LOCKSS: ensuring access through time

Victoria Ann Reich
Diretora Executiva do Programa LOCKSS, Stanford University Library, [www.lockss.org]. O Programa LOCKSS capacita bibliotecas localmente arquivo e controlar a sua coleções arrendadas, que assegurem o acesso perpétuo, controle local e à propriedade, e para a saúde e força da comunidade biblioteca.
E-mail: vreich@stanford.edu

Resumo

Especialistas retratam a preservação digital como muito cara e muito complicada para as comunidades a executarem por conta própria. Eles ganham muito perpetrando esta mensagem e as comunidades têm muito a perder ao tomar como absolutas suas palavras. Comunidades querem garantir o acesso permanente ao conteúdo que foi adquirido, pois desejam assegurar que uma cópia de ativos intelectuais (pagos e de acesso livre), em que eles investiram ao longo do tempo permaneça em sua própria jurisdição. O Programa LOCKSS (Lot Of Copies Keep Stuff Safe), da Universidade de Stanford, auxilia comunidades a construir e preservar suas próprias coleções digitais, passo essencial para garantir o acesso ao longo do tempo.

Palavras-chave

LOCKSS. Universidade de Stanford. Preservação digital. Acesso pós-cancelamento. Acesso perpétuo.

LOCKSS: garantindo acesso ao longo do tempo

Abstract

Some experts portray digital preservation as very expensive and too complicated for communities to do for themselves. They have a lot to gain by perpetrating this message and communities have a lot to lose by taking them at their word. Communities want to ensure post-cancellation access to content they have purchased; they want to ensure that a copy of the intellectual assets (fee and open access) in which they have invested through time resides in their own jurisdiction. The Stanford University LOCKSS Program (Lots Of Copies Keep Stuff Safe) helps communities build and preserve their own digital collections, an essential step to ensuring access over time.

Keywords

LOCKSS. Stanford University. Digital Preservation. Post Cancellation Access. Perpetual Access.

INTRODUCTION

There are two major reasons to preserve electronic books and journals: to facilitate access and to maintain scholarship's integrity (ROSENTHAL, 2007a, b). While paper can survive for long periods of time with benign neglect, digital content must be preserved to remain accessible. Digital content preservation is a continuous, active process.

Digital preservation systems vary considerably, with different business, social and technical implementations. Communities are advised to take direct, local custody of the content to ensure their access to important works. The LOCKSS Program, based at Stanford University Libraries enables communities to take custody, to preserve, and to make available scholarly assets while at the same time upholding and enforcing participating publisher's access and business models.

OBSTACLES TO PRESERVATION

Digital content falls into two major classes: offline digital content, materials that are not on the web and online digital content, materials that are on the web. Some actions needed to preserve these two types of digital content are the same.

• Digital preservation must be an active process, the bits and bytes are fragile and need continuous auditing and repair;

• Digital media is stored on fragile, easily alterable medium with a short service life (disk or magnetic tape;

• Digital content can be easily altered with little or no trace of what was an authoritative version.

Web published digital content such as journals, books, datasets, thesis and dissertations have additional preservation challenges:

• The authentic version of the content, the version the reader sees is held centrally on publishers' web site. The publisher's, or the intellectual property holder's permission must be obtained before starting any preservation action. This is often a labor-intensive process that requires considerable persistence over long periods of time.

• The large commercial publishers have closed proprietary publishing systems that track users and control use. These mechanisms impose mechanisms, for example one-time urls that make preservation difficult.

• Many small publishers are open access. A good number of them publish via open source platforms, for example OJS, thus avoiding these problems of the large publishers. However often small publishers are burdened with fragile business models and publish small amounts of content. This makes the preservation costs difficult to justify.

CAUSES OF DAMAGE AND LOSS

The digital preservation field has spent many resources attending to the risk of format obsolescence (ROSENTHAL, 2009). A common method for addressing the problems of format obsolescence is a process called file normalization, which by its very action destroys the authoritative version of the content. Format obsolescence may be a problem for offline digital media. It has not yet proven to be a problem for online digital content.

For online content or web assets, a format becomes obsolete when commonly available browsers and plugins can no longer render it. There is little evidence that this is happening in web formats widely used for web published books, journals, government documents, thesis and dissertations, etc.

