“Diffused Knowledge Immortalizes Itself”

Sir James Mackintosh 1765-1832
MOTIVATION

http://lockss.stanford.edu

LOTS OF COPIES KEEP STUFF SAFE
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 to 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 2\textsuperscript{nd} place all student research
• 7 other key research papers accepted so far

Investigating LOCKSS communication
• Scaling, attack resistance

Production Team:
\hspace{1em} deploys findings, builds system
LOCKSS Global Cache Monitor
Map View

Show the latest snapshot of LOCKSS Packet data.

Current as of: Sat Apr 13 00:00:00 PDT 2002.
Zoom in by clicking on the map.

Partners
50+ Publishers
90+ Libraries

<table>
<thead>
<tr>
<th>Descriptions of Published Papers</th>
<th>Status: Libraries and Endorsing Publishers</th>
<th>Technical Specifications</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Overview, Security, Network Integration, The Plug-in, OAI, Research - FAQ</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>April 5, 2004 - Production software released</th>
</tr>
</thead>
<tbody>
<tr>
<td>ask us to send you the software and instructions</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Librarians</th>
<th>LOCKSS Alliance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collection Development</td>
<td>Description</td>
</tr>
<tr>
<td>Humanities Project</td>
<td></td>
</tr>
<tr>
<td>Title Registry</td>
<td></td>
</tr>
<tr>
<td>User Interface Demo</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Related Work</th>
<th>Frequently Asked Questions</th>
<th>Government Documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Software License</td>
<td>Frequently Asked Questions</td>
<td>LOCKSS-DOCS</td>
</tr>
<tr>
<td>Credits Funders &amp; Technical Wizards</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Developers | |
|------------| |
| Software | License | |

| Support | |
|---------| |
| Help | |

| Resources | |
|----------| |
| Links | |
| e-news | |
| Research | |
| Funding | |
| Events | |
| Newsletters | |

| Contact | |
|---------| |
| Information | |
| Feedback | |
| Support | |
| Help | |
LOCKSS software turns a PC into a persistent web cache into a preservation tool

1 PC holds ~2,500 e-j years

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 (proxy servers)

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 (1,000 GB) = 2000 journal years

<table>
<thead>
<tr>
<th>Year</th>
<th>J-yr storage</th>
<th>TB/PC</th>
<th>J-yrs/PC</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004</td>
<td>$0.35</td>
<td>1.44</td>
<td>2,880</td>
</tr>
<tr>
<td>2005</td>
<td>$0.28</td>
<td>2.88</td>
<td>5,760</td>
</tr>
<tr>
<td>2006</td>
<td>$0.14</td>
<td>5.76</td>
<td>11,520</td>
</tr>
<tr>
<td>2007</td>
<td>$0.07</td>
<td>11.52</td>
<td>23,000</td>
</tr>
</tbody>
</table>
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 6.25" 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 MP3/E-U-Z video. Truly plug and play, this device requires no driver or software installation for Windows XP and later OS X users.

Triplet interface solution
But 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 capacity.
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
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
### Archive of 2003 Online Issues:

<table>
<thead>
<tr>
<th>Month</th>
<th>Volume</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>January</td>
<td>51 (1)</td>
<td>1 - 132</td>
</tr>
<tr>
<td>April</td>
<td>51 (4)</td>
<td>407 - 554</td>
</tr>
<tr>
<td>July</td>
<td>51 (7)</td>
<td>853 - 980</td>
</tr>
<tr>
<td>October</td>
<td>51 (10)</td>
<td>1249 - 1391</td>
</tr>
<tr>
<td>February</td>
<td>51 (2)</td>
<td>133 - 267</td>
</tr>
<tr>
<td>May</td>
<td>51 (5)</td>
<td>555 - 696</td>
</tr>
<tr>
<td>August</td>
<td>51 (8)</td>
<td>981 - 1112</td>
</tr>
<tr>
<td>September</td>
<td>51 (9)</td>
<td>1113 - 1248</td>
</tr>
<tr>
<td>March</td>
<td>51 (3)</td>
<td>271 - 404</td>
</tr>
<tr>
<td>June</td>
<td>51 (6)</td>
<td>697 - 952</td>
</tr>
<tr>
<td>November</td>
<td>51 (11)</td>
<td>1393 - 1574</td>
</tr>
<tr>
<td>December</td>
<td>51 (12)</td>
<td>1575 - 1712</td>
</tr>
</tbody>
</table>
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

LOTS OF COPIES KEEP STUFF SAFE
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

PUB
Collection Access
LOCKSS and Local Networks

*publisher is unavailable*

PAC File
or Proxy

LOCKSS