Preservation Tools For BLOGS

Victoria Reich
Director LOCKSS Program
Stanford University Libraries
Introduction

• This talk is in 3 parts
  – About LOCKSS
    • What is it?
    • How does it work?
  – Decentralized preservation
    • Authentication
    • Threats to digital content
  – How blogs could make preservation easier
Introduction

This talk is in 3 parts
  – About LOCKSS
    • What is it?
    • How does it work?
  – Decentralized preservation
    • Authentication
    • Threats to digital content
  – Making preservation easier
About LOCKSS

- An international community initiative
- Libraries taking charge
  - To continue their role as information stewards
  - As memory organizations in the digital environment

[Logo for LOCKSS]

www.lockss.org
A Bit More Detail...

- Distributed Digital Preservation System
  - Open source peer to peer software
  - Standards: OAIS, OpenURL, HTTP, WARC
- Stanford Libraries - administrative home
- Founded in 1998
- Supported by libraries - LOCKSS Alliance

www.lockss.org
Why Should You Care?

- Librarians use LOCKSS as digital stacks
  - Build and preserve e-collections
  - Subscription content - buy it once and keep it!
  - Open access - not locked up behind fee wall
- Provide 100% perpetual access
- Taking action for themselves
- Ensure library role-memory organization

www.lockss.org
What Is It Preserving?

• Authoritative versions of heterogeneous content
  – Journals, books, blogs, web sites, scanned files, audio, video …
• All Web formats and genres
  – Animations, datasets, moving images, still images, software, sound, text …
To Participate...

INSTALLING LOCKSS

Please read this page carefully before bringing your LOCKSS box on line. The page will tell you what you need to know before downloading and installing the software, and will lead you to instructions for doing so.

Bringing up a LOCKSS box is easy and does not require special technical expertise. The LOCKSS box works like an appliance (for example, a TIVO box is an appliance). Librarians have been bringing online LOCKSS boxes for eight years. The first time it may take you as little as an hour -- or it may take you a half day. If it takes longer than this -- contact lockss-support@lockss.org, something is wrong. Once your LOCKSS box is online it requires very little maintenance, about 15 minutes to 1 hour per month.

Watch this YouTube Video of how to bring a LOCKSS box online http://www.youtube.com/watch?v=0wdcnXrQkaI

The LOCKSS software turns a PC into a digital preservation appliance. An appliance does one thing, and because it only does one thing it can be easy to use. The key to making the LOCKSS appliance easy to use is that it boots and runs from a CD and stores all its configuration information on media such as a floppy disk that is write-locked while the system is running. There is no software of any kind, even an operating system, on the PC’s disk. The disk is used only to store preserved content.

The hardest things for a system administrator to do correctly are to install, configure and upgrade software so
How LOCKSS Works

• Publisher gives permission
  – By putting a Creative Commons license
  – Or a LOCKSS permission statement online