Evidence suggests that human factors, intentional and unintentional are the greatest cause of loss or corruption to digital materials (ROSENTHAL, 2011). Technology failures, economic failures and social failures also pose threats to digital content. When materials are held in a single centralized repository or when they are held in multiple repositories all administered by the same authority, it is easy for someone to tamper with the master copy without detection.

The Blue Ribbon Task Force on Sustainable Digital Preservation and Access identified economics as another major threat to preservation (BLUE RIBBON TASK FORCE ON SUSTAINABLE DIGITAL PRESERVATION AND ACCESS, 2010). Even with optimistic projections of future costs, there is not enough money to preserve everything that should be preserved (ROSENTHAL, 2012). This places particular importance on minimizing the cost of digital preservation systems. Every unnecessary dollar of cost means more content that will be lost.

The biggest threat to a communities' access to digital content is when communities do not have local control over the material. A community has no security when their continued access depends upon continued payment. They are at the mercy the content custodian's business models. Outsourcing responsibility for the safekeeping of important cultural and scholarly works is a risky proposition.

GUARDING AGAINST DAMAGE

Effective preservation systems are engineered to guard against these common causes of digital integrity and access loss. One way to implement safeguards against human folly, economic failures, and communities not having physical control over their collections is to build digital preservation systems that share these two attributes manifest in paper libraries:

• Communities take custody of content, separating payment from access and shielding them from publisher's and third party vendor's possible future business decisions;

• Each piece of content is replicated and copies are held under different administrative control. This makes it difficult for humans to alter copies of the content without detection, protecting the content from censorship or loss.

In the LOCKSS system, it is very difficult and expensive for someone to find and tamper with a significant number of the preserved copies without being caught. The copies are geographically distributed and independently held under many different administrations. Tamper-evidence engineering is a unique property of LOCKSS preservation and it is a keystone of our work (ROSENTHAL et al. 2005). LOCKSS' ACM award-winning open-source technology is built on a peer-to-peer software infrastructure (MANIATIS, 2003; STANFORD…, 2013c).

The LOCKSS Program protects digital content from this broad set of technical, economic and social threats.

THE LOCKSS PROGRAM (LOTS OF COPIES KEEP STUFF SAFE)

The LOCKSS program was founded in 1998 and is based at Stanford University Libraries. The Program is an open-source preservation system built on the principle that "lots of copies keep stuff safe." It balances communities' needs to protect their access to scholarship and the publisher's need to uphold their branding and keep users on their web sites.

The LOCKSS system is the world's first production quality Distributed Digital Preservation Network. Distributed Digital Preservation Networks have intentional geographical and organizational distributed infrastructures as essential design components and require community collaboration and cooperation for implementation. It is not possible to achieve the required robustness to ensure the long-term persistence of digital objects through centralized technical or organizational approaches. The LOCKSS Program set best practices for reliable preservation and persistent access to digital content via content replication, geographic distribution, infrastructure heterogeneity, modularity and organizational diversity.

The traditional LOCKSS model is for each participating institution to bring online their own LOCKSS box. They use their LOCKSS box to take custody of and preserve access to the digital content important for their community including open access and subscription journals and books, locally published dissertations, and imaged collections.

By using their own computers and network connections, institutions obtain, preserve and provide access to authorized copies of digital content. This process is analogous to libraries' using their own buildings, shelves and staff to obtain, preserve and provide access to paper content.

Stanford University Libraries LOCKSS Program in conjunction with several partners are investigating a new LOCKSS model where consortia bring online their own LOCKSS network to serve all their members.

Several consortia have asked the LOCKSS Program to consider implementing a regional or national level LOCKSS network hosted by a library consortia and/or the government. These LOCKSS networks would preserve open access and subscription journals, books, and other materials for long-term post-cancellation access. For some countries, a LOCKSS network is being considered in addition to other complimentary national collection building approaches.

Consortia and national groups are spending large amounts of money for access, and want to ensure post-cancellation access. They are investing in open

access content and wish to ensure this material is also available for the long term. Given global political uncertainties, it is becoming increasingly important for countries to ensure that a copy of the intellectual asset in which they have invested through time resides in their own jurisdiction. A consortia or national LOCKSS network helps to meet these needs.

