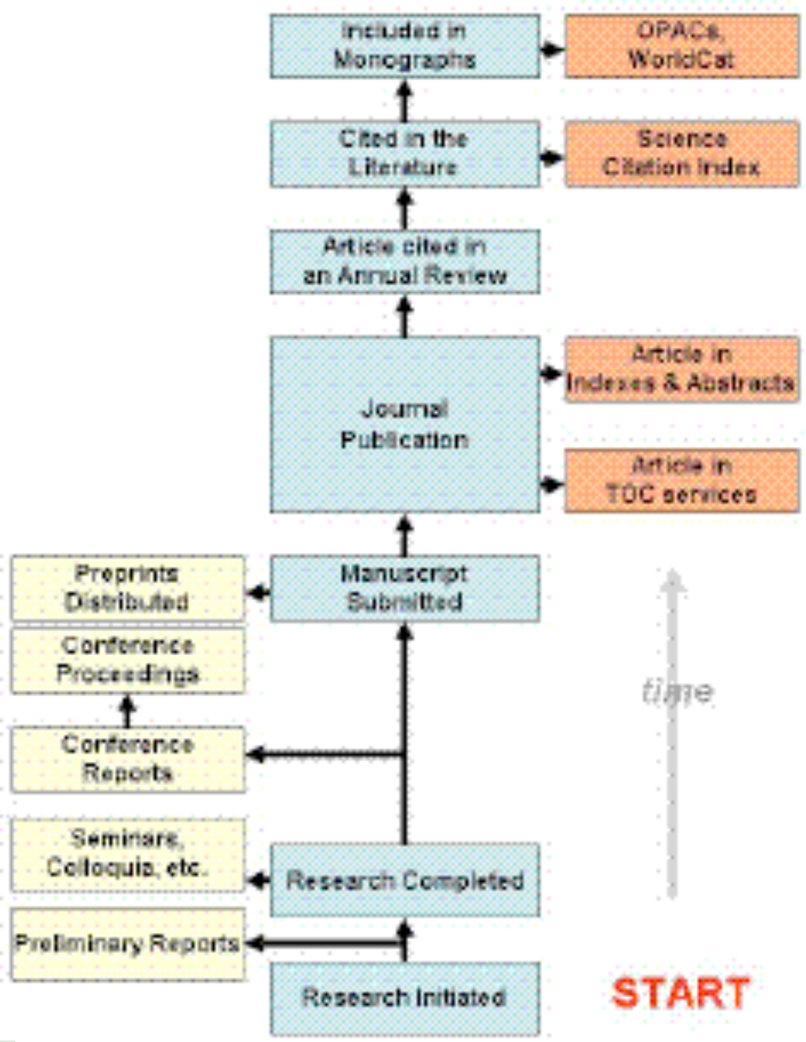


Expenditure Trends in ARL University Libraries, 1986–2004

Source: *ARL Statistics 2003–04*, Washington, DC: Association of Research Libraries, 2005.

SCHOLARLY COMMUNICATION: LINEAR MODEL

1. **Institutional R**
2. Asset Mgmt
3. Preservation DL
4. HathiTrust
5. Discussion



PD-INEL Source Undetermined

CHALLENGES FOR DEVELOPMENT OF IR

- Markey, et al. MIRACLE, 2007.

- Markey, et al. MIRACLE, 2007.

1. Institutional R

2. Asset Mgmt

3. Preservation DL

4. HathiTrust

5. Discussion

- Institutional focus
  - Discussion: what is the challenge here
- Repository focus
  - Discussion: what is the challenge here

