



Open Access Forever,
Or Five Years,
Whichever Comes First

Progress on Preserving the
Digital Scholarly Record

MacKenzie Smith, MIT Libraries



Open Access to What?

Not just published, refereed journal articles...

- Grey literature
theses, reports, white papers, government documents
- Supplementary material
datasets, images, simulations/visualizations
- Informal communications
email, blogs, podcasts, websites, wikis, presentations



Open Access for How Long?

“...the most counterproductive of the confluences among the five distinct aims [for OA] has been that between research self-archiving and *digital preservation*.... the current preservation burden is on its primary corpus, which is the published literature (online and on paper)... [OA] archives are a *supplement* to this primary corpus, for the purpose of maximizing its impact by maximizing access to it; not a *substitute* for it.”

[Stevan Harnad, June 2003 email]

“...OA just means free online access (immediate, *permanent*)

[Stevan Harnad, December 2006 email]



Open Access Means...

Citable

Should scholars cite what they cannot themselves access?

What if readers can't access the “official” version?

Trustworthy

To be valued, archives must be *trusted* by scholars; research publications must not randomly *disappear*



Open Access “D2D”

Library mission to provide

“**Discovery to Delivery**” of the scholarly record

- Scholars can’t “discover” what no one collected, cataloged, or published online (ACCESS)
- Libraries can’t “deliver” what they haven’t collected or maintained over time (PRESERVATION)



Publishing Trends

15 year half-life for STM articles, longer in other fields

Publishers going “electronic only”

- 2006 search of *Ulrich’s International Periodicals Directory* for all active, online, and refereed journal titles ... returned 14,338 hits (1,429 of which are open-access titles), representing **62% of all active and refereed titles listed**. (The remaining 38% were print only).
- 2003 Electronic Publishing Services Ltd. study projecting publishing trends to 2020 claimed that by 2016, **half of all serial publications will have migrated to electronic-only format**. They predicted that science, technology, and medical (STM) titles would be the first to switch.

[CLIR report – E-journal Archiving Metes and Bounds, Sep 2006]



Publishing Trends

Libraries going “electronic only”

- EBSCO estimates that for STM titles, online journal subscriptions will exceed print subscriptions by 2008. [Bruce 2005]
- 84% of respondents to a 2004 Publishers Communication Group survey of academic libraries worldwide said they cancel print when an electronic version is available
- Research libraries in an ARL member survey reported canceling print equivalents for bundled e-content in 153 out of 266 contracts (58%) for 2006 [Hahn 2006]
- Print repositories are being developed at the regional and national levels to ensure that at least one paper copy remains accessible, but increasingly institutions recognize that print is not an acceptable archival format for electronic content. [CLIR report – E-journal Archiving Metes and Bounds, Sep 2006]



Publishing Trends

Governments going “electronic only”

- U.S. e-government act of 2002
- 65 Percent of U.S. federal government publications produced *only online* in 2003
- Estimated 90 percent of U.S. federal government publications available *only online* by 2008

[California Digital Library <http://www.cdlib.org/inside/projects/preservation>]



Digital Preservation

“Digital Information Lasts Forever –
Or Five years, Whichever Comes First”

Jeff Rothenberg, 1997, RAND Video V-029

Combating link rot, bit rot, human error,
format obsolescence, lack of mission,
lack of funding...



Digital Preservation

“Who should do [digital preservation] then? Non-profit organizations? Do we need a new organization formed specifically to perform the function of e-journal archiving? Should there be one for each continent? Each nation? Each language? What are the implications for access, and what kind of economic model would work? Is it possible to have publishers to establish something like an escrow deposit for archival data that would be available to subscribers, regardless? What technical issues should be considered, and how do we prepare for unknown technological changes in the future which will affect storage techniques and access?”

[McKay, Blackwell Publishing, 1996]



Digital Preservation

Refers to the
management of digital information over time

Defined as the set of processes and activities
that ensure continued access to information
... in digital formats



Digital Preservation

“**Digital preservation**” means long-term, error-free storage of digital information, with means for retrieval and interpretation, for the entire time span that the information is required.

"**Retrieval**" means obtaining needed digital files from the long-term, error-free digital storage, without possibility of corrupting the continued error-free storage of the digital files.

"**Interpretation**" means that the retrieved digital files, files that, for example, are of texts, charts, images or sounds, are decoded and transformed into intelligible representations.

