



**“Diffused Knowledge
Immortalizes Itself”**

Sir James Mackintosh 1765-1832





MOTIVATION



<http://lockss.stanford.edu>

L O T S O F C O P I E S K E E P S T U F F S A F E





Paper Library System

Libraries act for their institution to

- Acquire copies of important “stuff”
- Keep copies on shelves
- Give access to local readers

Libraries cooperate to

- Supply copies to other libraries
 - a reader can easily find a copy
 - a “bad guy” has trouble finding and destroying all copies





Paper Library System

Libraries ensure content persists simply by supporting their local communities

A cooperative, affordable, decentralized, 'archive system' with LOTS OF COPIES





LOCKSS “Library System”

Libraries act for their institution to

- Acquire copies of important “stuff”
- Keep copies in transparent web caches
- Give access to local readers

Libraries cooperate to

- Detect and repair damage
 - a reader can easily find a copy
 - a “bad guy” has trouble finding and destroying all copies





LOCKSS “Library System”

Libraries ensure content persists simply by supporting their local communities

*A cooperative, affordable,
decentralized, ‘archive system’
with LOTS OF COPIES*





Long Lived: *slow, determined, indestructible*





LOCKSS

- Open source
- Peer to peer
- Persistent access preservation system
- Web delivered information

Production: Released April 2004

Support: Mellon, NSF, Stanford Libraries

Software: www.sourceforge.net

Teams: Production and Research





Research Team

Stanford, Harvard, HP Labs & Intel Labs

Award winning research: ACM 2004

- Best Paper SOSP
- Grand Finals 2nd place all student research
- 7 other key research papers accepted so far

Investigating LOCKSS communication

- Scaling, attack resistance

Production Team:

deploys findings, builds system





LOCKSS Global Cache Monitor

Map View

[Switch to Table View](#)

Show the latest snapshot of LOCKSS data.

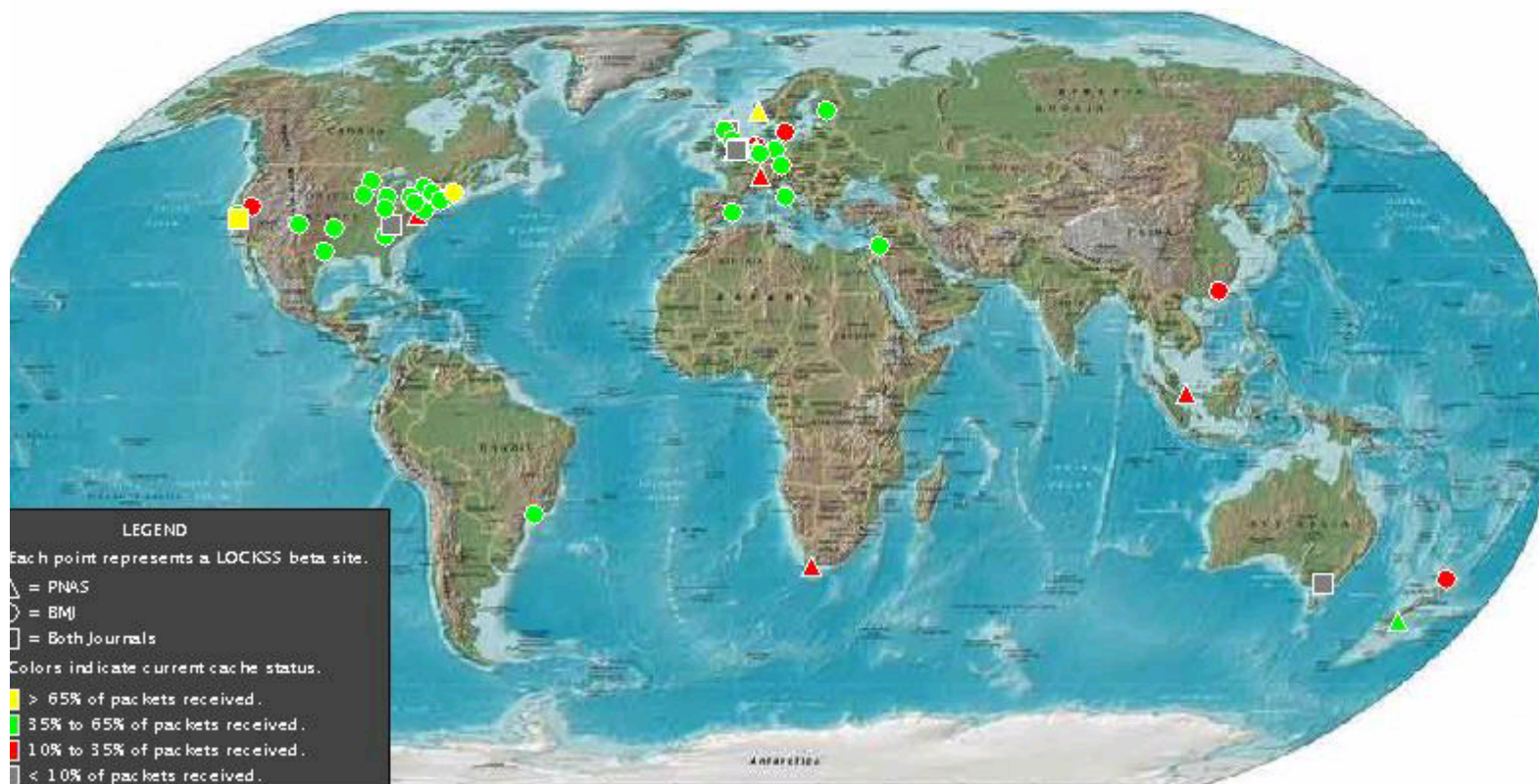
Current as of: **Sat Apr 13 00:00:00 PDT 2002.**

