USING CLOCKSS TO KEEP STUFF SAFE ACROSS TIME
Using CLOCKSS To Keep Stuff Safe Across Time

Peter Burnhill
Director, EDINA
University of Edinburgh
Member, CLOCKSS Board
www.clockss.org
Overview of Talk

• Introduction
• CLOCKSS
• Questions

• Time permitting, two other related activities
  1. LOCKSS Alliance
     – Empowering libraries to act for local content
  2. Preservation Registry Service (project with ISSN-IC)
     – Who is looking after what?
Some things that Information Services at University of Edinburgh does

1. National and International Engagement:
   • EDINA National Data Centre
     • Developing and Delivering National Online Services
     • Projects to enrich Integrated Information Environment
     • Technical Support to UK Access Management Federation
   • Digital Curation Centre
   • JANET Video Conference Service

2. Supporting a World-class University:
research, learning & teaching in UK universities & colleges

NDCs acting as two platforms for network-level services as part of JISC Integrated Information Environment

Digital Content & Metadata

National Data Centres

EDINA
Mimas

JISC Collections

JISC Sub-Committees

Tools & Infrastructure

UK funding councils for HE & FE

UK Research Councils
Some things EDINA does: scholarly communication

SUNCAT
UK serials union catalogue

the Depot
national facility assisting Open Access deposit of peer-reviewed papers
- links Institutional Repositories

National OpenURL Router
- links OpenURL resolvers
Looking at CLOCKSS

• As pilot project
• Now in operation
  – how CLOCKSS works
  – CLOCKSS uses LOCKSS in a ‘C’ way
• How you can help CLOCKSS succeed
What’s the Problem?

• First, the Good News!
  – Researchers and students now have online access to journal articles
    • to read & download: Any-where, Any-time …

• Next, the Bad News!
  – What is now in digital form may not always be available
    • computing failure
    • natural disaster [earthquake, flood or fire]
    • human folly [criminal/political action; financial loss; stupidity]
  – Stops ‘tipping point’ from print to online
    • Frustrates economic benefits of existing investment in digital
    • Not good for libraries, not good for publishers
Some Consequences of Web

• Essentials of supply chain have changed
  • licensed to access, not sale of content
• Libraries no longer take physical custody of much key content
  • online remotely, not on-shelf locally
• Role of libraries as trusted keepers of information and culture has been disrupted
  – Need assurance of continuity of access
    • of all content for future generations
    • of the back copies, post-cancellation of the licence
• Scholarly, cultural & intellectual heritage is at risk
What’s the Answer?

1. **Think**: Understand how we ensured continuing access to printed works over the long term
   - Human-readable format; *relatively* enduring media (paper)
   - Multiple copies held in multiple places (a network of libraries)

2. **Think again**: Understand what is different about the digital
   - Formats become obsolete; unseen digital decay (‘bit rot’)
   - Can easily be altered (authenticity), copied and transported (theft)

3. **Propose**: Develop digital preservation policy & practices that address threats & risks

4. **Act**: Implement policy & practices for global effect
   - Need to command consensus across stakeholders (Transparency)
   - Need to be sustainable, in organisational, technical & financial terms

5. **Reflect**: Test, monitor and report: Community & Transparency
Two Schemes

1. LOCKSS - ‘Lots of Copies Keeps Stuff Safe’
   a) Open source technology developed at Stanford University
      - a scheme for slowly checking integrity of information [tortoise]
   b) Organisational ‘franchise’ to empower libraries to be able to safe-guard collections of interest
      - focus on perpetual access for licensed/subscribed content

2. CLOCKSS
   • Collaborative action by publisher and library communities
     - ‘C’ for collaborative/closed/controlled, shared governance
     - Initially a two-year project, from February 2005
   • Uses LOCKSS technology in private dark network
   • Comprehensive target: ingest of publishers’ total content
     - focus on long-term and ‘open’ release in event of ‘trigger event’
Mission

“Ensuring access to published scholarly content over time
… a community-governed partnership of publishers and libraries
… working to achieve a sustainable and globally distributed archive.”
How CLOCKSS Works

