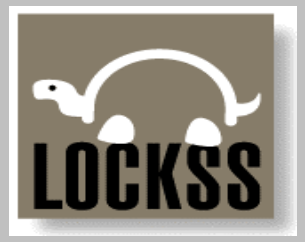


LOCKSS Project:
An Overview and Consideration of its Applicability to Government Information

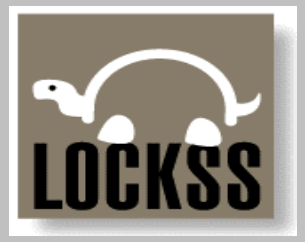
Chuck Eckman

Depository Library Conference

October 17, 2001

Definitions

- Archive
  - A place in which public records or other important historic documents are kept.
- Library
  - A place set apart to contain books, journals, etc. for reading, study, or reference.
- Preservation
  - The action of keeping from injury or destruction.

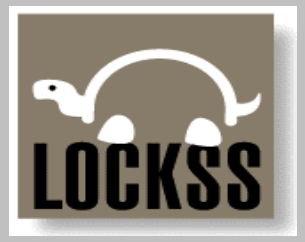


LOCKSS

- Stands for: *Lots of Copies Keep Stuff Safe*
- Is not an archive
  - Archives for hard to replicate materials
  - Sacrifice access to ensure preservation
- Could be described as a global library system
  - Libraries for easily replicated materials
  - Ensure access at some preservation risk

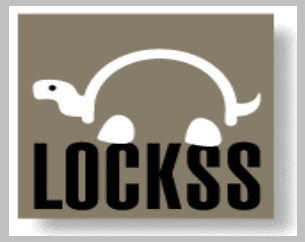
Overview of LOCKSS

- LOCKSS turns a pc into a browsable, low maintenance cache of online content
- Caches at different institutions communicate to repair damage and maintain content integrity
- “Digital Preservation Internet Appliance”
- E-Journals are initial testbed



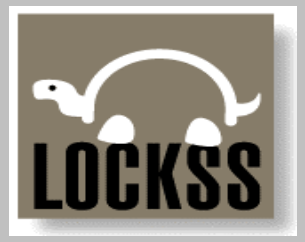
Librarians Role in Paper Environment

- Distribute and house copies worldwide
- Loan copies to libraries on request
- Readers find a copy easily
- It is hard to find and destroy all copies
- Implicit guarantors of authenticity



Librarians Prevent "Un-publication"

- Publisher takeovers, buyouts
- Malicious act
- Natural disaster
- Loss or Disappearance
- Official edict



The Problem: Electronic Journals

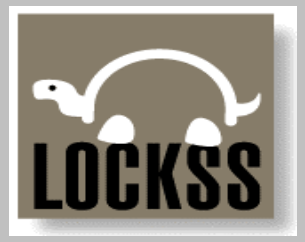
- E-Journals are the version of record
  - Online has more peer-reviewed text, non-text content
  - Paper journals are no longer the record of scholarship
- Online services (linking, searching) important to discovery process
- E-Content Easily “Unpublished”
  - Leased, not owned
  - Publishers promise “perpetual access”, unfounded

Publisher Concerns

- Brand and Services
  - Maintain brand and journal image
  - Have material available for future society members and other subscribers
  - Encourage librarians to initiate online access
- Security
  - Prevent illegal replication of content
  - Enforce access control
  - Maintain content integrity = original published version
- Direct access to readers
  - Collect use statistics
  - Exploit/explore new payment models

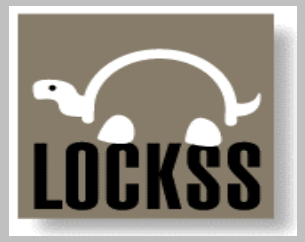
Librarian Concerns

- Custody of content
  - Have material available for future community members in perpetuity
    - Subscription cancellations, publisher buyouts
- Maintain services
  - Links resolve, full text integrated into local search environment
- Security
  - Maintain content integrity = original published version
- Direct access to readers
  - Collect use statistics



Solution: Global Online Library

- Where the content
  - Is locally stored & managed
  - Maintains integrity, functionality, accessibility
- Where the system
  - Has no central authority/point of failure
  - Tolerates faults, including attacks



LOCKSS Technical Requirements

- Be affordable
  - Cheap PC, open-source software
  - Low administration “appliance”
- Have low probability of failure
  - Many replicas, resists attack, no secrets
  - Scale to enormous rates of publishing
- Preserve access
  - Links resolve, searches work
  - Conform to publishers access controls
- Libraries take custody of content