Zoom in by clicking on the map.

Partners

50+ Publishers

90+ Libraries





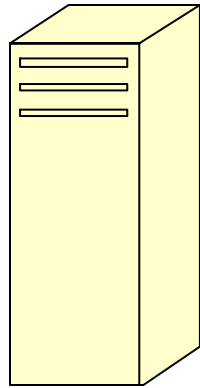
LIBRARIES and PUBLISHERS

™ "...let us save what remains: not by vaults and locks which fence them from the public eye and use in resigning them to the waste of time, but by such a multiplication of copies, as shall place them beyond the reach of accident." Jefferson, Thomas.
1] 1984. Thomas Jefferson to Ebenezer Hazard, Philadelphia, February 18, 1791. In Thomas Jefferson: Writings: Autobiography, Notes on the State of Virginia, Public and Private Papers, Addresses, etc., edited by Merrill D. Peterson. New York: Library of America

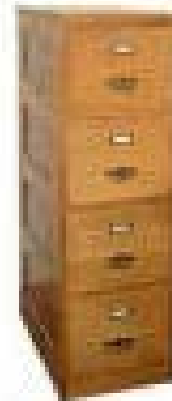
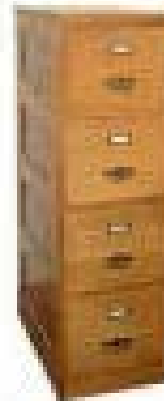
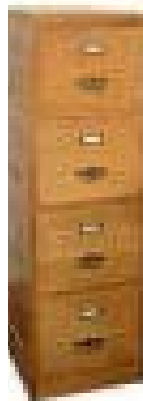
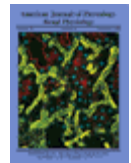
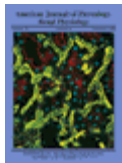
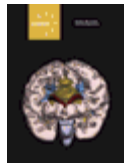
Descriptions Ref Published Papers LOCKSS Releases & News Articles	Status Libraries and Endorsing Publishers April 5, 2004 - Production software released ask us to send you the software and instructions	Technical Specifications Overview Security Network Integration The Plug-in OAIS Research - FAQ
Publishers Publisher Actions	Librarians Collection Development Humanities Project Title Registry User Interface Demo	LOCKSS Alliance Description
Related Work Other Work	Frequently Asked Questions Frequently Asked Questions	Government Documents LOCKSS-DOCS
	Software License	Credits Funders & Technical Wizards



