11th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology, 2018. ISSN 2336-5021. Available from: https://nusl.techlib.cz/en/conference/conference-proceedings

LOCKSS DISTRIBUTED DIGITAL PRESERVATION NETWORKS

Anthony Leroy

anthony.leroy@ulb.ac.be

Université libre de Bruxelles, Belgium

Abstract

As university libraries, preserving digital objects for future generations is one of our key missions.

This paper briefly discusses the essential features required for an ideal digital preservation solution to mitigate the many risks that endanger our digital assets.

The LOCKSS open source technology can help libraries build a robust distributed digital preservation network to ensure the very long-term availability of our scientific heritage. Many examples of existing implementations illustrate the wide variety of preservation networks currently based on the LOCKSS software.

Keywords

Digital preservation, LOCKSS, distributed preservation network

Introduction

In our digital era, the production of information grows at an unbridled rate. Access to the information has also become easier and faster than ever.

However, digital information also became much more fragile and vulnerable. A study from the University of British Columbia estimates that more than 80% of the research data at the origin of publications in zoology dating from the nineties is definitely lost (Vines, 2014).

The preservation of digital assets is a complex matter. Analog media such as paper or microfilms can be preserved for hundreds of years just by ensuring appropriate environmental storage conditions. By contrast, preserving digital objects in order to make sure that they will be reusable in the very long term requires elaborated strategies and rigorous processes to protect them against a large variety of threats. Most of these threats are not easily quantifiable and many catastrophic events have very low probability to occur but in the long term, some will inevitably happen.

Evaluating the risks

It is generally thought that the main risks concern hardware breakdowns, obsolescence or natural disasters and mitigation measures tend to focus mainly on those aspects.

However, in practice, it is observed that data losses find mostly their origin in human errors, external or internal computer attacks, financial or organizational problems (Rosenthal, 2005).

To mitigate these risks, the commonsense solution consists in making multiple copies. As a matter of fact, having more copies of the digital objects is the criteria which has the largest impact on the preservation solution reliability (Rosenthal, 2011)**.** Storing the copies on more reliable or more heterogeneous hardware will help but it will actually have a lower impact on overall reliability than having more copies.

Just how many copies are required depends on the threat model which includes the financial risk of managing too many expensive copies. Also, the law of diminishing returns states that over a certain number of copies, the cost of an extra copy is not paying off compared to the relative reliability improvement.

There is thus no single answer to this question but having many copies on consumer-grade media is much more reliable than having few copies on advanced expensive hardware.

Hence, the statement that « Lots of Copies Keep Stuff Safe » or LOCKSS (Maniatis, 2005).

Georeplication and Diversity for Better Preservation

Of course, having many copies is not enough. To protect them from small-scale disasters, it is necessary to disseminate them throughout the world, in places considered safe from natural and man-made hazards.

The management of each individual archive copy should also be left to parties which are autonomous and independent on the financial, administrative and organizational levels. It is also necessary to check the integrity of the data regularly and, if required, to migrate the files to long-term sustainability format.

No outsourced solution can guarantee data preservation according to these criteria. For reasons of economic profitability, commercial companies resort to the mutualization of technical and human resources and adopt the most profitable technology. The only guarantee offered to the customer of such third-party services is the existence of a contract stating that the supplier provides assurance to deploy economically “reasonable” efforts to preserve customer data and, possibly, in the event of an unfortunate loss, to grant a compensation.

The Role of University Libraries

The preservation of academic and scientific heritage is the responsibility of university libraries (Skinner, 2010). Digital objects of scientific interest that need to be shared and preserved are commonly stored in institutional repositories : theses, scientific publications and research data but also the grey literature for which the university is generally the only holder : internal reports, working papers, laboratory notebooks, white papers and research data.

Collaboration for an efficient preservation solution

A university library alone is unlikely to have at their disposal the human and technological resources needed to build an efficient preservation solution.

Collaboration between institutions to build a distributed preservation network is thus inevitable.

The organizational structure of a distributed preservation network can take multiple forms. The network can emerge from a pre-existing organization seeking for a preservation solution for their own locally created resources or it can be specifically created by a community for the preservation of objects of global interest.