The advantages include:

• Custody, ownership of content on national soil;

• Distributed national fault-tolerant preservation infrastructure;

• Community run and operated open source software;

• Greater efficiency and lower overall cost;

• Shared community service minimizes an individual libraries' barrier to entry and effort;

• Confirmation that libraries are receiving access to purchased content at the article level.

The LOCKSS collaborative approach addresses the economic, legal, technical and social challenges of building and preserving digital collections. Over the last 15 years it has proven to be practical and economically sustainable.

CONTENT PRESERVED

Many communities are running their own LOCKSS networks to preserve and ensure access to their collections.

Fee Based Journals & Books

Librarians invest in digital preservation to ensure post cancellation access, or perpetual access. The LOCKSS system restores to a community the ability to build and preserve local collections. It brings the traditional purchase-and-own model to fee based digital materials. When a community takes custody of content to which they subscribe, access is separated from payment and perpetual access assured. Materials stored in a local LOCKSS box remain available when the publisher goes away for any reason including publisher merger, bankruptcy, subscription cancellation, and network outage. LOCKSS provides 100% post cancellation access.

The Global LOCKSS Network is the largest, and oldest LOCKSS network preserving access to fee based journals and books. The Stanford University LOCKSS team releases content is to LOCKSS Alliance participants approximately twice per month (STANFORD…, 2013a).

Open Access Journals & Books

Many publishers, large (ELSEVIER 2013) and small are publishing open access materials. There are less-than-reputable open access publishers publishing content that do meet many research libraries' collection development policies (BEALL, 2013). There are also thousands of small open access publishers making available very important research and scholarship. These materials are particularly at risk. Most small publishers have vulnerable business models. A strong long-term strategy is for countries or regions to preserve the open access content produced in their own country.

Brazil is doing just that. The government agency, the Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT) in partnership with six Brazilian universities have implemented a six node LOCKSS network called, the Brazilian Network Services Digital Preservation - Cariniana

(Cariniana, 2013). The Cariniana project is preserving the more than 1000 Brazilian open access journals made available via SEER (SEER, 2013).

Government's & Information

It is easy to alter, tamper, remove, or otherwise censor without detection, online documents held

in centralized repositories. The reported number of documents being removed from the web is going up (GOOGLE, 2013) and occasionally document tampering is identified (FGI, 2013). The LOCKSS software is the only digital preservation approach designed to be tamper resistant. Government documents are vulnerable to tampering.

Librarians in Canada (CGI-PLN 2013) working with the Stanford LOCKSS Program and the Internet Archive have implemented a Canadian LOCKSS network to preserve their government documents. The content is initially harvested by Archive-It (Archive-It 2013) and then moved into the Canadian Government Documents LOCKSS Network for preservation.

United States librarians, working with the LOCKSS Program and the U.S. Government Printing Office have implemented a LOCKSS network that re-implements in the digital environment the Federal Depository Library Program (STANFORD…, 2013g; NEWS, 2010).

Locally Published Materials

Universities publish a wide variety of digital content. Some is born digital; some is digitized from analog media. A number communities are running LOCKSS network to preserve locally published content including electronic theses and dissertations, scanned newspapers, digitized photograph collections, audio video, business records and datasets.

Canadian colleagues have contributed code to the open source LOCKSS software to enable LOCKSS to interoperate with ContentDM and DSpace.

Examples of communities running LOCKSS networks to preserve locally published materials include:

• Alabama Digital Preservation Network: Alabama libraries are collaborating to preserve a wide variety of historic archival materials, including image collections and databases (ADPN, 2013).

• Council of Prairie and Pacific University Libraries: Canadian University libraries are collaborating to preserve collections important to the provinces of British Columbia, Alberta, Saskatchewan and Manitoba. This group focuses on preserving freely available born digital Web content including government documents, journals and small presses (COPPUL, 2013).

• MetaArchive Cooperative: Run by the non-profit Educopia Institute, this international membership organization coordinates cultural memory institutions that are collaborating to preserve very high value locally created digital materials (METAARCHIVE, 2013).