• The CLOCKSS Archive Network has several Nodes
  – two (2) computer servers per node.
• Each of these ‘CLOCKSS Boxes’ is ‘dark’
  – secure machine-to-machine interface, configured with LOCKSS software.
• Journal content from publisher sites is routinely ingested
  – distributed to every CLOCKSS Box.
• Using the LOCKSS software, these Boxes automatically and continuously chat to one another across the Internet
  – monitoring and self-correcting the preserved content
  – ensuring authenticity over the very long term.
• When the Board determines a trigger event has occurred, the relevant content is moved to a CLOCKSS Hosting Platform
• Orphaned journal content is made available for free to the world.
Library Organisations in Pilot

7 Nodes in CLOCKSS Archive Network
each with two ‘CLOCKSS Boxes’:
- Indiana University
- New York Public Library
- OCLC
- Rice University
- Stanford University
- University of Edinburgh
- University of Virginia

Operational CLOCKSS will have about 12 Archive Nodes
- globally located across geo-, political-, and legal- boundaries
Publishers in CLOCKSS Pilot

American Chemical Society
American Medical Association
American Physiological Society
Elsevier
Nature Publishing
Taylor & Francis
Wiley Blackwell

+ Waiting list to join operational CLOCKSS
First Recipient of Outstanding Collaboration Award

In 2007
Governance

A not-for-profit legal entity
• 501(c)(3) company based in California, USA

Three-tiered structure:
1. CLOCKSS Board
   • Meeting twice a month (by tele-conference)
2. Executive Committee
   • Elected by Board
3. Council of Members
Board Responsibilities

• Build Community
  – Libraries, Publishers & other stakeholders

• Oversee Operation
  – Stewardship of preserved content
  – Technology watch

• Manage Trigger Events and hosted content

• Promote digital preservation practices

• Build and manage Endowment & revenue
Defining a ‘Trigger Event’

When title, or part of, is no longer available

• Publisher ceases operations
  – Titles not available from any other source

• Publisher ceases to publish a title
  – Title not offered elsewhere

• Publisher removes back issues
  – Content not offered elsewhere

• Publisher’s delivery platform fails for a sustained period.
Managing a Trigger Event

1. Decision taken at Board level
   – Requires 75% SuperMajority vote; no single veto

2. Transferred to Host Platform
   – Treated as though ‘out of copyright’

3. Made available free-to-web
   – No authentication required
Testing the Trigger Process

Recent decisions by Sage Publications gave CLOCKSS opportunity to practice, discover and test

   - 3 Volumes of Web-rendered content ingested into CLOCKSS.
     - test decision process and transfer to Hosts for open access.
     - Article from this content needed to be ‘retracted’
       - authors declared data to be incorrect.

2. *Auto/Biography.* international and interdisciplinary journal addressing theoretical, epistemological, and empirical issues relating to autobiographical and biographical research.
   - Content received from SAGE Publications
     - preserved in CLOCKSS as XML, PDF and other formats.
   - HTML representation had to be generated from the XML files
     - using XSLT with preserved PDF files added.
Graft Public Copies

Free, Public Access to Journal, Graft

The recent decision by SAGE Publications to discontinue its journal, Graft: Organ and Cell Transplantation, provides the CLOCKSS Initiative with an opportunity to show how CLOCKSS works. In doing so, CLOCKSS offers continuing and public access to all the SAGE-published articles (three volumes from 2001 to 2003) of Graft that are preserved in the CLOCKSS archive.

Follow the links below to access this material at either of the two CLOCKSS hosting platforms based in Europe and the U.S. (for use worldwide, free, and without need of subscription):

- EDINA (University of Edinburgh):
  - Volume 4 (2001)
  - Volume 5 (2002)
  - Volume 6 (2003)

- Stanford University Libraries:
  - Volume 4 (2001)
  - Volume 5 (2002)
  - Volume 6 (2003)

The Graft content is copyright SAGE Publications and is licensed under a Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 United States License.
PDF Version

Automatic download
BEGIN manual download

Downloading the PDF version of:
Graft Cooper 4 (1): 6. (712K)

This file is in Adobe Acrobat (PDF) format. If you have not installed and configure Reader on your system, see Help with Printing for instructions.

Having trouble reading a PDF?

