The CLOCKSS Archive Architecture

David S. H. Rosenthal

LOCKSS Program, Stanford Libraries

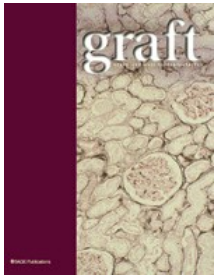
The CLOCKSS Archive – Digital Preservation

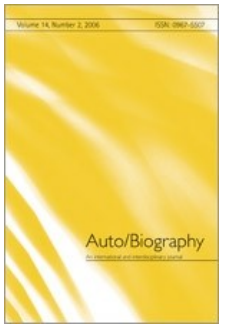
CLOCKSS is a **dark archive** founded by the world's leading international libraries and publishers to keep **digital preservation** in the hands of the **community**.

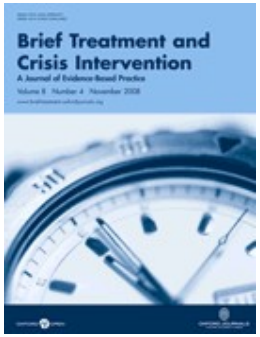
1. Community-governed archive
   - The Board is 50% libraries and 50% publishers.
2. Globally distributed libraries preserving content.
   - Geographically, geologically, politically distributed.
3. Low fees to encourage participation.
   - Leverage library infrastructure.
   - Using LOCKSS technology for preservation.
4. Triggered content is made available to the whole community to preserve continued access to abandoned/orphaned content.

One Key Value:
Open Access “Triggered” Content

- Graft
  - Sage
- Auto/Biography
  - Sage
- Brief Treatment & Crisis Intervention
  - OUP