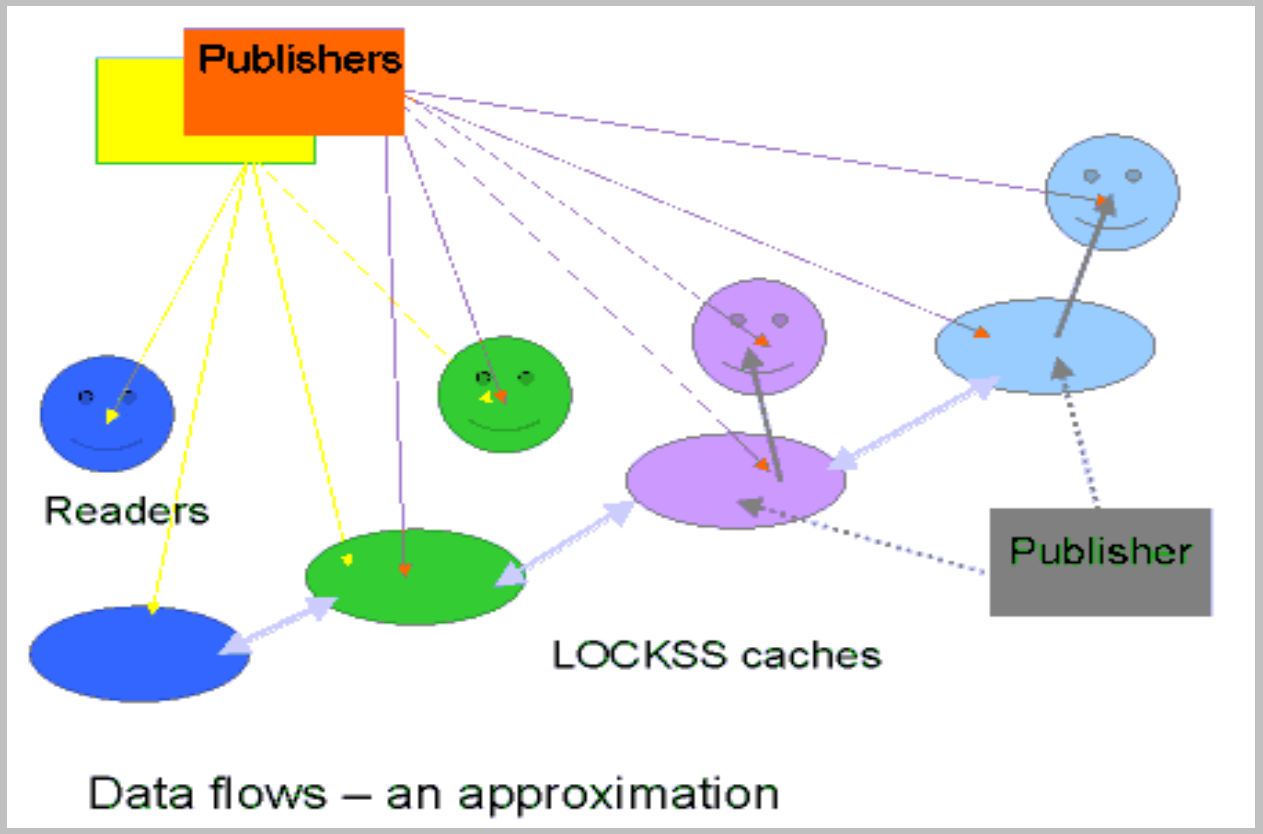
LOCKSS Works for Content

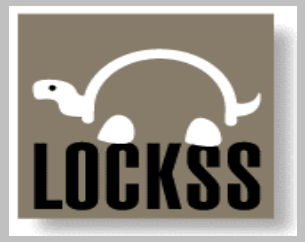
- Multiple Formats
  - gif, jpeg, html, pdf, video, audio
- Delivered through http
- That has an authoritative version
  - Not intended for dynamic content
  - Good match for peer-reviewed articles (and perhaps some government documents)

How the Data Flows

- Libraries install LOCKSS web caches that
  - Never get flushed
  - Hold authorized content
- These caches
  - Pre-fetch content as published
  - Continuously validate against other caches
  - Repair gaps from publisher and other caches
- Readers access content
  - From publisher servers or their local caches (depending on configuration)

Data Flows


Publishers
Readers
Publisher
LOCKSS caches
Data flows – an approximation



LOCKSS Project Status

- Support
  - National Science Foundation
  - Sun Microsystems Labs
  - Stanford University Libraries
  - Mellon Foundation
- Software
  - Technical design complete
  - Prototype working
  - Alpha test ended 4/5/01
  - Beta test started 4/6/01

Beta Test Libraries, 1

Africa/Middle East

- Israel: Hebrew University
- South Africa: University of Stellenbosch

Asia/Pacific

- Australia: University of Melbourne
- Hong Kong: University of Science & Technology
- New Zealand: University of Auckland, Univeristy of Otago Singapore: National University

Europe

- Belgium: University of Ghent
- England: British National Library, Cambridge University, Imperial College, University of Leeds
- Germany: University of Munich
- Finland: Helsinki University of Technology
- Italy: IEI-CNR: Italian National Council of Research Netherlands: Koninlijke Bibliotheek, University of Maastricht, University of Amsterdam
- Norway: University of Bergen Scotland: Edinburgh University, University of Glasgow
- Spain: University of Alicante Sweden: Lund University