1 PC holds
~2,500 e-j
years



LOCKSS software turns a PC
into a persistent web cache
into a preservation tool



600MHz-128MB RAM-Bootable CD drive-Floppy disk drive





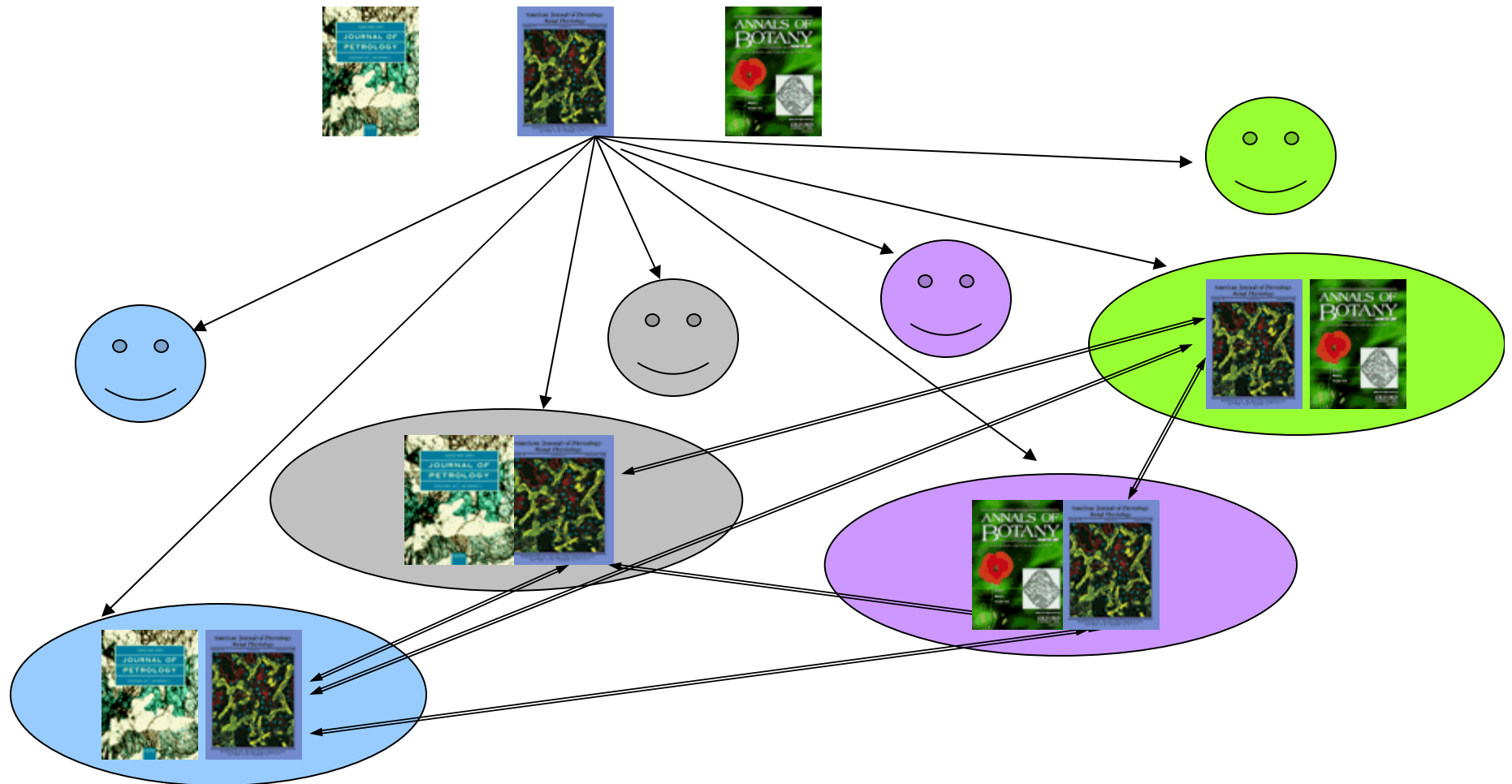
LOCKSS Caches

- Crawls and collects HTTP content
 - All formats (PDF, HTML, JPEG, TIF, Audio, Video)
- Preserves content integrity
 - Independent collection
 - Cooperate to audit and repair damage
- Provides access
 - Via web browser
 - Content is never “dark”