[Wikipedia, 2006]



Trustworthy Repositories Audit and Certification Checklist

- Organization Infrastructure

Governance, organizational structure, mandate/mission, scope, roles & responsibilities, policy, funding, contracts, licenses, transparency

- Digital Object Management

Acquisition, ingest, metadata, storage, preservation strategies, access

- Technologies, Technical Infrastructure, and Security

Repository system management, disaster recovery, replication/synchronization, integrity auditing, security,



Preservation Threat Model

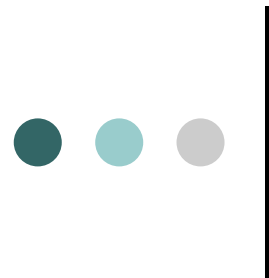
- Media faults
- Component faults (hardware, software, network, etc.)
- Media/hardware obsolescence
- Software/format obsolescence
- Natural disaster
- Human error
- External or internal attacks
- Economic failure
- Organizational failure

D.Rosenthal, LOCKSS chief scientist, et al.



General principles

- Replication – multiple copies
- Migration – of hardware, storage media, collections
- Transparency – open source software, [meta]data
- Diversity – different organizations, preservation strategies
- Auditing – checksums, file reads
- Economy – cheaper and easier is more likely to work

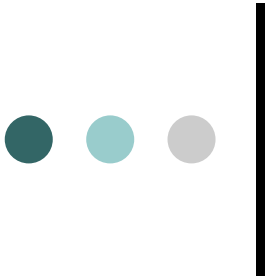


Component faults, natural disasters

Failure of server, network, software, 3rd party services, or entire data center

Possible Solution

- Separate digital asset storage from management system
- Replicate collections at multiple, geographically distributed sites
 - Technically possible now, e.g. via SRB
 - Federations work by mutual mirroring



Link Rot

- Percent of Web-based references in scientific articles from 3 major journals inaccessible within *2 years* of publication: **21%**
- Proportion of 1998 websites gone in 1999: **44%**
- Life of an average website: **44 days**

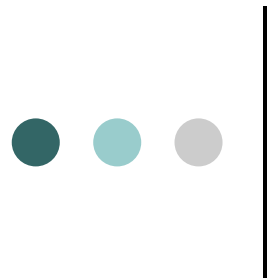


Media faults, obsolescence

Extensive (e.g. entire SAN) or limited (e.g. CD-ROM bit rot); visible or latent at the point of request

Possible Solution

- Audit data frequently (e.g. checksum checking)
- Replicate collections at multiple sites (e.g. via SRB)
- Automate repairs from replications

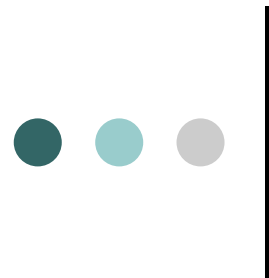


Software/format obsolescence

Software to read digital files no longer runs
(e.g. VisiCalc, AppleWorks, Wordstar, Ami Pro)

Possible Solution

- User education on best formats to archive via published support policies
- Format migration (on ingest, on-the-fly)
- Integration with PRONOM or the GDFR
- Format-specific preservation plans
(e.g. Microsoft Excel, Word, Powerpoint; HTML, XHTML)



Human Error, External or internal attacks

Digital collections are deliberately or accidentally damaged by people or their (software) agents

Possible Solution

- Event tracking system (“History”) to record significant Information Lifecycle events
- Regular data audits



Legal Constraints

Law prevents preservation by non-copyright-holder (i.e. creation of “derivative works”) without permission, or modifying proprietary format

Possible Solution

- Clear rights at the time of submission
- Keep earlier version (e.g. preprints)
- Keep non-proprietary formats