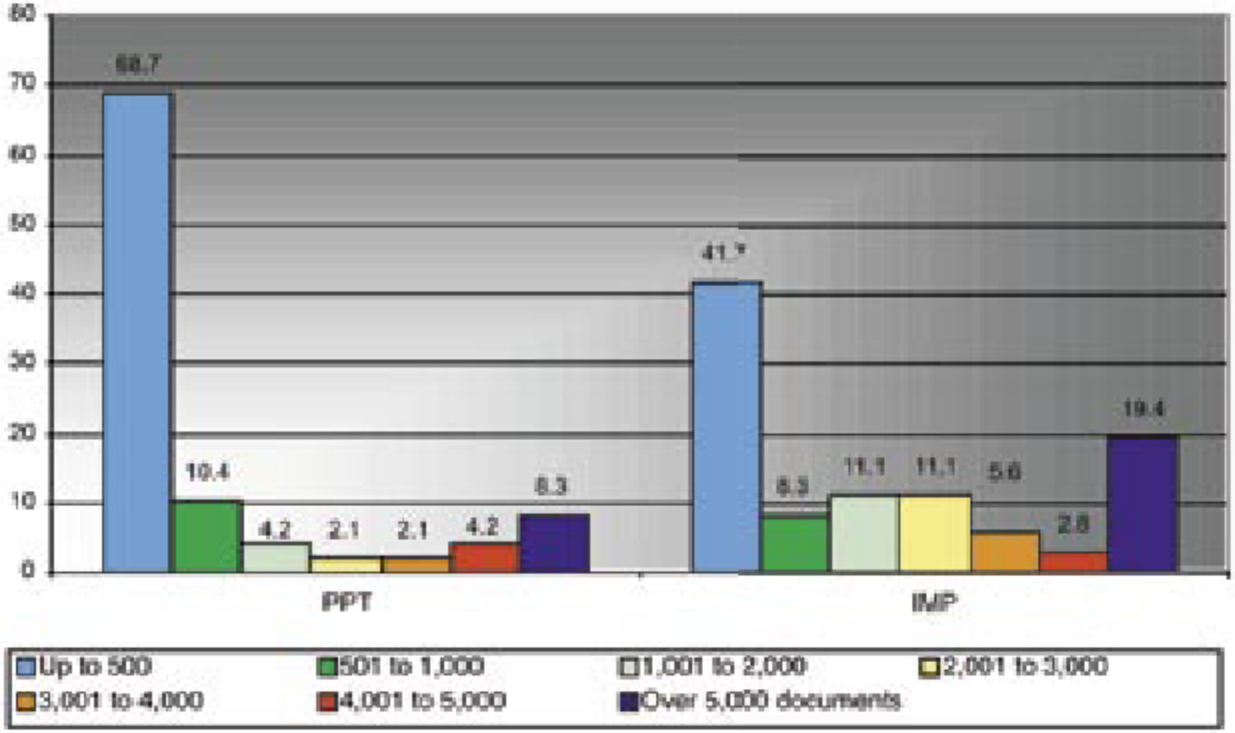
- Incentives to contribute
- Content types and level of support
- Policy frameworks for maintenance
- Centralization versus decentralization
- Technology limitations (until recently)

• **Markey, et al. MIRACLE, 2007.**

MIRACLE – CENSUS OF IR

1. **Institutional R**
2. Asset Mgmt
3. Preservation DL
4. HathiTrust
5. Discussion

Figure 6.1. Number of digital documents in IRs



PD-INEL [Council on Library and Information Resources]

IR TECHNOLOGIES

1. **Institutional R**
2. Asset Mgmt
3. Preservation DL
4. HathiTrust
5. Discussion

- DSpace
  - http://www.dspace.org/
- Fedora
  - http://www.fedora-commons.org/
- ePrints
  - http://www.eprints.org/
- Digital Commons (BE Press)
  - http://www.bepress.com/ir/
- eScholarship (CDL)
  - http://www.cdlib.org/services/publishing/escholarship.html

- OpenDOAR – Directory of Open Access Repositories http://www.opendoar.org/

MANAGING DIGITAL ASSETS IN HIGHER ED.

- Waters, Managing Digital Assets, ARL 244 (2006).