Approximate Data Flows

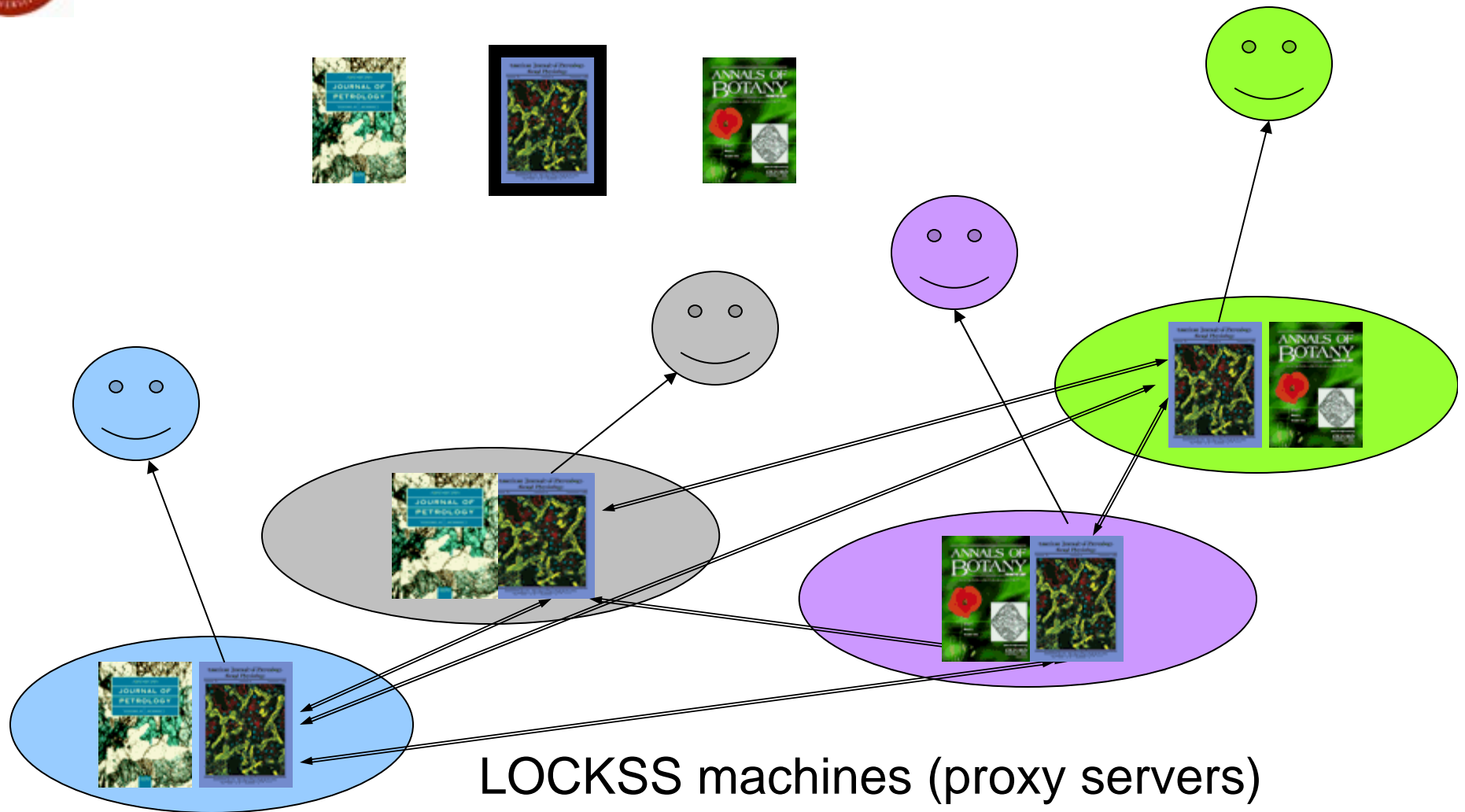


LOCKSS machines





Approximate Data Flows



Prevent the publisher from revoking