Economic, Organizational failure

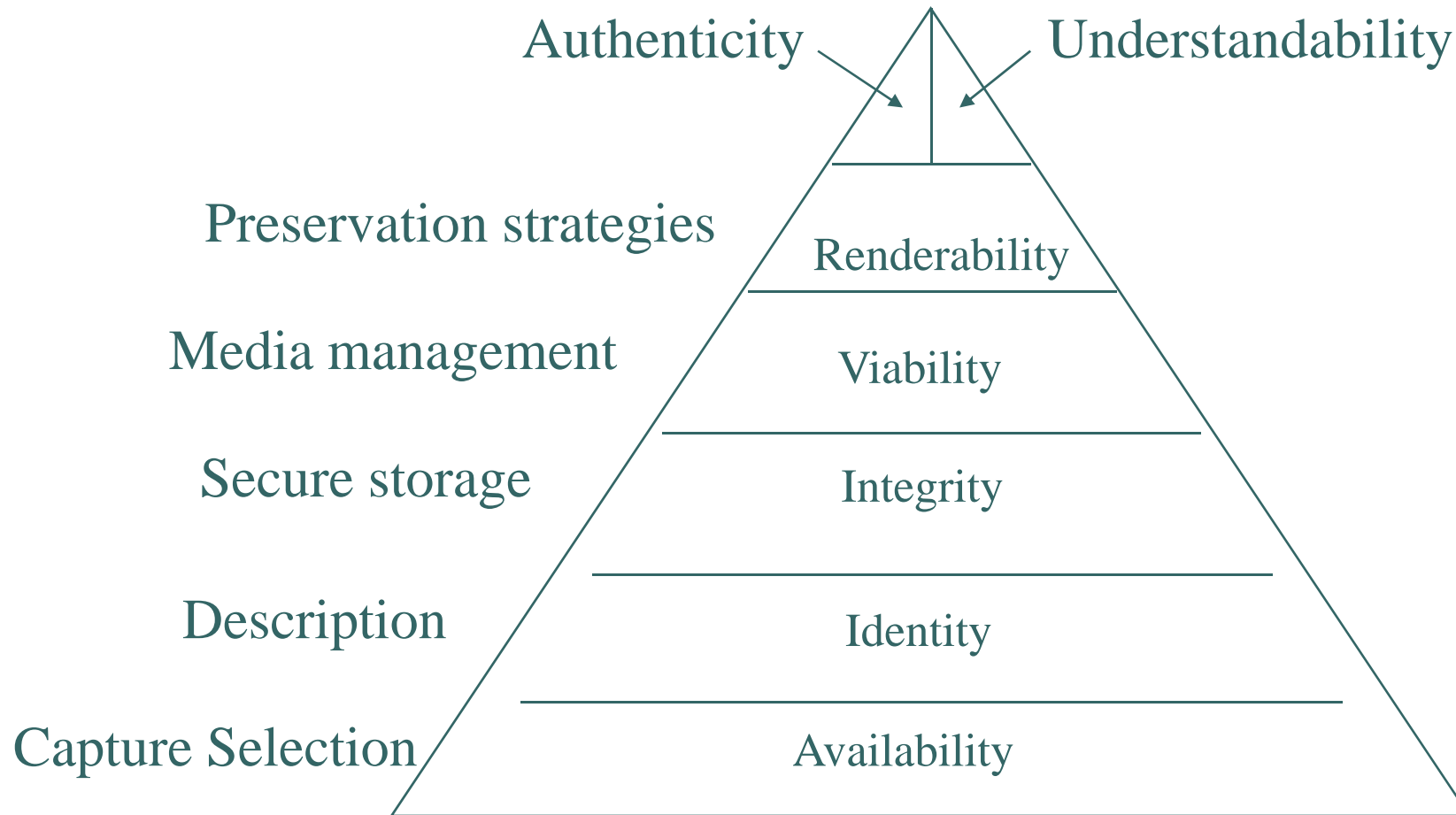
Archival organization can no longer care for the collections, due to high maintenance costs, bankruptcy, incompetence, acquisition or reorganization, etc.