PDFs are designed to be printed out and read, but if you prefer to read them online, you increase the view size to 125%.
Graft
http://gft.sagepub.com

Xenotransplantation-A Closer Look
David K.C. Cooper
Graft 2001; 4: 6

The online version of this article can be found at:
http://gft.sagepub.com

Published by:
SAGE Publications
http://www.sagepublications.com

Downloaded from http://gft.sagepub.com at LOCKSS on December 9, 2007
© 2001 SAGE Publications. All rights reserved. Not for commercial use or unauthorized distribution.
Sustainability (1)

Control present costs and predict future costs

• Leverage existing technology
  – Storage costs are falling
  – LOCKSS preserves Web-published content now
  – Defer re-formatting costs until when needed

• Leverage existing infrastructure
  – University Research Libraries as Stewards
  – Internet allows multi-location and tele-communication
Sustainability (£)

Once you end preservation operation, you risk all
‘Free-to-Web’ service requires different model

1. Raise an endowment [a capital fund]
   • Long term digital preservation should not wholly depend upon recurrent revenue raising
   • as economic times get tough, preservation unlikely to be priority.

2. Fees from both sectors
   • Continue volume-related fees for ingest
   • Use Library fees to contribute to endowment
   • End or lower annual fees after 5 years
## Revenue from Publishers

### Mixed Model: Turnover + Ingest per article

<table>
<thead>
<tr>
<th>Revenue ($m)</th>
<th>Fees (US $)</th>
<th>Ingest Fee: $0.25/article</th>
<th>Max Fee $75,000/year</th>
</tr>
</thead>
<tbody>
<tr>
<td>200+</td>
<td>25,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50 - 200</td>
<td>15,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 - 50</td>
<td>9,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 - 10</td>
<td>5,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 - 5</td>
<td>2,500</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 1</td>
<td>1,000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Back File Ingest is FREE
Revenue from Library Sector

<table>
<thead>
<tr>
<th>Revenue ($m)</th>
<th>Fees (US $)</th>
</tr>
</thead>
<tbody>
<tr>
<td>25 - 30</td>
<td>15,000</td>
</tr>
<tr>
<td>20 - 25</td>
<td>12,000</td>
</tr>
<tr>
<td>15 - 20</td>
<td>9,000</td>
</tr>
<tr>
<td>13 - 15</td>
<td>7,800</td>
</tr>
<tr>
<td>11 - 13</td>
<td>6,600</td>
</tr>
<tr>
<td>9 - 11</td>
<td>5,400</td>
</tr>
<tr>
<td>7 - 9</td>
<td>4,200</td>
</tr>
<tr>
<td>5 - 7</td>
<td>3,000</td>
</tr>
</tbody>
</table>

About 0.05% of a Library’s Materials Budget
Twelve Things About

1. **Collaborative**: This joint initiative by publisher and library communities won the ALA's Association for Library Collections & Technical Services (ALCTS) Outstanding Collaboration Award in 2007

2. **Governance**: Publishers and librarians work as equals in shared decision making on Council, Board and Executive

3. **Global**: Network of globally distributed archive nodes spans geographic, political and legal ‘tectonic plates’ and boundaries, not relying upon legal deposit in each country

4. **Comprehensive**: Aim is to have all publishers’ content routinely ingested, starting with journal content, including branding and publisher’s look and feel

5. **Stewardship**: ‘CLOCKSS Boxes’ of journal content in Archive Nodes located in established research library organisations

6. **Dark**: Digital content is held securely, in trust, closed until there is agreed trigger event
Twelve Things About

7. **Access For All.** Content that is deemed ‘orphaned’ or otherwise suitable for release via a CLOCKSS Host, is made available free to the public, without need for any prior subscription, fee or registration.

8. **Robust & Resilience.** LOCKSS technology, for continuous and systematic audit and repair, is proven open source software acknowledged by ACM Award in 2004.

9. **Sustainable business model.** Key role for five-year plan to build financial endowment in order to reduce dependence upon recurrent revenue in tough economic times.

10. **Cost effective.** Defer format migration until content is triggered, saving front loaded re-formatting costs for all ingested content.

11. **Advocacy.** Do not wait until the Eleventh Hour. Add your voice in favour of digital preservation to ensure long-term access to scholarly content.