access rights to back content





You're Crazy

A research library's serial
collection on a PC

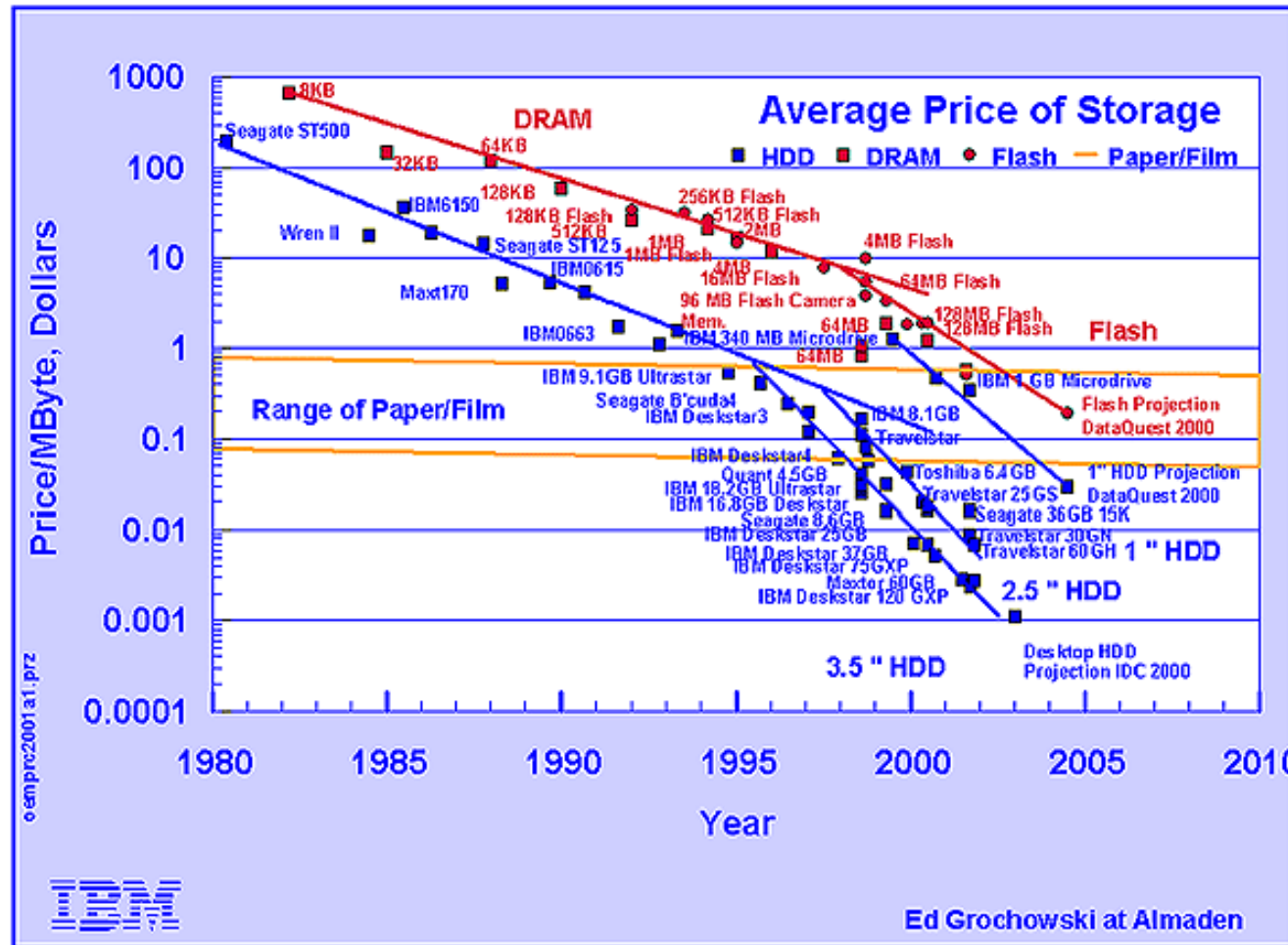
?





