Digital preservation of scholarly content, focusing on the example of the CLOCKSS Archive

Based on a breakout session to be held at the 38th UKSG Annual Conference, Glasgow, March–April 2015

In this article the author explores the development of digital preservation, including consideration of what it is and what it is not, and looks at the challenges that preservation of multiple formats of digital scholarship brings. After setting definitions and considering the different types of archive currently available, the mission and principles of the CLOCKSS Archive are explained. The wrap-up covers the changing nature of scholarly communication, a few of the challenges and the approach to these taken by the CLOCKSS Archive.

Introduction

RANDY S KIEFER

Executive Director
CLOCKSS Archive

Preservation of digital content can be defined in a number of ways. Long-term preservation refers to the processes and the procedures required to ensure content remains accessible well into the future. A Jisc-commissioned report published in 2008[1] provided distinct and helpful definitions regarding digital preservation. *Continuing* or *perpetual access* is an attempt to replicate the situation with paper journals where libraries receive, make available and preserve the material for ongoing reference, regardless of whether or not the subscription is continued. *Long-term preservation*, on the other hand, can be viewed as an issue not just for the subscribing library, but for society as a whole, ensuring that the scholarly record continues to be accessible to future generations.

Why is there a need for the preservation of digital content? Globally, the market demand by libraries is that they want to be assured that there is an independent third-party preservation of electronic content. Centrally managed preservation of collections, preserved on national soil for safekeeping, provides security for the regional community. Publishers also want to be good stewards of their content and respond to the library market.

Preservationists become the keepers of content, and then the disseminators of the content in the cases where trigger events occur. Common trigger events can include the demise of the publisher, usually due to bankruptcy where there is no pick-up of their assets; discontinuation of a journal where a publisher removes all internet access; or, a disaster disrupting the publisher's availability for an extended period of time. These are the key concerns.

In the event of such triggers, the value of preservation for libraries is that of an insurance policy for all of the resources. Preservation provides all libraries with access to archived content when it becomes lost, orphaned, or abandoned.

'the value of preservation for libraries is that of an insurance policy'

At this point, it would be useful to identify systems or operations that do not constitute genuine preservation. Commercial hosting is not preservation. This includes aggregation databases, journal-hosting platforms and distribution platforms for e-books. These are not preservation modes, and they are not archives. Commercial hosting that publishers have with a number of entities,

and the relationship with its rights to content, end when the publisher no longer pays for the service. Aggregators are not preservation archivists. They arrange distribution (secondary publishing) of content on behalf of the publisher, with a royalty income element. The access is limited to subscribers, and they may or may not continue when the content and/or publisher is no longer available. Some examples include Gale, EBSCO and ProQuest. Other aggregation platforms include Project Muse, Project Euclid and other specialized topical collections based on academic discipline. Again, should a publisher drop out of them, they have the right to remove the content, or it might be contractually required. They are not in the business of long-term preservation.

Digital preservation archives

Generally, there are two types of digital preservation archives. The first type is the global archive, like the CLOCKSS Archive, the Global LOCKSS Network and Portico. The second type is the regional archive, including the British Library, the National Library of the Netherlands (Koninklijke Bibliotheek: KB) and the Swiss National Library. There are many national libraries run by the central government of a country that are actively involved in preservation – some of the efforts cover digital content and many are actively working on printed content.

Preservation by CLOCKSS

The example of a global archive, CLOCKSS (Controlled LOCKSS), will be examined in some detail. CLOCKSS was founded in 2006 as a project with a number of research libraries and global publishers considering how they could leverage the open source software LOCKSS (Lots Of Copies Keep Stuff Safe) into a dark archive. These participants developed a structure including the rules, regulations and processes needed to create a dark archive. The underlying concept of the dark archive choice is to protect the digital content from any degradation that can occur when there is constant access to the content. In addition, the distribution of the content across the globe adds another layer of protection and involvement beyond one region.

> 'the dark archive … is to protect the digital content from any degradation'

The mission of CLOCKSS, a not-for-profit joint venture between the world's leading academic publishers and research libraries, is to build a sustainable, geographically distributed dark archive with which to ensure the long-term survival of web-based scholarly publications for the benefit of the greater global research community. So, it is intended to be for the benefit of the entire world! Content that is no longer available from any publisher (the 'triggered content' already described) is made accessible free of charge. CLOCKSS assigns this abandoned and orphaned content a Creative Commons license to ensure that it remains available forever.

The founding principles of the CLOCKSS Archive, which continue today, are:

- community governance
- global approach – decentralized preservation
- proven technology using the open source software LOCKSS
- a commitment to open access (OA) in the long term.

Community governance

Community governance is a shared responsibility across the academic community. Designated libraries host the CLOCKSS archive node servers. These libraries also are represented on the CLOCKSS Board of Directors, matched by an equal number of publishers. So, there is a 50-50 split between libraries and publishers on the Board, each with one vote, and it has very international representation.[2] There is currently one publisher vacancy on the Board.