WORKFLOW

Implementing a LOCKSS network requires three things: for a publisher to give permission for the target content to be preserved; for a library to bring online a LOCKSS box that has authorized access to the content; and for that LOCKSS box to be registered with one of a number of associated LOCKSS Alliance networks (STANFORD…, 2013b).

Publishers grant the LOCKSS system legal permission to ingest, preserve and access their intellectual content by putting online a LOCKSS permission statement. This LOCKSS permission statement is bundled and preserved with the content. The content and the legal rights are preserved together. Paper contracts are lost over time, preserving the legal agreement with the content minimizes any future misunderstandings.

A LOCKSS network implements an extremely effective form of cooperative collection development. When one library or library group secures a publisher's agreement to participate in LOCKSS, all LOCKSS libraries with authoritative access to that content have permission to locally ingest and preserve the material.

A library uses the freely available, open source LOCKSS software to turn a mid-range PC, or virtual computing environment into a digital preservation appliance called a LOCKSS box (STANFORD…, 2013).

A LOCKSS box performs five main functions:

• It ingests content from target websites using a web crawler similar to those used by search engines;

• It preserves content by continually comparing the content it has collected with the same content collected by other LOCKSS boxes, and repairing any differences;

• It delivers authoritative content to readers by acting as a web proxy, cache or via Metadata resolvers when the publisher's website is not available;

• It provides management through a web interface that allows librarians to select new content for preservation, monitor the content being preserved and control access to the preserved content.

• It dynamically migrates content to new formats as needed for display.

LOCKSS Program staff at Stanford University analyzes the target content's URL structure, file formats and delivery mechanisms. They design, implement and update a tailored, content-specific preservation action plan that serves publishers, librarians and readers.

The publisher permits the LOCKSS system to collect, preserve and provide access to the content by putting a LOCKSS manifest page on the content's website (STANFORD…, 2013d). The manifest page contains a LOCKSS permission statement and links to the issues (or other parts) of the content as they are published. The required manifest page is ingested and preserved with the original content and negates the need for paper contracts.

Software called a LOCKSS plugin tells each institution's LOCKSS box where to find the publisher's LOCKSS manifest page, and how far to follow the chains of web links. A LOCKSS plugin encapsulates a publisher's content model by listing parameters specific to each publishing platform. The LOCKSS team builds, tests and distributes plugins to LOCKSS boxes registered with the LOCKSS Alliance.

Every LOCKSS box is located at an IP address that falls within its parent University's IP address range. Authorized LOCKSS boxes independently collect subscription or open access content directly from the publisher's website. The publisher authorizes or denies a LOCKSS box's access to content through their access control system. Publishers register LOCKSS activity on their web logs and have access to real time statistics through their own systems.

Once ingest is complete, the LOCKSS technology ensures that each LOCKSS box has collected all intended content, thus preserving the authoritative version. The LOCKSS software continually monitors the content in each LOCKSS box to ensure it is properly preserved though a cooperative preservation process that compares one LOCKSS box's content with the same content on other LOCKSS boxes (MANIATIS, 2003). When content is damaged or lost the system arranges for content repair from another LOCKSS box.

The administrator of each LOCKSS box can monitor the preservation status of the content in their box, by looking at delivered content and the management tools available through the LOCKSS box web administrative interface (STANFORD…, 2013e).

PROVIDING CONTINUOUS ACCESS

An institution's LOCKSS box can provide readers with continual, seamless access to branded publisher content (STANFORD…, 2013f). The LOCKSS system preserves content at its original URL, critically retaining the content's relationship to

other web resources. An institution's LOCKSS box delivers content to authorized readers when the publisher's website is unavailable including when a subscription is canceled, the network is busy or the publisher's server is down. The LOCKSS Program works to preserve and to deliver the publisher's original artifact to readers, in other words – LOCKSS preserves and delivers what the publisher published.

LOCKSS boxes provide four main ways for readers to access the content they preserve: by proxying (acting like a web cache), by serving (acting like a web server), by serving through integration with an OpenURL resolver, or via the Memento standard (RFC 7089).

• Proxying: Institutions often run web proxies to allow off-campus users to access subscription content. Libraries integrate their LOCKSS box into a proxy (PAC Files, EZ Proxy, ICP, Squid) to ensure a reader's URL request is seamlessly fulfilled when the content is unavailable from the publisher's website.