• Publishers include
  – Bloggers!
How LOCKSS Works

www.lockss.org
# How LOCKSS Works

## Archive-It ARC/WARC Data for Collection #903

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

<table>
<thead>
<tr>
<th>Archive File</th>
<th>MD5</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARCHIVEIT-903-STANFORD-MAC-200811613389-74012-crawling99.us.archive.org.arx</td>
<td>d6e6c0f236335f2c2b009372b015d90ef</td>
</tr>
<tr>
<td>ARCHIVEIT-903-STANFORD-MAC-200811613389-74012-crawling99.us.archive.org.arx</td>
<td>a0d5c595156d3e4ec9e9f8b7082d499</td>
</tr>
<tr>
<td>ARCHIVEIT-903-STANFORD-MAC-200811613389-74012-crawling99.us.archive.org.arx</td>
<td>3ba6e6b4f2d9f07c4eb667f1dd81a3532</td>
</tr>
<tr>
<td>ARCHIVEIT-903-STANFORD-MAC-200811613389-74012-crawling99.us.archive.org.arx</td>
<td>3233b0fe2e347e135d8b7082d3e7772d6</td>
</tr>
<tr>
<td>ARCHIVEIT-903-STANFORD-MAC-200811613389-74012-crawling99.us.archive.org.arx</td>
<td>e162739a965c2a5817b9656e54331ed9120</td>
</tr>
<tr>
<td>ARCHIVEIT-903-STANFORD-MAC-200811613389-74012-crawling99.us.archive.org.arx</td>
<td>4ecf599ac9e3d3a477b9649e3686e75d</td>
</tr>
<tr>
<td>ARCHIVEIT-903-STANFORD-MAC-200811613389-74012-crawling99.us.archive.org.arx</td>
<td>9393e4a960da2a62c1d06b04d70d704</td>
</tr>
<tr>
<td>ARCHIVEIT-903-STANFORD-MAC-200811613389-74012-crawling99.us.archive.org.arx</td>
<td>d6c5f023a27db4b4161f545b5c0c2d5</td>
</tr>
<tr>
<td>ARCHIVEIT-903-STANFORD-MAC-200811613389-74012-crawling99.us.archive.org.arx</td>
<td>8c2671d922714ed87c072487b3b12539</td>
</tr>
<tr>
<td>ARCHIVEIT-903-STANFORD-MAC-200811613389-74012-crawling99.us.archive.org.arx</td>
<td>7d0446d09e5640f6966551e25bc50</td>
</tr>
<tr>
<td>ARCHIVEIT-903-STANFORD-MAC-200811613389-74012-crawling99.us.archive.org.arx</td>
<td>e3e8d64a9ed1594e36022c8629e707</td>
</tr>
<tr>
<td>ARCHIVEIT-903-STANFORD-MAC-200811613389-74012-crawling99.us.archive.org.arx</td>
<td>7ef15807a5bf2e2671120d6dfe31646</td>
</tr>
<tr>
<td>ARCHIVEIT-903-STANFORD-MAC-200811613389-74012-crawling99.us.archive.org.arx</td>
<td>9c6f933536779704894943c5b86</td>
</tr>
<tr>
<td>ARCHIVEIT-903-STANFORD-MAC-200811613389-74012-crawling99.us.archive.org.arx</td>
<td>43690889a770aa427f6c64f11739b</td>
</tr>
<tr>
<td>ARCHIVEIT-903-STANFORD-MAC-200811613389-74012-crawling99.us.archive.org.arx</td>
<td>59b2c2ea6b773e92800e63983e0e7539</td>
</tr>
<tr>
<td>ARCHIVEIT-903-STANFORD-MAC-200811613389-74012-crawling99.us.archive.org.arx</td>
<td>26ed30ef3275242792906c6d29b021</td>
</tr>
<tr>
<td>ARCHIVEIT-903-STANFORD-MAC-200811613389-74012-crawling99.us.archive.org.arx</td>
<td>c4d5e0cb23eda72f2b6664ae9e5520e</td>
</tr>
<tr>
<td>ARCHIVEIT-903-STANFORD-MAC-200811613389-74012-crawling99.us.archive.org.arx</td>
<td>5a31ccf9a5e69e6f19537f5f792f92</td>
</tr>
<tr>
<td>ARCHIVEIT-903-STANFORD-MAC-200811613389-74012-crawling99.us.archive.org.arx</td>
<td>306b06e7b1814e614dd9680f7f97</td>
</tr>
<tr>
<td>ARCHIVEIT-903-STANFORD-MAC-200811613389-74012-crawling99.us.archive.org.arx</td>
<td>e080520307a55c11f64e0c9f9e</td>
</tr>
<tr>
<td>ARCHIVEIT-903-STANFORD-MAC-200811613389-74012-crawling99.us.archive.org.arx</td>
<td>47af6019d62e70d1d6e03e2f225b9d44</td>
</tr>
</tbody>
</table>