Distributed preservation networks need to rely on a common technological infrastructure to support, at low cost and with limited human intervention, the coordination of the multiple archive copies which are stored on independent preservation nodes.

One of the most important services that a preservation infrastructure should provide is a mechanism to automatically and regularly check the integrity of the preserved copies.

The LOCKSS technology precisely offers one of the most sophisticated integrity monitoring services.

LOCKSS: a state-of-the-art technology

The LOCKSS technology is implemented as an open-source software originally developed for the global LOCKSS network to provide a solution for the post cancellation access or perpetual access to subscription e-journals and e-books.

It is an awarded technology : in 2014, it received the first ever perfect score when audited for the TRAC certification of the CLOCKSS archive (Rosenthal, 2014).

What makes LOCKSS software unique is precisely its robust integrity check and repair protocol enabling a secure audit mechanism between independently administered preservation nodes to test the integrity of the distributed copies. In the event of a corruption, the altered copy can be automatically repaired based on the valid copies (Maniatis, 2005).

LOCKSS mainly addresses bit-level preservation ensuring the preserved bits remain unchanged. The advanced logical-level preservation activities, such as format normalization ensuring that the bits will still be interpreted correctly, are left at LOCKSS network users' discretion.

LOCKSS: a vibrant community

The organizations using LOCKSS networks to preserve their digital content constitute a large and vibrant international community. Currently, counting a dozen networks and hundreds of institutions, the LOCKSS community is constantly growing.

While the Global LOCKSS network is probably the most famous LOCKSS preservation network, there are many other networks exploiting the same technology to preserve a large variety of content with various scopes, goals, governance and membership models (Reich, 2009).

Some examples are provided hereunder, classified by types of preserved content.

LOCKSS Networks Preserving General Interest Scientific Content

- The Global LOCKSS Network (GLN) is the first and largest LOCKSS network, counting more than 200 participating institutions all over the world. It preserves the content from over ten thousand e-journals from a wide variety of commercial and learned society publishers (Publishers & Titles, 2018).
- The Controlled LOCKSS (CLOCKSS) network is a dark archive jointly governed by academic publishers and university libraries to ensure the long-term survival of scholarly publications (CLOCKSS, 2018).
- *The Public Knowledge Project Preservation Network (PKP PN) is a dark archive preserving OJS journals. The PKP PN currently preserve more than 700 OJS journals and primarily targets the content not preserved elsewhere* (PKP Preservation Network, 2018).

LOCKSS Networks Preserving Government Information

- The Canadian Government Information (CGI) Digital Preservation Network, preserves digital collections of Canadian government documents (Wakaruk, 2013).
- The USDocs network preserves US digital government documents.

LOCKSS Networks Preserving Scientific Journals of Local Interest to a Community

- The Council of Prairie and Pacific University Libraries (COPPUL) Network preserves e-journals from member university libraries and small journals in Western Canada (COPPUL, 2015).
- The Cariniana Network preserves over 1000 Brazilian open access journals from Sistema Eletrônico de Editoração de Revistas (SEER) (A Cariniana e a Aliança LOCKSS da Stanford University, 2018).

LOCKSS Networks Preserving Content of Local Interest to a Community, including grey literature

- The Alabama Digital Preservation Network (ADPNet) preserves digital content locally created in Alabama. It proposes a low-cost digital preservation solution for academic institutions, state agencies, and cultural heritage organizations in Alabama (Alabama Digital Preservation Network, 2018).
- The WestVault Network provides a distributed digital preservation storage network spread across 4 western Canada provinces to preserve critical digital content submitted through an ownCloud instance (WestVault, 2018).
- The MetaArchive Distributed Digital Preservation Network is an international dark archive run by the Educopia institute to preserve high value locally created digital materials for more than 60 member libraries, archives, and museums (MetaArchive Cooperative, 2018).
- The SAFE LOCKSS Network is an international distributed archive preserving the born-digital open-access collections from a variety of institutional repositories managed by the participating member (Leroy, 2015).