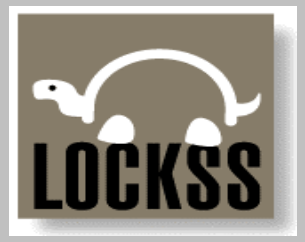
Best Test Libraries, 2

North America

- Canada: University of Toronto
- United States: Stanford University, University of California Berkeley, Columbia University, University of Tennessee, Los Alamos National Laboratory, Harvard University, Carnegie Mellon University, Cornell University, Emory University, Indiana University, Library of Congress, University of Chicago, University of Minnesota
- United States (cont'd): University of Texas Austin,Yale University, University of Oklahoma Health Science Center, University of Nevada Reno, New York Public Library, Case Western Reserve University, Iowa State University, National Agricultural Library, University of Virginia

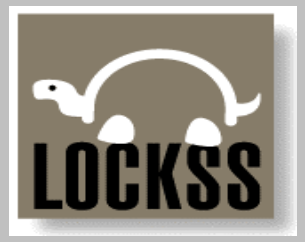
South America

- Brazil: BIREME (the Latin American and Caribbean Centre on Health Sciences Information)



Best Test Publishers

- American Association for the Advancement of Science, American Physical Society, American Physiological Society, Federation of American Societies for Experimental Biology, Biophysical Society, Annual Reviews, Rockefeller University Press, The Endocrine Society, American Society for Biochemistry and Molecular Biology, American Association for Clinical Chemistry, National Academy of Sciences, British Medical Journal, American Psychiatric Publishing Inc., Oxford University Press, Company of Biologists Ltd, New England Journal of Medicine, American Society for Clinical Investigation, Radiological Society of North America, Society for General Microbiology, The Histochemical Society, American Thoracic Society, BMJ Publishing Group, American Society of Neuroradiology, Lipid Research Inc., American Society for Investigative Pathology, American Society of Plant Physiologists, The Royal College of Psychiatrists, Society for the Study of Reproduction, American Society for Microbiology, Cold Spring Harbor Lab Press, American Society for Pharmacology & Experimental Therapeutics, Society for Molecular Biology and Evolution, American Society for Nutritional Sciences, BioMedCentral, Genetics Society of America, Investigative Ophthalmology and Visual Science, Botanical Society of America, American Heart Association, American Society of Hematology, The American Physical Society



Beta Test Status

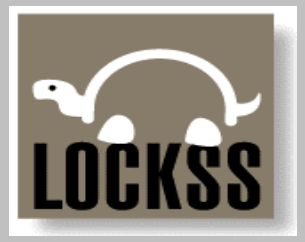
- What's worked so far
  - 60+ machines have collected a static test journal
  - Files missing during collection identified and repaired
- What we have learned
  - Polling protocol works at 60+ scale
    - Difficulties with "transparent" web proxies
    - UI & instructions need improvement
- What's next
  - Multiple, dynamic journals
  - Simulated small & large failures
  - Simulated "bad guy" attacks

LOCKSS & E-Documents

- Modeled on the depository program
- Contrast problem statement
- Contrast stakeholder interests
- Special content challenges
- Possible LOCKSS applications

The Problem: Electronic Documents

- E-Documents not necessarily the version of record
  - Legal issues are diminishing but remain
  - Budgetary pressures are enhancing status of e-documents
- Online services (linking, searching) important to discovery process
- E-Content Easily “Unpublished”
  - Accessed, not owned
  - Agencies promise “perpetual access”, unfounded

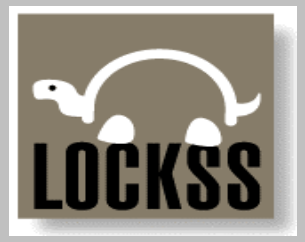


Agency Concerns

- Brands and services
  - Maintain “officialness” of content
  - Ensure access to future citizens and constituents
  - Encourage librarians to rely increasingly on electronic
- Security
  - Enforce access control (in limited circumstances)
  - Maintain content integrity = official and authorized version(s)
- Direct access to readers
  - Collect use statistics

Government Document Librarian Concerns

- Custody of content
  - Have material available for future community members in perpetuity
    - **agency closure, withdrawal of publication from agency server**
- Maintain services
  - Links resolve, full text integrated into local search environment
- Security
  - Maintain content integrity = all published versions
- Direct access to readers
  - Collect use statistics



E-Documents Challenge LOCKSS

- Publication patterns often irregular
  - Editions, monographs, corrections
- Frequency of update
  - From never to very high (dynamic databases)
- Wide range of data structure
  - Neither uniform, nor “vertical”
- Wide array of formats

Possible Applications

- Use LOCKSS for E-Documents with regular publishing frequencies (journals, annuals, biennials, dailies etc.)
- Test version-verification capabilities of LOCKSS on E-Documents
- Test monograph and irregular publication capabilities of LOCKSS with E-Documents
- ??

Further Information

- http://lockss.stanford.edu
- For information on participation as a Beta site, contact:
  - Vicky Reich, Assistant Director and Digital Librarian, HighWire Press, Stanford University Libraries and Academic Resources, 650.725.1134, vreich@stanford.edu. (Please copy Chuck Eckman on any correspondence ceckman@stanford.edu)