Hardware Costs

*HDD
prices
decline
by 50%
a year*



<http://www.almaden.ibm.com/sst/html/leadership/g05.htm>





Terabytes of E-Journals

Median e-journal size is less than 0.5 GB/ year

1 Terabyte (1000 GB) = 2000 journal years

	<u>J-yr storage</u>	<u>TB/PC</u>	<u>J-yrs/PC</u>
2004	\$0.35	1.44	2,880
2005	\$0.28	2.88	5,760
2006	\$0.14	5.76	11,520
2007	\$0.07	11.52	23,000





Netscape

File Edit View Go Communicator Help

LACIE
Made for ideas.

Products | Support | LaCie

Hard Drives Optical Tape Color Add-Ons


0 item(s)

Search

LaCie USA > Products > Hard Drives > Bigger Disk > Bigger Disk


Bigger Disk

Unprecedented 1 terabyte capacity




- User Benefits :
- Unique triple interface solution
- FireWire 800, FireWire and USB 2.0 hard drive
- Sturdy aluminum alloy design
- Stackable desktop or rackmount configuration

1 terabyte for \$1,199.00



largest capacity available
The LaCie Bigger Disk, with the largest hard drive capacity available, is a unique innovation that packs an amazing 1 terabyte of storage space in a manageable 5.25" form factor. With this unsurpassed storage capacity, the LaCie Bigger Disk allows users to store nearly two years of continuous music and up to one month of non-stop MPEG-2 video¹. Truly plug and play, this device requires no driver or software installation for Windows XP and Mac OS X users.



triple interface solution
Built with a unique triple interface solution, the LaCie Bigger Disk can be connected to any computer equipped with FireWire 800, FireWire 400, iLink/DV, Hi-Speed USB 2.0 or USB 1.1, making it the most universal drive ever. The FireWire 400 port can be used to conveniently chain a DV camera to your computer for direct digital video capture. The LaCie Bigger Disk can also be daisy chained to most storage, audio and video devices to achieve even greater storage.

Document: Done





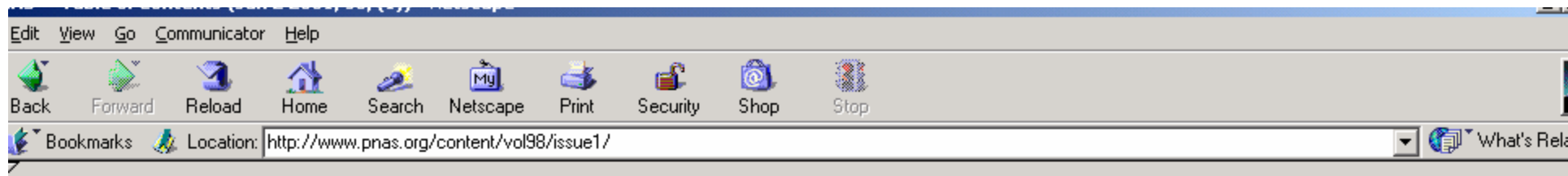
Look and Feel to Readers

Configure LOCKSS as a web proxy

Example:

- PNAS table of contents page
 - from web (9/11/02)
 - from LOCKSS cache





NAS Online

ARCHAEA

An International
Microbiological
Journal

[ME](#) [HELP](#) [FEEDBACK](#) [SUBSCRIPTIONS](#) [ARCHIVE](#) [SEARCH](#) [TABLE OF CONTENTS](#)

stitution: STANFORD UNIV MED CENTER || [Sign In as Member / Individual](#)

Want to be notified by email when new content goes on-line? [\[Sign up for eTOCs\]](#)

Other Issues: [←](#) [→](#)

Table of Contents: Jan 2 2001; 98 (1) [\[Index by Author\]](#) [\[Cover\]](#)

[COMMENTARIES](#)

[PERSPECTIVES](#)

[Physical Sciences:](#)

[Mathematics](#)
[Statistics](#)

[Social Sciences:](#)

[Anthropology](#)

[Biological Sciences:](#)

[Biochemistry](#)
[Biophysics](#)
[Cell Biology](#)
[Developmental Biology](#)
[Ecology](#)
[Evolution](#)
[Genetics](#)