• Basic Serving: In the basic serving model, articles are accessed using a local URL pointing to the LOCKSS box. The LOCKSS box checks if the publisher will provide content to fulfill a reader's request. If the content is not available from the publisher, the LOCKSS box serves its own copy to the reader.

• OpenURL Serving: Libraries can integrate their LOCKSS box with their library catalog and OpenURL resolver by adding their LOCKSS box as a target to an OpenURL Resolver. In the OpenURL model content is accessed using bibliographic information.

• Memento: The Internet Engineering Task Force (IETF) has standardized Memento, a mechanism created by Michael Nelson and Herbert van de Sompel by which browsers can access preserved versions of websites. With funding from the Mellon Foundation, the LOCKSS team has implemented Memento so that LOCKSS boxes will conform to this standard to access past content.

When content has a persistent URL and a persistent URL service is available, a LOCKSS box configured with either proxying and/or momento can use that persistent URL to serve content.

Three audit and verification tools detail what content is in a library's LOCKSS box and the content's preservation status.

• On demand, a LOCKSS box produces a KBART (Knowledge Bases And Related Tools) report of the locally preserved content;

• A LOCKSS box displays detailed preservation status for each Archival Unit. (An Archival Unit is typically a volume of a journal, or a complete book);

• A LOCKSS box administrator can use a properly configured web browser from an authorized IP address to view preserved content through an "audit proxy". The viewer sees the content as it was collected by the LOCKSS system.

Librarians administer their institution's LOCKSS boxes through a web browser that allows them to easily select new content for preservation, monitor content's preservation status and a variety of other functions.

Post cancellation access to all preserved content is ensured as the content is under the library's local custody.

MIGRATING OBSOLETE FORMATS

LOCKSS preserves all web published formats (animations, datasets, moving images, still images, software, sound, text) and genres (journals, books, blogs, websites, scanned files, audio, video). The LOCKSS software is format-agnostic and preserves all content in its original format, as delivered from the publisher, including the format metadata that enables a browser to render the content.

The LOCKSS technology demonstrated its ability to handle format obsolescence (ROSENTHAL et al., 2005). When a browser requests content from a LOCKSS box, it uses the part of the HTTP standard called "content negotiation" to specify the formats it can render. If the requesting browser cannot render the format of the preserved content, the LOCKSS box invokes appropriate format migrators to create a temporary copy in a format the browser can render. After use, the LOCKSS box discards the temporary access copy.

The LOCKSS Program's "migration on access" approach has significant advantages over "format normalization" as it preserves the original artifact, uses much less overhead, saves money and takes advantage of the most up to date technology. Preserving the content in its original format satisfies archival requirements. It allows the LOCKSS system to be frugal with storage space. LOCKSS Program staff knows of no preservation system that discards the original bits after migrating them to a new format. Migrating and keeping both the original and the migrated copy multiplies the storage requirements for a preservation system by the number of migrations.

Preserved content that is migrated at the time the reader requests access on can use the most recent, and presumably best, technology available. Performing migration only when and if it is needed reduces the resource cost. Content can be migrated directly from the original to the current format, minimizing the effects of format conversion artifacts. The format converters, once developed, can themselves be preserved to document the original format.

SUMMARY

The Stanford University LOCKSS Program's open source distributed digital preservation approach builds tamper evident infrastructure to protect access to subscription, open access and special digital collections. Local infrastructure and collections ensure continual access. The more content is replicated, the greater its chances of surviving for the long term. The collaborative approach to addressing the economic, legal, technical and social challenges of building and preserving collections of digital content is practical and affordable. The LOCKSS Program enables communities to preserve important assets on their own soil.

Data de submissão: 23-09-2013 **Data de aceite: 23-03-2014**

REFERENCES

ADPN. Alabama digital preservation network. 2013. http://www.adpn.org/.

Archive-It. Web archiving services for libraries and archives. 2013. http://www.archive-it.org/.

BEALL, J. Scholarly open access. 2013. http://scholarlyoa.com/publishers/.

BLUERIBBONTASKFORCEONSUSTAINABLE DIGITAL PRESERVATION AND ACCESS. Sustainable economics for a digital planet: ensuring long-term access to digital information; final report. 2010. Disponível em: <http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf >.