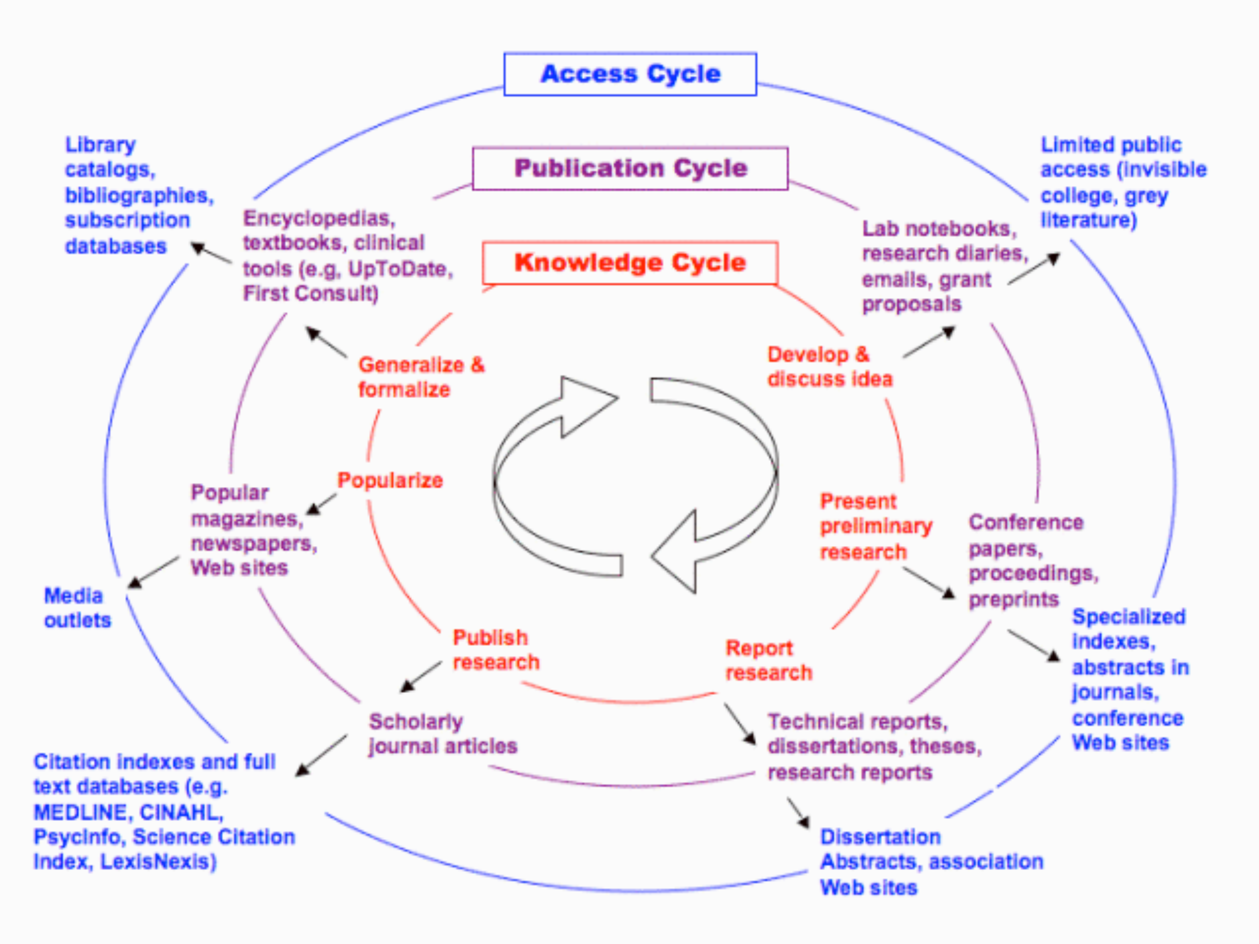
1. Institutional R
2. **Asset Mgmt**
3. Preservation DL
4. HathiTrust
5. Discussion

- The ‘processed’ publication
- Intellectual property
- Search
- Research methods
- New publication emphases
- New collection emphases
- Interactions among systems

THE SCHOLARLY KNOWLEDGE CYCLE

- Lyon, “eBank UK.” http://www.ariadne.ac.uk/issue36/lyon/

1. Institutional R
2. **Asset Mgmt**
3. Preservation DL
4. HathiTrust
5. Discussion



Access Cycle
Publication Cycle
Knowledge Cycle
Library catalogs, bibliographies, subscription databases
Encyclopedias, textbooks, clinical tools (e.g. UpToDate, First Consult)
Lab notebooks, research diaries, emails, grant proposals
Limited public access (invisible college, grey literature)
Generalize & formalize
Develop & discuss idea
Popularize
Popular magazines, newspapers, Web sites
Present preliminary research
Conference papers, proceedings, preprints
Media outlets
Publish research
Report research
Specialized indexes, abstracts in journals, conference Web sites
Scholarly journal articles
Technical reports, dissertations, theses, research reports
Citation indexes and full text databases (e.g. MEDLINE, CINAHL, PsycInfo, Science Citation Index, LexisNexis)
Dissertation Abstracts, association Web sites