[Immunology](#)

[Medical Sciences](#)
[Microbiology](#)
[Neurobiology](#)
[Physiology](#)
[Plant Biology](#)

[CORRECTIONS](#)

Find articles in this issue containing these words:

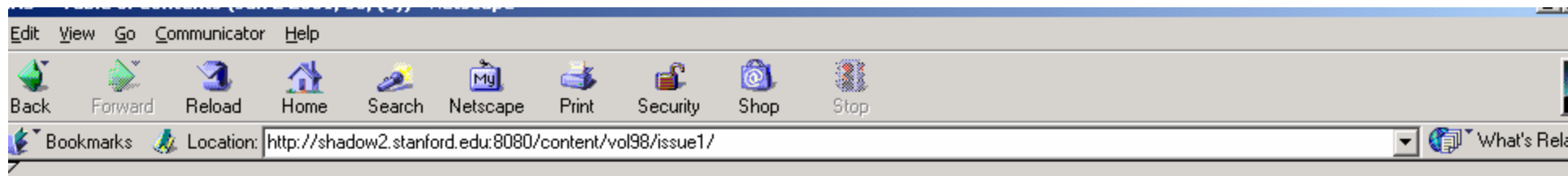
Enter

[\[Browse & Search All Issues\]](#)

To see an article, click its [\[Full Text\]](#) link. To review many abstracts, check the boxes to the left of the titles you want, and click the 'Get All Checked Abstract(s)' button. To see one abstract at a time, click its [\[Abstract\]](#) link.

[Get All Checked Abstract\(s\)](#)

Document: Page



NAS Online

[ME](#) [HELP](#) [FEEDBACK](#) [SUBSCRIPTIONS](#) [ARCHIVE](#) [SEARCH](#) [TABLE OF CONTENTS](#)

nt to be notified by email when new content goes on-line? [\[Sign up for eTOCs\]](#)

Other Issues: [←](#) [→](#)

Table of Contents: Jan 2 2001; 98 (1) [\[Index by Author\]](#) [\[Cover\]](#)

[COMMENTARIES](#)

[PERSPECTIVES](#)

[Physical Sciences:](#)

[Mathematics](#)

[Statistics](#)

[Social Sciences:](#)

[Anthropology](#)

[Biological Sciences:](#)

[Biochemistry](#)

[Biophysics](#)

[Cell Biology](#)

[Developmental Biology](#)

[Ecology](#)

[Evolution](#)

[Genetics](#)

[Immunology](#)

[Medical Sciences](#)

[Microbiology](#)

[Neurobiology](#)

[Physiology](#)

[Plant Biology](#)

[CORRECTIONS](#)

Find articles in this issue containing these words:

Enter

[\[Browse & Search All Issues\]](#)



What to Collect and Preserve?

- E-Journals
 - Titles you've paid for and are leasing
 - Freely available titles
- Other genres
 - Newspapers, Gov Docs

http delivered - serial - stable URLs
– authoritative version





Easy for publishers to participate

Publisher give permission (copyright materials) to:

- Libraries
- LOCKSS crawler

*Blanket license permissions
no individual library negotiations*





Publisher License

Permit libraries

- Collect materials as published for preservation
- Use material consistent with original license terms
- Provide copies for audit and repair to other caches only if they've had copy in the past



[HOME](#) [HELP](#) [FEEDBACK](#) [SUBSCRIPTIONS](#) [ARCHIVE](#) [SEARCH](#)

QUICK SEARCH:		[advanced]
Author:	Keyword(s):	
Go		
Year:	Vol:	Page:

Archive of 2003 Online Issues:

◀ 2003 ▶		
January January ; 51 (1): 1 - 132	February February ; 51 (2): 133 - 267	March March ; 51 (3): 271 - 404
April April ; 51 (4): 407 - 554	May May ; 51 (5): 555 - 696	June June ; 51 (6): 697 - 852
July July ; 51 (7): 853 - 980	August August ; 51 (8): 981 - 1112	September September ; 51 (9): 1113 - 1248
October October ; 51 (10): 1249 - 1391	November November ; 51 (11): 1393 - 1574	December December ; 51 (12): 1575 - 1712