Cariniana. Brazilian Network Services Digital Preservation. 2013. < http://www.ibict.br/pesquisa-desenvolvimento-tecnologico-e-inovacao/rede-brasileira-de-servicos-de-preservacao-digital>.

CGI-PLN 2013. Canadian Government Information Private LOCKSS Network. 2013. http://plnwiki.lockss.org/wiki/index.php/CGI_network.

COPPUL. Council of prairie and pacific university libraries. 2013. <http://coppullockssgroup.pbworks.com/w/page/11478105/FrontPage>.

ELSEVIER. Open access publishing. 2013. < http://www.elsevier.com/about/open-access/open-access-options>.

FGI. Free Government Information. Why we need digital deposit. 2013. <http://freegovinfo.info/node/3988>.

GOOGLE 2013. Google sees 'alarming' level of government censorship. <http://news.cnet.com/8301-1023_3-57454920-93/google-sees-alarming-level-of-government-censorship/>.

MANIATIS, P. et al. Preserving peer replicas by rate-limited sampled voting. 2013. (Best paper presented at the 19th ACM Symposium on Operating Systems Principles (SOSP), Bolton Landing, NY. October 2003). <http://berkeley.intel-research.net/maniatis/publications/SOSP2003.pdf>.

METAARCHIVE. 2013. <http://www.metaarchive.org>.

NEWS. Stanford University. Stanford helps to digitally preserve mountains of documents. 2010. <http://news.stanford.edu/news/2010/june/government-document-preservation-061510.html>.

OCKERBLOOM, J. Mediating Among Diverse Data Formats. 1998. Thesis (PhD) Carnegie Mellon Computer Science Department, Pittsburgh. <https://www.cs.cmu.edu/~spok/thesis.ps>.

RFC 7089. <http://www.ietf.org/rfc/rfc7089.txt>.

ROSENTHAL, D. S. H. Why preserve e-journals? Post-cancellation access. 2007a. <http://blog.dshr.org/2007/06/why-preserve-e-journals-post.html>.

_____. Why preserve e-journals? To preserve the record. 2007b.

<http://blog.dshr.org/2007/06/why-preserve-e-journals-to-preserve.html>.

_____. How well are we ensuring the longevity of digital documents? 2009. (Keynote paper presented at the CNI Spring Task Force Meeting, April 6-7, 2009, Minneapolis MN). <http://blog.dshr.org/2009/04/spring-cni-plenary-remix.html>.

_____. How Few Copies? 2011. (Paper presented at Screening the Future 2011, Netherlands Institute for Sound and Vision, March 14 -15, 2011). < http://blog.dshr.org/2011/03/how-few-copies.html>.

_____. Modeling the economics of long-term storage. 2012. (Paper presented at the Personal Digital Archiving Conference, Internet Archive, San Francisco). February 22-24, 2012. http://blog.dshr.org/2012/02/talk-at-pda2012.html.

ROSENTHAL, D. S. H., et al. Requirements for digital preservation systems: a bottom-up approach. D-Lib Magazine, 11:11. doi:10.1045/november2005-rosenthal.

SEER. Sistema Eletrônico de Editoração de Revistas. 2013. <http://seer.ibict.br/>.

STANFORD University LOCKSS Program. Build A LOCKSS Box. 2013. <http://www.lockss.org/support/build-a-lockss-box/>.

_____. Global LOCKSS network publishers and titles. 2013a. <http://www.lockss.org/community/publishers-titles-gln/>.

_____. How to join. 2013b. < http://www.lockss.org/join/>.

_____. LOCKSS source code: Sourceforge. 2013c. <http://sourceforge.net/projects/lockss/>.

_____. Prepare your content. 2013d <http://www.lockss.org/support/prepare-your-content/>.

_____. Use a LOCKSS box. 2013e. <http://www.lockss.org/support/use-a-lockss-box/>.

_____. View your preserved content. 2013f. <http://www.lockss.org/support/use-a-lockss-box/view-your-preserved-content/>.

_____. Digital Federal Depository Library Program. 2013g.

<http://www.lockss.org/community/networks/digital-federal-depository-library-program/>.