Evidence-Based Nursing

- **Van de Somple, et al. Rethinking Scholarly Communication (2004) D-Lib Magazine**

VALUE CHAIN OF SC SYSTEM

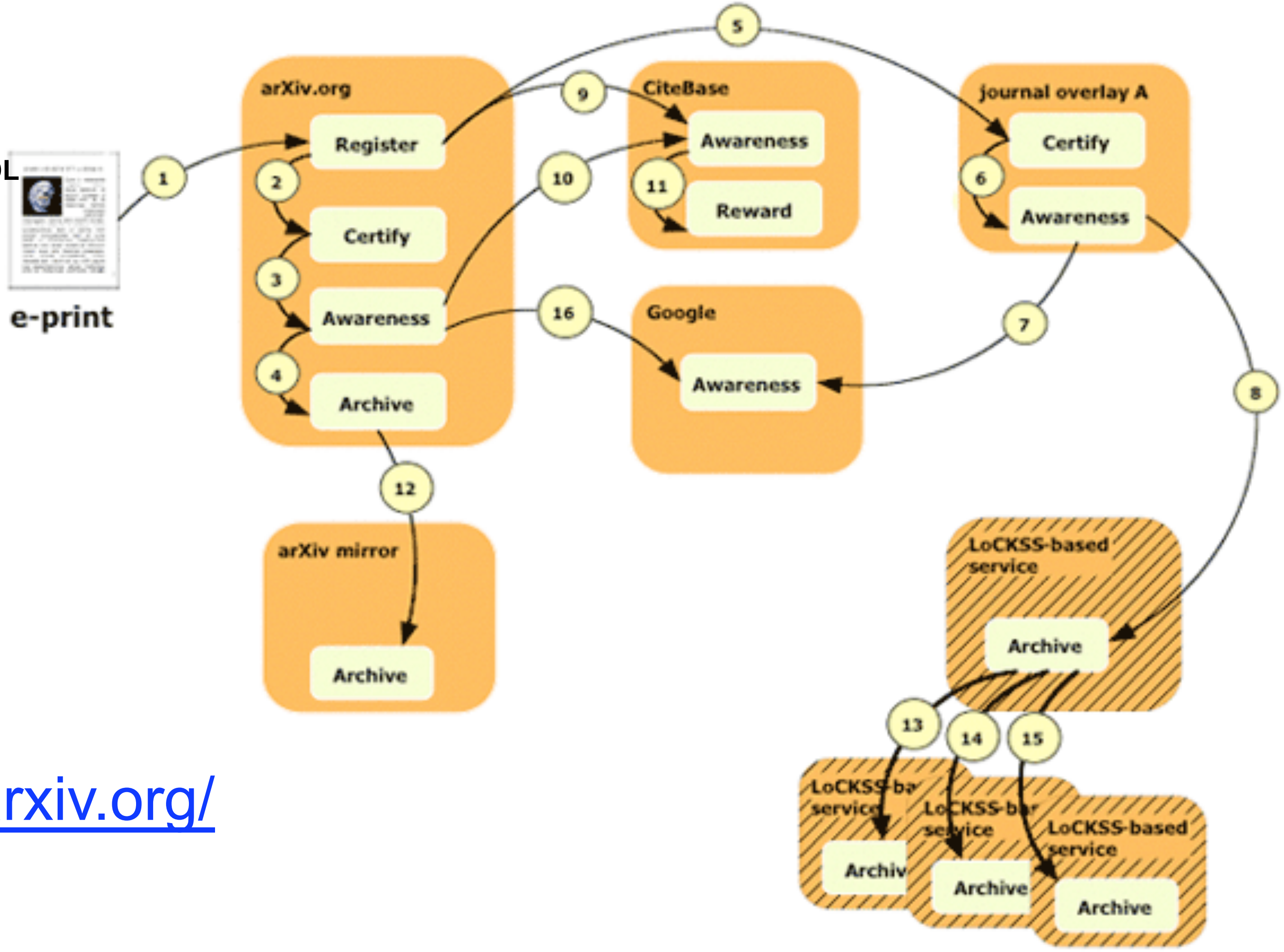
1. Institutional R
2. Asset Mgmt
3. **Preservation DL**
4. HathiTrust
5. Discussion

- *Registration*, which allows claims of precedence for a scholarly finding.
- *Certification*, which establishes the validity of a registered scholarly claim.
- *Awareness*, which allows actors in the scholarly system to remain aware of new claims and findings.
- *Archiving*, which preserves the scholarly record over time.
- *Rewarding*, which rewards actors for their performance in the communication system based on metrics derived from that system.