Decentralized preservation

CLOCKSS looks to academic libraries to provide stewardship and preservation in accordance with the principle of decentralized preservation. In this way, libraries are reinforcing their established social value as memory organizations by being engaged in preservation. They are also ensuring against social and geophysical risks because they are spread around the world in regions that cover Australia, Hong Kong, Japan, Canada, the US, Scotland, Italy and Germany. As well as distributing the risk geographically, the CLOCKSS Archive has actively involved universities from different political and social backgrounds.

'CLOCKSS looks to academic libraries to provide stewardship and preservation'

Open source technology

The deployment of proven, open source technology by CLOCKSS is based on the use of the LOCKSS software, which has been securely and safely preserving web-published content for well over 15 years. The LOCKSS software continues to evolve to handle web advances to preserve new content types and the LOCKSS technology continues to be adapted for the dark archive functionality used by CLOCKSS. One of the key features is the strictly limited access control, which eliminates any ability of outsiders to access the content on the archive nodes. The existing 12 preservation servers talk to each other, and only to each other, to compare and update any content. Additional functionality has been added to the LOCKSS software to enable a copy to be extracted at the time content is triggered. In the broadest sense, the CLOCKSS Archive is a Private LOCKSS Network (PLN), and there are several other PLNs preserving different types of content.

Commitment to open access

The CLOCKSS Archive is committed to open access. While we are a dark archive that permits no access at the time of the trigger event due to the non-availability of archived content, the CLOCKSS Archive Board will intervene with a vote to trigger the content, which may be set in motion by a request from the community. With the Board's approval, the content is made available OA to those who want to host the content under a Creative Commons license. There are two main hosts: Stanford University and the University of Edinburgh, though anyone may host the triggered content.

'The CLOCKSS Archive is committed to open access'

The CLOCKSS community

The CLOCKSS Archive's designated community can be identified in three parts. The first includes the scholars, students and readers of electronic academic content (the end users). The second is made up of the libraries that purchase and manage this content on behalf of those end users. And finally, of course, there are the publishers of the content.

CLOCKSS provides indirect, independent services to the community served by its Archive. Essentially, this is content insurance for the libraries. They know if content is in our system and they are subscribing to it, when the content is no longer available for those reasons previously noted, CLOCKSS will be able to provide it. Libraries have peace of mind about the content from participating publishers, because publishers have provided the content for preservation to the CLOCKSS Archive.

In February 2011, the author was hired as the first full-time Executive Director. By 2014, the CLOCKSS Archive had nearly 200 participating publishers in 29 countries, including Egypt, Greece and Romania. As of today (the beginning of 2015), it has 722 supporting libraries in 42 countries, and the CLOCKSS network is expanding. In January 2015, Brazil was in the process of completing its application process to become the 13th node, and the Board looks forward to welcoming Brazil to its network. As the CLOCKSS Board has authorized 15 nodes in total, there are opportunities for growth into other regions such as Africa, other parts of Asia and Europe.

CLOCKSS was audited by the Center for Research Libraries (CRL) in the US using the Trusted Repository Audit Checklist (TRAC). The audit was performed from September 2013 through June 2014 and included an on-site visit and audit of the system. The CLOCKSS Archive was certified as a 'Trusted Digital Repository', scoring 13 out of a possible 15, tied on performance with one other archive. It scored a perfect five in the Technology and Security category – the only archive to score perfectly in that category. This certification is a very important accomplishment for the Archive and involved a strong commitment of resources from a small team to prepare all the materials for CRL to review.

Challenges ahead

Going forward, there are many issues to be resolved regarding preservation of electronic content. Three basic challenges to content preservation can be pinpointed, each one of them intertwined with the others. With 12 archived nodes located around the world, replicating the content 12 times, the CLOCKSS Archive has to be very selective in what it captures for preservation – primarily journal content and e-books. This precious commodity of space requires ongoing rigorous review of its collection development policy. In recent months the Board of Directors of CLOCKSS has begun a new review of those policy guidelines. The issue most related to this is the publication of databases.

Database content preservation

Academic content in databases is a growing area, some of which is not peer reviewed. Among the various items that are being collected in databases are data sets, figures, single entries and ongoing revisable entries. Databases can be broken down into three categories. The first is a closed database, where all the individual records are already contained in the database table. This type is relatively straightforward to preserve because the content is fixed and does not change over time. The second type of database is open, where entries can be made continuously through time. One type of open database is the appended-style database where every new record is added to the end of the database. The data content can be separated into time intervals, and preserved in distinct blocks of data based on a date interval, for example. The other type of open database is the continually updated database. This particular type of database is nearly impossible to preserve and one can only gather a snapshot at a given point in time. To the best of this author's knowledge, there is currently no known way to capture a continuously updated database.