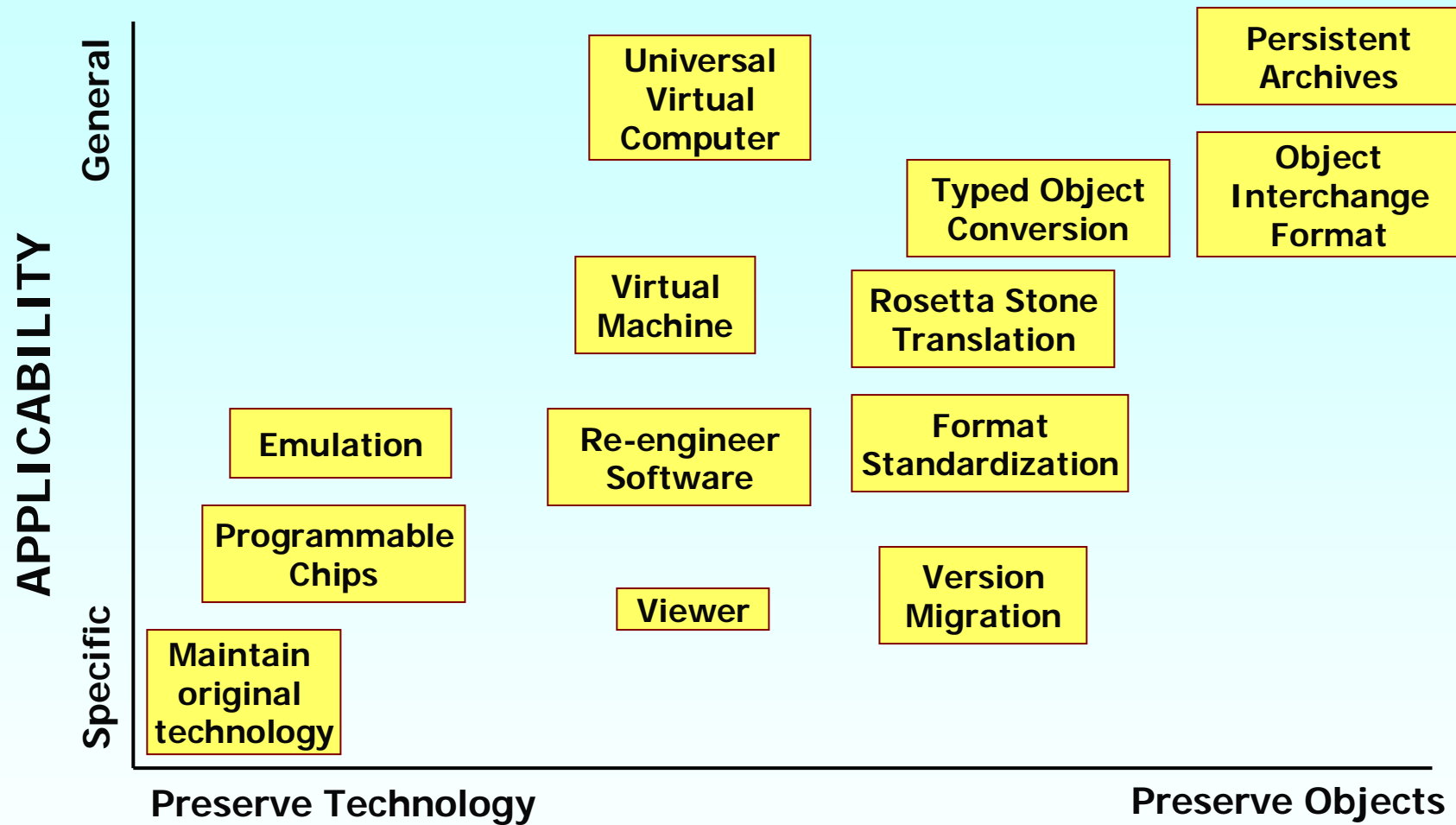
Possible Solution

- Keeps costs low
 - share software maintenance, leverage international infrastructure, minimize human curation
- Safety in numbers
 - large community of open source adopters with vested interest
- Collections or entire archive transferable to another organization
 - standard export formats (METS, IMS-CP, MPEG21, OAI-ORE)
 - standard export mechanisms (OAI-PMH, WebDAV, JSR170, Atom/RSS)



The Preservation Pyramid





OBJECTIVE

Source: Thibodeau, 2002.





Progress on Preservation

Emerging international infrastructure, e.g.

- EU PLANETS, CASPAR projects
- UK Digital Curation Centre
- Australian PANDAS
- US Library of Congress NDIIPP, NARA ERA
- US NSF Cyberinfrastructure, UK e-Science
- Internet Archive
- WebCite



Progress on Preservation

Emerging tools and services, e.g.

- Format identification, validation, metadata extraction
 - US GDFR/JHOVE, UK PRONOM/DROID
 - National Library of New Zealand toolkit
 - NIST software database
 - CRiB
- Format normalization on ingest
 - DAITSS (FCLA, USA)
 - Xena (Australian National Archives)
 - DiVA (Uppsala University, Sweden)



Progress

- Content Replication
 - LOCKSS
 - San Diego Supercomputer Center's SRB/iRODS
- Migration on-the-fly
 - Multivalent browser
 - LOCKSS
- Emulation
 - IBM Digital Library
 - UVC



Progress

Emerging commercial services, e.g.

- IBM Digital Library system
- OCLC digital archive
- Portico (Ithaka Harbors, Inc.)
- Iron Mountain, Akamai
- Archivas, Permabits, Bamboo
- Chronopolis grid storage (SDSC, USA)



Progress

Open Source Repositories and Services, e.g.

- DSpace, EPrints, Fedora
all actively developing preservation support
- DAITSS, LOCKSS
- SRB/iRODS
- Xena, JHOVE, DROID



Digital Preservation

Begin by creating *local policies* within a framework
(e.g. TRAC, DRAMBORA)

- Collection policies
- Format support policies
- Access control and availability policies
- Backup/disaster recovery (SLA) policies



OA and Preservation

- Access and Preservation are closely related
- Open Access makes preservation easier, less expensive
- Preservation makes Open Access possible!