[www.lockss.org](http://www.lockss.org)
How LOCKSS Works

Publisher

HTTP Server
Presentation Files
LOCKSS Permission Statement

LOCKSS Network
LOCKSS Box 1
LOCKSS Box 2
LOCKSS Box 3

www.lockss.org
How LOCKSS Works
### Archival Units

<table>
<thead>
<tr>
<th>Volume</th>
<th>Content Size (MB)</th>
<th>Disk Usage (MB)</th>
<th>Peers</th>
<th>Agreement</th>
<th>Status</th>
<th>Last Poll</th>
</tr>
</thead>
<tbody>
<tr>
<td>Academie Psychiatry Volume 30</td>
<td>93,400,185</td>
<td>103.9</td>
<td>peers</td>
<td>0</td>
<td>Waiting for Poll</td>
<td>00:14:49</td>
</tr>
<tr>
<td>Accounting History Volume 5</td>
<td>3,125,086</td>
<td>5.4</td>
<td>peers</td>
<td>0</td>
<td>100.00% Agreement</td>
<td>02:30:36</td>
</tr>
<tr>
<td>Acta Sociologica Volume 48</td>
<td>7,992,857</td>
<td>11.4</td>
<td>peers</td>
<td>0</td>
<td>86.93% Agreement</td>
<td>00:05:26</td>
</tr>
<tr>
<td>Action Research Volume 3</td>
<td>6,443,633</td>
<td>9.6</td>
<td>peers</td>
<td>0</td>
<td>95.77% Agreement</td>
<td>12:47:23</td>
</tr>
<tr>
<td>Active Learning in Higher Education Volume 6</td>
<td>5,146,272</td>
<td>7.7</td>
<td>peers</td>
<td>0</td>
<td>Waiting for Poll</td>
<td>12:03:57</td>
</tr>
<tr>
<td>Adaptive Behavior Volume 13</td>
<td>13,264,401</td>
<td>23.4</td>
<td>peers</td>
<td>0</td>
<td>99.39% Agreement</td>
<td>17:11:49</td>
</tr>
</tbody>
</table>

1055 Archival Units
11 awaiting recalc
How LOCKSS Works

Article provided by the publisher

Article provided by LOCKSS

www.lockss.org
Two Kinds Of LOCKSS Networks

• **Public** LOCKSS network
• **Private** LOCKSS networks
  – Communities share preservation responsibilities
    • You preserve my stuff, I’ll preserve yours
  – Arizona State Library and Archives - PeDALS
  – Data-PASS - social science data sets
  – U.S. Federal Government Documents

www.lockss.org
How LOCKSS Works

Box 1

Box 7

www.lockss.org
Introduction

• This talk is in 3 parts
  – About LOCKSS
    • What is it?
    • How does it work?
  – Decentralized preservation
    • Authentication
    • Threats to digital content
  – Making preservation easier

www.lockss.org
Centralized Collection & Preservation

- Hub and spoke model
  - Change at hub propagates along spokes
  - Single point of corruption
- Centralized authentication
  - Single point of subversion
  - Only publisher can sign web pages
  - Too expensive for blogs and dynamic content
  - Public key infrastructure lacking
Decentralized Collection & Preservation

• Independent collection, independent replicas
  – Very hard to change all copies w/out being caught
    • ACM research award for showing how to make this true

• Distributed ‘authentication’
  • Compare multiple independent collections
    – If consensus, it’s what publisher meant, else:
      • Still changing? (e.g. blog comments, stop after a while)
      • Defective?
Threat Model

• What are you preserving stuff from?
• Many preservation models ignore threats known to cause data loss

http://www.dlib.org/dlib/november05/rosenthal/11rosenthal.html
Threat Model

- Media failure
- Hardware failure
- Software failure
- Network failure
- Obsolescence
- Natural Disaster

But…. 
Threat Model

- Media failure
- Hardware failure
- Software failure
- Network failure
- Obsolescence
- Natural Disaster

- Operator error
- External Attack
- Insider Attack
- Economic Failure
- Organization Failure
Introduction

• This talk is in 3 parts
  – About LOCKSS
    • What is it?
    • How does it work?
  – Decentralized preservation
    • Authentication
    • Threats to digital content
  – Making preservation easier
Making Preservation Easier

- Use Creative Commons (CC) license
  - Permits archiving through time
  - By any interested institution
  - Facilitates archive interoperability
- You may not think it worth preserving
  - CC allows others to disagree
  - Without having your lawyers & their lawyers talk
    - 1 lawyer-hour > 5TB of storage.
Making Preservation Easier

- Use meaningful URLs
  - Canonical, hierarchical
  - Don’t use session IDs in links, put in cookies
  - Blogger does this right
- Dynamic elements add complexity
  - What canonical content must be preserved, ads?
  - Use semantic markup & style sheets
- Don't put content behind a fee wall
Questions Welcome!

Thank you.