The SAFE Network

The SAFE Archive Federation (SAFE) LOCKSS Network provides an interesting example of an organization solely built for the purpose of preserving content from the institutional repositories of participating institutions (including a good proportion of grey literature materials).

The network is a federation of completely independent institutions sharing a common view on how their own digital collections must be preserved. It is based on a light organizational structure around a simple memorandum of understanding. The budgets of participating institutions remain fully independent. Each member of the network simply agrees to make available a portion of their preservation node storage to keep copies of their partners' content.

SAFE is an international network ensuring an efficient replication of the archives in completely independent sites on the organizational level. SAFE has preservation nodes in Belgium, Canada, Germany, Sweden and Switzerland.

Each member keeps full technical control on their preservation node as only local administrators can manage the content of their network node.

Partners are however able to check the status of their preserved content over the network thanks to a monitoring tool collecting and aggregating status information from the nodes.

The LOCKSS technology is particularly well suited to enable this type of collaboration.

The future of LOCKSS

The original LOCKSS software was designed almost two decades ago at a time when collecting digital objects from static web pages was straightforward. It then gradually evolved to fit the needs of the increasingly dynamic web content, making the software more and more complex and difficult to maintain.

In 2017, LOCKSS was awarded a Mellon Foundation Grant to modernize the LOCKSS codebase by rearchitecting the software as a collection of Web Services with fully documented REST-APIs, a project named LOCKSS Architected As a Web Service (LAAWS) (Guicherd-Callin, 2018).

The new architecture relies on the state-of-the art open-source community software for the non-core-business modules (such as web crawling or content dissemination) and will align to the web archiving standard WARC. It will also support large-scale distributed storage to cope with the ever-increasing preservation storage needs.

In particular, the core LOCKSS component that provides the state-of-the-art data integrity monitoring service will be de-siloed into an independent web service. This will undoubtedly facilitate the reuse of LOCKSS in diverse contexts where advanced integrity check is needed; which will certainly result in a significant increase of the LOCKSS user base in the coming years.

Coincidently, the LOCKSS community is consolidating by creating users and developer groups to share best practices and develop community tools across networks.

Conclusion

The LOCKSS technology empowers university libraries to fulfill their essential mission of preserving digital knowledge by collaborating with other institutions to build a robust distributed preservation network. Many examples have been provided to illustrate the wide variety of existing networks currently employing the LOCKSS software.

The future of LOCKSS is exciting: the software is being re-architected to meet state-of-the art technology standards and the user community is looking forward to welcome and support the new preservation networks that will emerge to secure the access to unique knowledge for future generations.

Acknowledgment

We would like to acknowledge the great work of the LOCKSS team at Stanford University in developing and promoting this remarkable technology.

Of course, none of this would be possible without the groundbreaking work of Dr. David Rosenthal. His thoughtful ideas and clear vision greatly inspire the whole community. His blog, blog.dshr.org, is an invaluable source of information for anyone interested in digital preservation.

References

Alabama Digital Preservation Network, 2018. *Alabama Digital Preservation Network* [online]. Alabama Digital Preservation Network [Accessed 25 September 2018]. Available from: **http://www.adpn.org/**

CLOCKSS, 2018. *CLOCKSS* [online]. [Accessed 24 September 2018]. Available from: **https://clockss.org/**

COPPUL, 2015. COPPUL Private LOCKSS Network Governance Policy. In: *Council of Prairie and Pacific University Libraries (COPPUL)* [online]. COPPUL, 2015 [Accessed 25 September 2018]. Available from: **https://coppul.ca/pln-governance**

COPPUL, 2018. WestVault. In: *Council of Prairie and Pacific University Libraries (COPPUL)* [online]. COPPUL, 2018 [Accessed 25 September 2018]. Available from: **https://coppul.ca/westvault**

GUICHERD-CALLIN, Thib, 2018. LOCKSS Software Re-Architecture. In: *34th International Conference on Massive Storage Systems and Technology* [online]. Santa Clara [Accessed 25 September 2018]. Available from: **http://storageconference.us/2018/Presentations/LOCKSS-tutorial-4.pdf**