12. **Act Now.** You can get involved. CLOCKSS needs your support in order to fulfill its mission for your future scholars. Letter of Intent at [www.clockss.org](http://www.clockss.org). Early supporters will be assigned charter status.
My time has ticked by …

Questions welcome

p.burnhill@ed.ac.uk  edina.ac.uk

Info@clockss.org  www.clockss.org
Extra Time

Two related activities

1. LOCKSS Alliance
   • Empowering libraries to act for local content

2. Preservation Registry Service
   • Who is looking after what?
LOCKSS Alliance

• Empowers libraries to build and preserve collections of interest

http://www.lockss.org
UK LOCKSS Alliance

- JISC & CURL/RLUK funded 2-year pilot
- Self-funding membership started in August
- Technical support based at EDINA
2. Preservation Registry?

• Many objects need preserving; many schemes emerging
• How can libraries & policy-makers assess who is doing what, for what, and how?
• JISC funded a scoping study into e-journals preservation registry
  • Rightscom / Loughborough University, 2007
    – Confirmed expressed need among libraries
    – Warned of potential burden on digital preservation agencies
    – Recommended that UK Union Catalogue of Serials (SUNCAT) or SHERPA (Open Access) get involved.
  • SUNCAT is hosted and managed at EDINA
Piloting an E-Journals Preservation Registry Service (PEPRS)

Two year project, starting August 2008.
- Scope, develop & test a registry service
- Establish and test an Information Architecture
- Seek consensus across stakeholders
- Technical & financial sustainability

*Partners:* EDINA and ISSN International Centre (Paris)

- Funded by JISC
  - with review in 18 months about transition into service
- Support of Council and Directors of ISSN Network
Early ideas about e-journals preservation registry service

*Only just begun, but:*

- Use E-Journals Register, sourced from ISSN Register
  - Over 50,000 e-journals now have ISSN

- Need to agree what users want to know
  - descriptors of digital preservation policy & practices

- Use network interoperability (to search or to harvest)
  - for up-to-date, reliable information held by preservation agencies on and statements about policies and coverage

- ‘Titles’ is easy, but ‘Holdings’ is difficult!
  - role for DOI and Onix for Serials

- Make sure that the e-journals you care about get an ISSN identifier!
  - The Directory of Open Access Journals (DOAJ) requires it
One Moment …
Format Obsolescence

“If a format is widely adopted, it is less likely to become obsolete rapidly, and tools for migration and emulation are more likely to emerge from industry without specific investment by archival institutions….Evidence of wide adoption of a digital format includes bundling of tools with personal computers, native support in Web browsers.”
  http://www.digitalpreservation.gov/formats/sustain/sustain.shtml

Web formats become obsolete when the majority of browsers no longer render that format.
Why Format Migration “on the fly”

- Preserve historical context
  - Original look & feel
- Reduce the cost of ingest.
  - Preserve more material per dollar.
- Postpone costs of migration.
  - Technology costs less and money can be invested.
- Migrate material upon reader request.
  - Most material not used, most content not processed.

What the readers sees is the result of the best possible technology at time of access
Format Migration

“on the fly”
When content is requested
Process is transparent to the reader

http://www.dlib.org/dlib/january05/rosenthal/01rosenthal.html
Library Costs

- Software -- free
- Hardware -- basic PC
- Staff time <15 minutes/month
- Fees on sliding scale
- *Not* a subscription
# LOCKSS Alliance

<table>
<thead>
<tr>
<th>Institution Size</th>
<th>Dues/Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Research Universities (Very High Research Activity)</td>
<td>10,800</td>
</tr>
<tr>
<td>Research Universities (High Research Activity)</td>
<td>9,600</td>
</tr>
<tr>
<td>Doctoral/Research Universities</td>
<td>8,200</td>
</tr>
<tr>
<td>Master’s Colleges and Universities (Large Programs)</td>
<td>5,200</td>
</tr>
<tr>
<td>Master’s Colleges and Universities (Medium Programs)</td>
<td>4,443</td>
</tr>
<tr>
<td>Master’s Colleges and Universities (Small Programs)</td>
<td>3,685</td>
</tr>
<tr>
<td>Baccalaureate Colleges</td>
<td>2,160</td>
</tr>
<tr>
<td>Associate’s Colleges</td>
<td>1,080</td>
</tr>
</tbody>
</table>
Some Early Statistics

How readers found Stanford-hosted *Graft* content:

- Google: 38%
- Links: 33%
- OpenURL resolvers: 7%

Over two-thirds accessed via non-academic IPs

- 121 PDF downloads
- 163 readers