'Academic content in databases is a growing area'

A further challenge with databases is their status among academic libraries regarding their worthiness to be preserved. Large data sets and data files also present a size problem, not only in collecting the data, but in distributing the data. At the moment, there is a working group on the Board of the CLOCKSS Archive developing new guidelines for databases. When CLOCKSS began, its focus was on journals and then expanded to e-books. The challenge of databases is one not only for CLOCKSS, but for every preservation organization.

'The challenge of databases is one … for every preservation organization.'

Changing standards and formats

Another challenge to digital preservation is the problem of ever-changing internet standards and formats. New programming languages like Ajax and HTML5 have to be addressed as content moves in that direction. Each of those formatting tools allows for much more dynamic interactive content. Once again, the challenges are in the changing nature of the content. In many cases, the actual content of articles and chapters has not really changed, but the issues are with screen interactions introduced to provide additional context or direction. These tools just allow you to access supplemental or related materials. The LOCKSS team at Stanford University, which supports the LOCKSS software, is working under a grant from The Andrew W Mellon Foundation to develop adaptive tools for the LOCKSS harvesting process. Since Stanford University and the CLOCKSS Archive are both

non-profit in nature, they continue to rely on outside funding grants in different initiatives to help keep the software up to date. The LOCKSS software is an open source software, and there are a number of developers at various PLNs working on additions and changes to the basic software to solve different problems of parsing, organizing and retrieving content. This model of adapting the software has worked for well over 15 years, and we fully expect it to continue with our various partners around the globe.

Collection development policy

One of the organizational challenges for CLOCKSS is to come up with a method to develop a consensus of what content needs preservation. The advent of the OA publisher and the continued growth of commercial publishers pose significant challenges. Commercial content is on an exponential path and has been for more than two decades. From the CLOCKSS point of view, the support we receive from libraries helps us to develop and maintain the archived node network and the managing infrastructure. The fees that we receive from publishers, both annual fees and transaction fees (ingest fees), are applied to the technical services needed to preserve content. We consider the growth of supporting libraries as key to our funding and to the growth of content under preservation. As our Board works on developing updated guidelines for the selection of publishers, we face a balancing act. In most models of services provided to the community, the funding is based on fee tiers, where the largest pay the most and the smallest pay the least. In short, the largest publishers subsidize the work it takes to preserve a small publisher. The work for either size publisher is essentially the same. Although the time spent gathering the content will vary from small to large, the inherent cost is in the initial set-up, and this effort is generally equal.

'One of the … challenges for CLOCKSS is … a method to develop a consensus of what content needs preservation'

Conclusion

Digital preservation of content will continue to be challenged by the rapid growth of content, the quality of the content being published, the nature of the content (such as databases, figures, etc.) and by the changing nature of programming that affects the display and storage of the content. The CLOCKSS Archive continues to evolve in many different ways to meet those challenges. We always look forward to any and all input from the library community regarding our future direction.

Competing interests: The author has declared his affiliation as Executive Director, CLOCKSS Archive

References and notes

1. Morrow, T, Electronic journals – continuing access and long-term preservation: roles, responsibilities and emerging solutions. Breakout session presentation, UKSG conference 2009: http://www.uksg.org/event/conference09/breakout_sessions/ (accessed 17 February 2015).

2. The current (2015) CLOCKSS Board of Directors is made up as follows:

   The libraries are represented by: Roxanne Missingham (Australian National University), Peter Schirmbacher (Humboldt University), Carolyn Walters (Indiana University), Jun Adachi (National Institute of Informatics), Chip Nilges (OCLC), Kerry Keck (Rice University), Michael Keller (Stanford University), Ellis Sada (Università Cattolica del Sacro Cuore), Geoff Harder (University of Alberta), Peter Burnhill (University of Edinburgh), Peter Sidorko (University of Hong Kong) and Carla Lee (University of Virginia).

   The publishers are represented by: Vida Damijonaitis (American Medical Association), Rita Scheman (The American Physiological Society), Alicia Wise (Elsevier), Graham McCann (IOP Publishing), John Carroll (Nature Publishing Group), Mark Heaver (Oxford University Press), Carol Richman (SAGE Publications), David K Marshall (Society for Industrial & Applied Mathematics), Wim van der Stelt (Springer), Ed Cilurso (Taylor & Francis) and Craig Van Dyck (John Wiley & Sons). There is currently one publisher vacancy on the Board.

Randy S Kiefer
Executive Director
CLOCKSS Archive
E-mail: randy.kiefer@clockss.net

ORCiD: http://orcid.org//0000-0002-2298-2418

To cite this article:

Kiefer, R S, Digital preservation of scholarly content, focusing on the example of the CLOCKSS Archive, *Insights*, 2015, 28(1), 91-96; DOI: http://dx.doi.org/10.1629/uksg.215

Published by UKSG and Ubiquity Press on 5 March 2015