LEROY, Anthony and Patrick HOCHSTENBACH, 2015. SAFE PLN: An International Preservation and Access Solution. In: *D-Lib Magazine: In Brief* [online]. July/August 2015 [Accessed 25 September 2018]. Available from: **http://www.dlib.org/dlib/july15/07inbrief.html**

MANIATIS, Petros, Mema ROUSSOPOULOS, T. J. GIULI, David S. H. ROSENTHAL and Mary BAKER, 2005. The LOCKSS peer-to-peer digital preservation system. *ACM Transactions on Computer Systems* [online]. **23**(1), 2-50 [Accessed 24 September 2018]. DOI: 10.1145/1047915.1047917. ISSN 0734-2071. Available from: **http://portal.acm.org/citation.cfm?doid=1047915.1047917**

METAARCHIVE COOPERATIVE, 2018. MetaArchive Resources. In: *MetaArchive* [online]. Atlanta: Educopia Institute [Accessed 25 September 2018]. Available from: **https://metaarchive.org/documentation-resources/**

PKP Preservation Network. In: *Public Knowledge Project* [online]. Simon Fraser University Library, 2014 [Accessed 24 September 2018]. Available from: **https://pkp.sfu.ca/pkp-pn/**

A Cariniana e a Aliança LOCKSS da Stanford University. In: *Portal da Rede Cariniana* [online]. 2018 [Accessed 25 September 2018]. Available from: **http://cariniana.ibict.br/index.php/noticias/377-a-cariniana-e-a-alianca-lockss-da-stanford-university**

Publishers & Titles (GLN). In: *Lots Of Copies Keep Stuff Safe* [online]. Stanford: Stanford University, 2018 [Accessed 24 September 2018]. Available from: **https://www.lockss.org/community/publishers-titles-gln/**

REICH, Victoria and David ROSENTHAL, 2009. Distributed Digital Preservation: Private LOCKSS Networks as Business, Social, and Technical Frameworks. *Library Trends*. **57**(3), 461-475 [Accessed 24 September 2018]. DOI: 10.1353/lib.0.0047. ISSN 1559-0682.

ROSENTHAL, David S. H., Thomas ROBERTSON, Tom LIPKIS, Vicky REICH and Seth MORABITO, 2005. Requirements for Digital Preservation Systems. *D-Lib Magazine* [online]. **11**(11) [Accessed 24 September 2018]. DOI: 10.1045/november2005-rosenthal. ISSN 1082-9873. Available from: **http://www.dlib.org/dlib/november05/rosenthal/11rosenthal.html**

ROSENTHAL, David, 2011. How Few Copies? In: *DHSR's blog* [online]. 2011-03-15 [Accessed 24 September 2018]. Available from: **https://blog.dshr.org/2011/03/how-few-copies.html**

ROSENTHAL, David, 2014. TRAC Certification of the CLOCKSS Archive. In: *DHSR's blog* [online]. 2014-07-24 [Accessed 24 September 2018]. Available from: **https://blog.dshr.org/2014/07/trac-certification-of-clockss-archive.html**

SKINNER, Katherine, Matt SCHULTZ and METAARCHIVE COOPERATIVE (U.S.), 2010. *A guide to distributed digital preservation*. Atlanta, Ga.: Educopia Institute. ISBN 978-0-9826653-0-5.

VINES, Timothy H., Arianne Y.K. ALBERT, Rose L. ANDREW, et al., 2014. The Availability of Research Data Declines Rapidly with Article Age. *Current Biology* [online]. **24**(1), 94-97 [Accessed 24 September 2018]. DOI: 10.1016/j.cub.2013.11.014. ISSN 09609822. Available from: **https://linkinghub.elsevier.com/retrieve/pii/S0960982213014000**

WAKARUK, Amanda, 2013. Introducing the CGI-PLN: Using the LOCKSS Program to Preserve DSP Content in a Changing Environment. In: *Government Information Day* [online]. Toronto [Accessed 24 September 2018]. Available from: **https://era.library.ualberta.ca/items/33a42ce9-a1d8-42b9-ba2c-3b8b0a776dbe**