ARXIV ECOLOGY AND SERVICE PATHWAYS

- Van de Sompel, D-Lib Magazine, September 2004.

1. Institutional R
2. Asset Mgmt
3. **Preservation DL**
4. HathiTrust
5. Discussion



http://arxiv.org/

PD-INEL [D-Lib Magazine September 2004 Volume 10 Number 9 ISSN 1082-9873]()

• **Wheatley, “IR and Preservation” (2004).**

INST. REPOSITORIES & PRESERVATION

1. Institutional R
2. Asset Mgmt
3. **Preservation DL**
4. HathiTrust
5. Discussion

- Digital Preservation Coalition – Technology Watch Report
- Preservation Functions
  - Unique, persistent identification (naming)
  - Ingest capabilities (acquisition)
  - Representation system (access knowledge)
  - Technology Watch (vigilance)
  - Rendering (data to display capability)
  - Repository structure (storage, db, apps)
  - Recording change metadata

- Rusbridge, “Excuse Me…” (2006)

EXCUSE ME… FALLACIES

1. Institutional R
2. Asset Mgmt
3. **Preservation DL**
4. HathiTrust
5. Discussion

- Digital preservation is very expensive [because]
- File formats become obsolete very rapidly [which means that]
- Interventions must occur frequently, ensuring that continuing costs remain high.
- Digital preservation repositories should have very long timescale aspirations,
- 'Internet-age' expectations are such that the preserved object must be easily and instantly accessible in the format de jour, and
- the preserved object must be faithful to the original in all respects.

- **RLG/NARA Audit Checklist, 2007**

AUDITING AND CERTIFYING ARCHIVES

1. Institutional R
2. Asset Mgmt
3. **Preservation DL**
4. HathiTrust
5. Discussion

- Organization
- Repository functions, processes, and procedures
- Designated community and useability of information
- Technologies and technical infrastructure

TRAC http://www.crl.edu/content.asp?l1=13&l2=58&l3=162

- Rieger, Large-Scale Digitization (2008)

LARGE-SCALE DIGITIZATION

Assessment Criteria

- Selection for Digitization and Preservation Reformatting
- Content Creation
  - Image-Quality Procedures for Large-Scale Digitization Initiatives
  - Preservation Metadata
  - Descriptive and Structural Metadata
  - Quality Control
- Technical Infrastructure
- Organizational Infrastructure

6 of 13 Recommendations

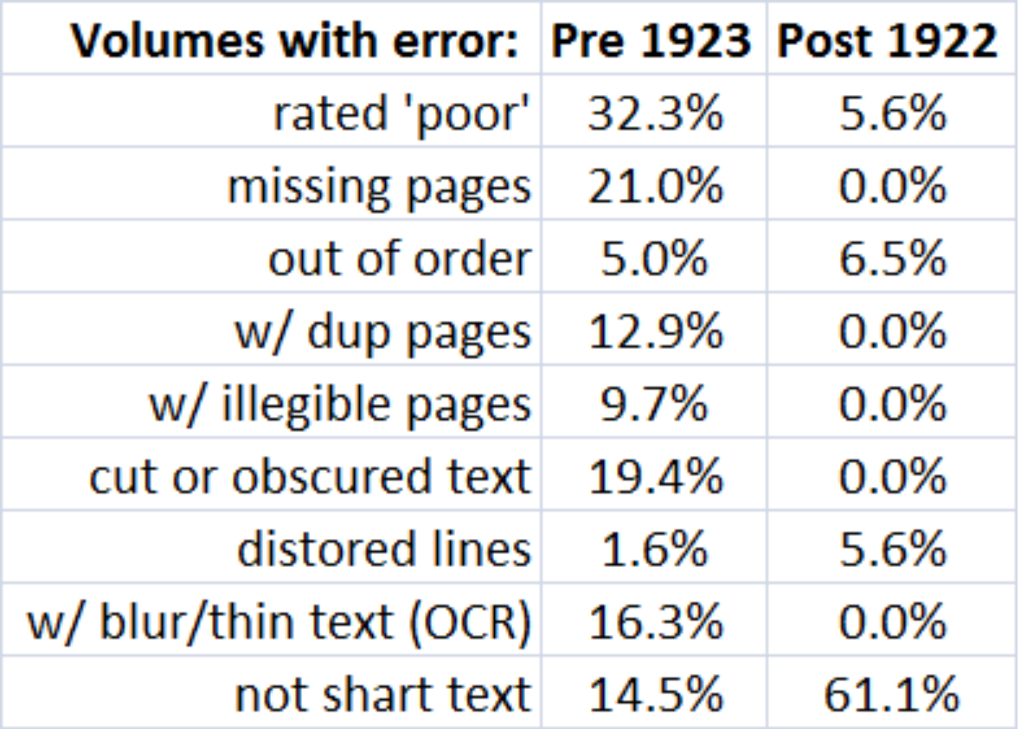
- Reassess Digitization Requirements for Archival Images
- Develop a Feasible Quality Control Program
- Balance Preservation and Access Requirements
- Enhance Access to Digitized Content
- Understand the Impact of Contractual Restriction on Preservation Responsibilities
- Lend Support for Shared Print-Storage Initiatives

