

UK LOCKSS Workshop

Monday 24 November 2003

Workshop Report

Brief summary of the LOCKSS system

The LOCKSS (Lots of Copies Keeps Stuff Safe) initiative was developed in Stanford and is an alliance of publishers and libraries to provide low-cost caches of content licensed by libraries from publishers, so that they effectively “own” the material they have purchased, in much the same way as they did in the print environment, and can assert the traditional library role of custodianship of scholarly digital materials. Because material is stored at several sites, the risk of loss of content in the event of disaster will be minimised as there will be another copy stored elsewhere.

Background to the UK LOCKSS Workshop

The Workshop was convened by JISC to take the opportunity of a visit to the U.K by Vicky Reich and David Rosenthal to explore the technical development of LOCKSS, its move to a new funding model, and the potential of LOCKSS for preserving content collected in the UK. The LOCKSS system is suitable for all file formats delivered via http protocol. This includes licensed e-journals, which have been the subject of a recent JISC Study looking at implementation of archiving clauses in the NESLI Model Licence, as well as other genres of digital content. A recommendation in this report, which JISC intends to proceed with, is for a technical evaluation of the LOCKSS system. Another JISC funded study is of e-print archives and it was felt that LOCKSS may hold potential in the context of institutional repositories as well as licensed e-journals.

Presentation on LOCKSS

The workshop began with a presentation by Vicky Reich and David Rosenthal to provide background and context for proposed further development of the LOCKSS system.

It was explained that in addition to e-journals licensed from large commercial publishers, the low cost of LOCKSS also makes it particularly attractive for journals produced by smaller publishers who lack the resources to protect the material they are creating. In addition to e-journals, other genres which could potentially benefit from the LOCKSS system includes government documents, electronic theses and dissertations, newspapers, medical grey literature, and snapshots web content. The system works well for material delivered through http and that has, or can be made to have an authoritative version.

There are two levels of technical expertise required, the “easy” layer required to maintain the LOCKSS system at participating institutions, and the more highly specialised programming expertise which currently resides at Stanford but which would be beneficial to develop outside Stanford and the U.S. This level of expertise is not required by everyone but it would be valuable for all LOCKSS partners to have

access to shared technical expertise in the U.K, as well as the U.S. This expertise is required to test and develop new models for the system as it evolves. It was suggested that the new Digital Curation Centre potentially could act in this role for the UK.

The LOCKSS system has received extensive beta testing over the past two years, and has involved more than 50 publishers and more than 80 libraries. U.K libraries involved in LOCKSS beta testing are the British Library and the universities of Cambridge, Edinburgh, Glasgow, Imperial College, and Leeds.

LOCKSS caches collect material from the publisher's web site as it is published, via a targeted crawl. Access to the content in LOCKSS caches is provided by configuring them as proxy servers. The LOCKSS architecture involves three layers, the platform, the preservation layer (where the proxy cache resides), and the plug-in.

A high level of trust has been developed between the LOCKSS system and publishers, one of the reasons why participants can only begin to collect material as it is published, not retrospectively. It is also part of the agreement with publishers that libraries can only provide their caches to external sources for purposes of audit or repair, not for access which has not been paid for.

Storage costs are declining and similarly, the storage costs for LOCKSS caches can also be expected to decline over time. In 2003 costs, one journals year of storage was calculated at \$0.70 (assuming a generous estimate of 0.5GB per year as the median e-journals size).

Future Development

After two years of beta testing, a production quality system is now being built and is being rigorously tested using a java code tool.

Further development work was required for the user interface, which has now improved but still needs further work on. It is planned that the user interface will show the status of content. An alert management system has been designed to notify participants if there is a problem requiring human intervention. Also under development is what has been termed a "neighbourhood watch" system so that participants can keep an eye on each others caches in case one server goes down.

Requests have also been received for LOCKSS to interact with central repositories, and negotiations have been held with the California Digital Library, who subscribe to many e-journals and so prefer to have a small number of large machines capable of storing these rather than a large number of small machines.

It was suggested that there is currently insufficient LOCKSS activity within the U.K for it to be sufficiently robust so there is a need to increase the number of UK LOCKSS participants.

Funding has been predominantly from grant aid to date but a new business model will need to be developed which relies less heavily on grant funding and moves to a fee from publishers and libraries which will allow further development and testing of the

software, training others to use the technology, and growing the collection development activity.

In terms of format migration, the LOCKSS system was considered to be long-lasting because browsers will be expected to be able to support both new and old formats for a long time. Evolution of formats on the web tends to be very limited (e.g. moving from html to xml). LOCKSS is capable of either performing bulk format conversion or on the fly, as material is requested. Further testing of a “migration on request” approach is planned for next year. It would be much more complicated to replace the transport protocol but this would take a very long time to become an issue as the web is far too big to coordinate instant upgrades.

In summary, the LOCKSS system permits libraries to own the content they are currently leasing from publishers, employing inexpensive hardware and open source software with low system administration costs. It offers a pragmatic and readily available solution to a daunting challenge. Several digital approaches are needed but LOCKSS can be used to save content which would otherwise disappear while waiting for an optimum solution, and is low cost. Participation in the LOCKSS Alliance is optional but will ensure that the system will continue to develop.