CLOCKSS

Collaborative Partners – 12 Libraries

Asia/Pacific

Australia: ANU

China: University of Hong Kong

Japan: NII

Europe

Germany: Humboldt University

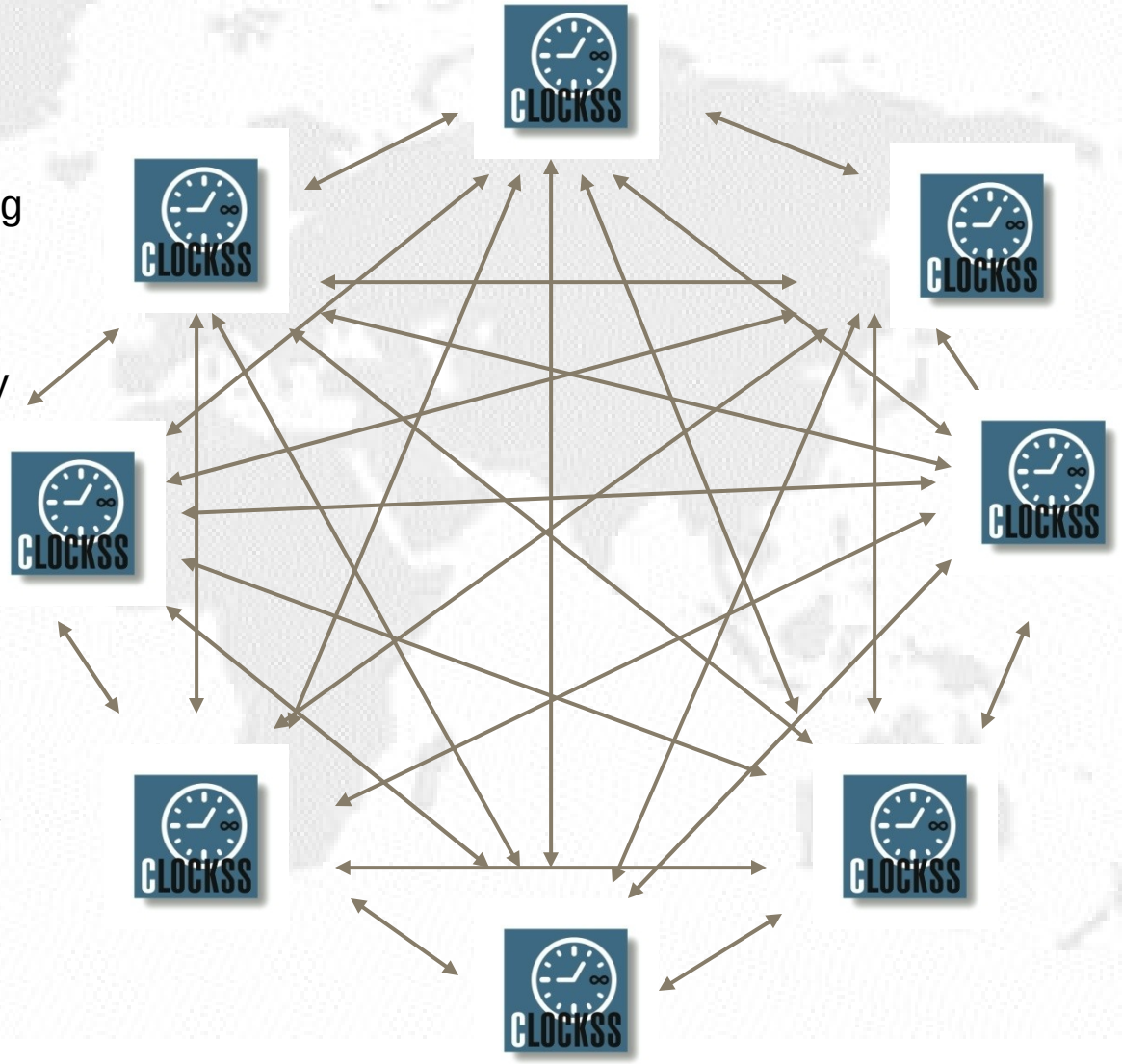
UK: University of Edinburgh

Italy: Università Cattolica del Sacro Cuore

North America

Canada: University of Alberta

United States: Indiana University, Rice University, Stanford University, University of Virginia, OCLC



Content Preserved – Journals & Books

American Academy of Pediatrics, American Anthropological Association, American Institute of Physics, American Medical Association, American Physiological Society, American Society of Civil Engineers, American Society for Pharmacology and Experimental Therapeutics, American University Washington College of Law, Association for Computing Machinery, bepress, BioMed Central, Bloomsbury Qatar Foundation, Boston College Law School, British Institute of Radiology, Casalini Libri, Clueb, Co-Action Publishing, Edinburgh University Press, Elsevier, European Respiratory Society, Fabrizio Serra editore, Geological Society of America, International Union of Crystallography (IUCr), IOP Publishing, Liverpool University Press, Maney Publishing, National Academy of Sciences, Nature Publishing Group, Oxford University Press, Pacific Affairs, Pion Ltd, Rockefeller University Press, Royal Society Publishing, RSC Publishing, SAGE Publications, Society for the Study of Reproduction, Springer, Taylor & Francis, Thieme Publishing Group, Wiley-Blackwell.

Governed By The Community

- Board of Directors
- Advisory Council

2007 ALA ALCTS
Outstanding Collaboration

CLOCKSS is a tax-exempt, 501(c)3, not-for-profit organization

Tasks

- Ingest
  - Presentation Content via web crawling
  - “Source” Content via FTP from publisher
- Preserve
  - Network of up to 15 CLOCKSS boxes gets ingested content
  - Continuously audit between boxes and repair any damage
- Disseminate
  - Triggered presentation content extracted, links rewritten
  - Triggered source content XML rendered, web structure created

Ingest

- Presentation content
  - Small network of “ingest” LOCKSS boxes crawls publisher's web site
  - Vote on what they collected, resolve differences via re-crawl
    - CLOCKSS knows it has what a reader's browser would have seen
  - After agreement, export content to CLOCKSS network
  - CLOCKSS boxes crawl using “ingest” boxes as proxy
- “Source” content
  - Publisher packages content in their favorite format, puts on FTP server
    - Varies by publisher: XML + PDF + metadata, PDF + metadata, ...
  - Source ingest machine collects it, checks it, exports to CLOCKSS boxes
    - CLOCKSS knows it has what the publisher wanted it to have
  - CLOCKSS boxes crawl the source ingest machine

Dissemination

- There is no hurry
  - Mandatory 6-month delay for board to approve trigger
  - Extract triggered content by crawling a CLOCKSS box
  - Process, add CC license, then export via Apache
- Presentation content
  - Rewrite internal links
- “Source” content
  - If available, render XML to HTML for abstracts, full-text, ...
  - Create web structure (ToC pages, etc)
  - Insert HTML (if available), PDF (if available)

What's Cool About CLOCKSS

- Free, open access to 'triggered' archived content
  - Keep open access content, open access over time
  - Good for authors, good for societies, good for scholars
- Community-governed archive
  - Librarians and publishers work together as equals
- Globally distributed libraries preserving content
  - Geo-graphically, geologically, geo-politically
  - Re-enforce library's memory role on a worldwide scale
- Low costs
  - Leverage library infrastructure
  - Using LOCKSS for preservation

Thank you!

David S. H. Rosenthal
LOCKSS Program
Stanford University Libraries

www.clockss.org