- Henry and Smith, “Ghostlier Demarcations” (2010)
- Gevinson, “Quality Assessment Summary” (2010)

QUALITY ISSUES

1. Institutional R
2. Asset Mgmt
3. **Preservation DL**
4. HathiTrust
5. Discussion

- Ghostlier Demarcations (quality of scans)



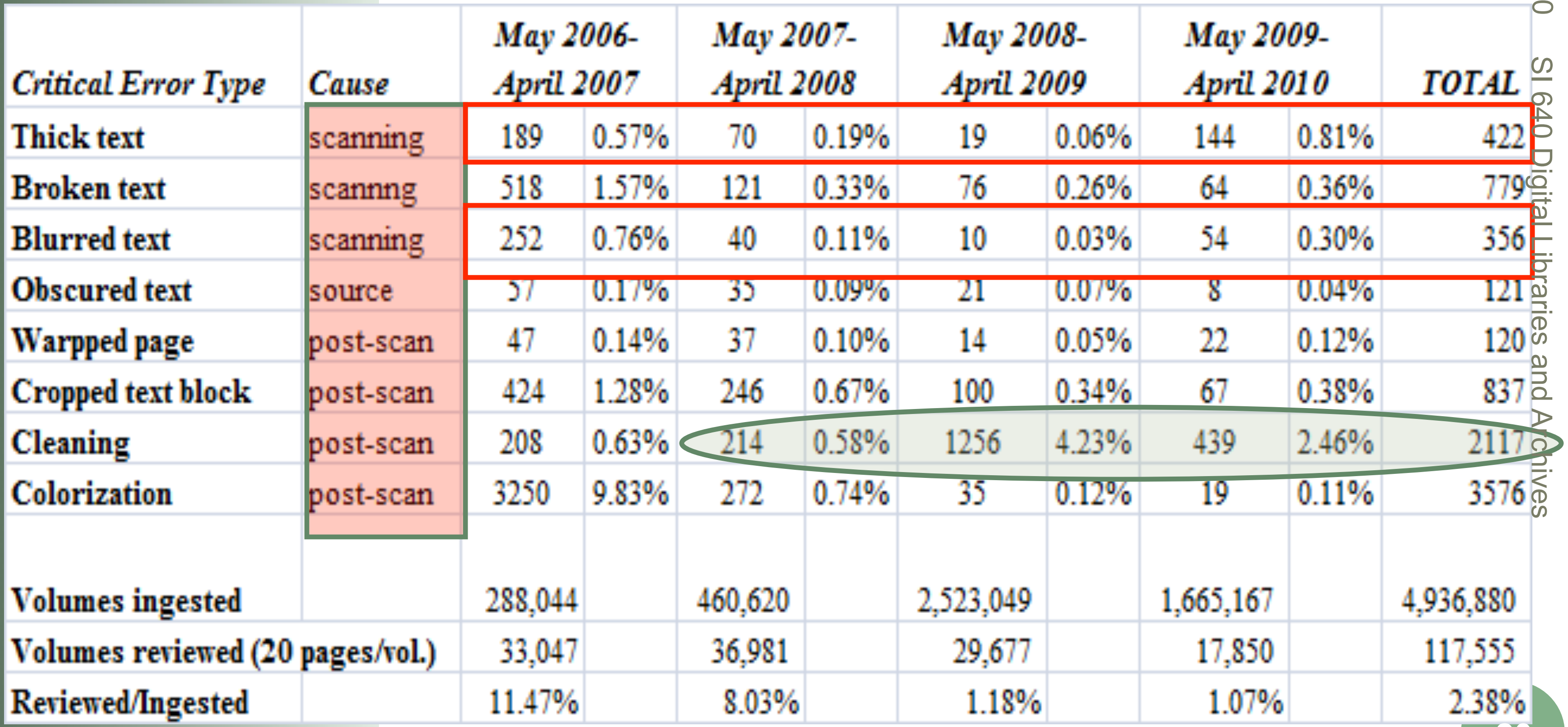
Volumes with error:	Pre 1923	Post 1922
rated 'poor'	32.3%	5.6%
missing pages	21.0%	0.0%
out of order	5.0%	6.5%
w/ dup pages	12.9%	0.0%
w/ illegible pages	9.7%	0.0%
cut or obscured text	19.4%	0.0%
distored lines	1.6%	5.6%
w/ blur/thin text (OCR)	16.3%	0.0%
not shart text	14.5%	61.1%