LOCKSS system has permission to collect, preserve, and serve this Archival Unit



Distributed Repository Model Technology

Uses many “unreliable repositories” (PCs)

- Robustness through redundancy
- Inexpensive consumer hardware
- Low sys admin overhead (less 1 hour/mo)

Leverages web technology

- HTTP delivered and displayed content, all formats
- No need to replicate publisher's system
- Automated content ingestion over time

No single point of failure





Distributed Repository Model Business

Costs shared widely

- Total system is never a line item
- Low management overhead
- Low capital cost

IP issues simplified

- Straight forward blanket license terms
- No “negotiated” access
- Locally owned collections

No single point of failure

Budget cuts = key threat to long term access





LOCKSS and “Central Repositories”

Benefits

- System stability improves with some reliable peers
- Diversity improves reliability and attack-resistance

Requirements

- Implement LOCKSS repository interface
- Run system on mega-servers
- More metadata may be needed for access





LOCKSS Alliance

Publishers and libraries work together

- Define policies and best practice
- Develop and share technology
- Share core team costs
 - For limited time, to give model a chance
 - Contributions not required to participate, but
 - Critical amount of support required
 - Suggested contributions on web site





Taking Action

LOCKSS Program

- is in a nascent stage of development
- needs the community's support to go forward
- shows great promise

There are few actions librarians can take now to preserve digital information for future generations.

The risks of going forward are few. The risks of doing nothing are extremely high.





<http://lockss.stanford.edu>

L O T S O F C O P I E S K E E P S T U F F S A F E





Frequent Questions

OAIS

*Formal statement of Conformance to
ISO 14721:2003 May 2004*

Format Migration





Format Migration

Replacing web format takes a long time

- Both servers and browsers to be updated
- Society pays conversion for popular formats

During this long time we can

- Update cache software with converter
- Preserve content in original format
- Convert on output from old to new format
- Rewrite intra-journal links on output

.jpg to .png test conversion mid 2004





Metadata

Format metadata

- Collected from HTTP headers and the HTML
- Sufficient for browsers (now & near term)
- Demonstrate format migration based on this metadata
- Incorporate Harvard's JHOVE

Bibliographic metadata

- For Ingest OAI metadata crawler.
- For Export OAI metadata export capability
- Exploring automatically extracting OAI bibliographic metadata from the text





When HTTP is no longer supported as a protocol?

- Servers will export content using old and new transport protocol.
- LOCKSS caches can be upgraded to support both old and new transport protocols

No “flag day on web”

Long period of format overlap for common
formats



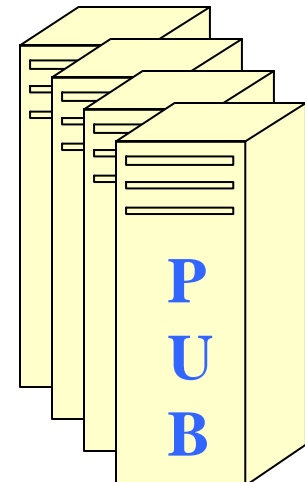
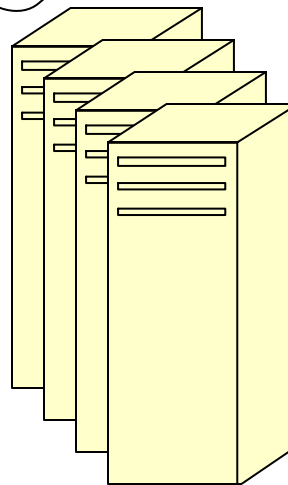
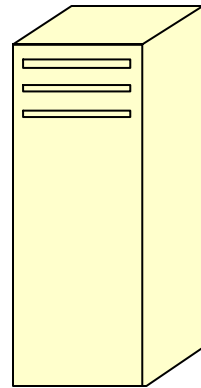


Collection Access LOCKSS and Local Networks *publisher is available*



**PAC File
or Proxy**

LOCKSS





Collection Access LOCKSS and Local Networks *publisher is unavailable*