PD-INEL Source Undetermined

INCIDENCE OF CRITICAL ERROR IN HATHITRUST

- Conway, “Validating Quality in HathiTrust” (iPres2010)

University of Michigan Quality Review, 2006-10



Critical Error Type	*Cause*	*May 2006- April 2007*		*May 2007- April 2008*		*May 2008- April 2009*		*May 2009- April 2010*		*TOTAL*
Thick text	scanning	189	0.57%	70	0.19%	19	0.06%	144	0.81%	422
Broken text	scannng	518	1.57%	121	0.33%	76	0.26%	64	0.36%	779
Blurred text	scanning	252	0.76%	40	0.11%	10	0.03%	54	0.30%	356
Obscured text	source	57	0.17%	35	0.09%	21	0.07%	8	0.04%	121
Warpped page	post-scan	47	0.14%	37	0.10%	14	0.05%	22	0.12%	120
Cropped text block	post-scan	424	1.28%	246	0.67%	100	0.34%	67	0.38%	837
Cleaning	post-scan	208	0.63%	214	0.58%	1256	4.23%	439	2.46%	2117
Colorization	post-scan	3250	9.83%	272	0.74%	35	0.12%	19	0.11%	3576
Volumes ingested		288,044		460,620		2,523,049		1,665,167		4,936,880
Volumes reviewed (20 pages/vol.)		33,047		36,981		29,677		17,850		117,555
Reviewed/Ingested		11.47%		8.03%		1.18%		1.07%		2.38%

PD-INEL Source Undetermined

• York, “Building a Future” (2010)

HATHITRUST CASE STUDY

1. Institutional R
2. Asset Mgmt
3. Preservation DL
4. **HathiTrust**
5. Discussion

- University of Michigan, School of Information (MSI 2008)
  - Web Development Librarian
  - Interface Programmer
  - Special Projects Librarian

  - HathiTrust ---

Source Undetermined

Google Images (NISO/BISG 4th Annual Forum 2010)

Thank you!

Paul Conway
Associate Professor
School of Information
University of Michigan
www.si.umich.edu

Additional Source Information

for more information see: http://open.umich.edu/wiki/CitationPolicy

Slide 6: University of Washington, http://www.lib.washington.edu/images2/graphics/economics_graph.png/view

Slide 7: Source Undetermined

Slide 9: Council on Library and Information Resources, http://www.clir.org/pubs/reports/pub140/body.htm

Slide 12: Evidence-Based Nursing, http://ebp.lib.uic.edu/nursing/node/7

Slide 14: D-Lib Magazine September 2004 Volume 10 Number 9 ISSN 1082-9873,

http://www.dlib.org/dlib/september04/vandesompel/09vandesompel.html

Slide 19: Source Undetermined

Slide 20: Source Undetermined

Slide 21: Source Undetermined image removed