**World Library and Information Congress:
69th IFLA General Conference and Council**

1-9 August 2003, Berlin

Code Number: **188-E**
Meeting: **173. National Libraries - Workshop**
Simultaneous Interpretation: **-**

Cultural custodianship in the information age: LOCKSS and other digital repositories

Michael A. Keller
Stanford University, Stanford, USA

National Libraries, traditionally charged with developing, maintaining, and providing access to the collections of record of their national imprints, are now facing the challenge of performing the same functions for digital publications. These national collections of record in many cases have arisen operationally from the receipts of books and other publications desiring national copyright protection. Now that a very large portion of the most prolific publishing nations' output is born and distributed digitally, new operations and mechanisms are necessary to collect, maintain, and provide access to the digital collections of record. However necessary the new operations and mechanisms are, they well may be unfunded mandates, requirements set either by national legislatures or by the missions of the national libraries, but not provided with sufficient funds to achieve them. The politics of unfunded mandates are very well known to each national librarian and certainly to me, as a citizen of the United States as much as a senior officer of Stanford University, but my talk will not dwell on that aspect of the topic, as much for selfish reasons of career stability as for the press of time.

At some level all libraries, certainly most university libraries, and absolutely all national libraries bear responsibility for cultural custodianship, the collection and preservation of artifacts of communication from our own and previous generations to those yet to be born. And along with the collection and preservation of those artifacts comes providing access to them along with interpreting and distributing them or their surrogates.

Before the decade of the 1990s, we focused upon written and printed artifacts, from cuneiform tablets recording business transactions in Mesopotamia 5,000 years ago to fanzines feeding the latest fads five minutes ago. Now we must focus on the digital artifacts of communication, e-journals and e-documents of all formats and genres. And we should regard seriously the failure of the industrialized societies to save the earliest digital artifacts, because we have already lost digital information resources that citizens and public policy makers would find valuable in understanding the course, and the curse, of human exploitation of the planet. I refer to the NASA Landsat photos of the Amazon River basin from the mid-1970s that would help us compare the size and effect of the Amazon rain forest on the global climate then and now. That the digital artifacts are seemingly ephemeral in their physical manifestation, pulses of energy in magnetic, optical, and network environments, does not excuse us from capturing those pulses and protecting them so that the true, authentic versions can be examined and re-examined over time – long, long periods of time. Just as we can now examine, consider, re-consider, and interpret reiteratively documents from the Middle Ages, so should our successors in 3003 and even 10,003 have the same opportunities for contemporary records.

Why should we worry about the long term survival and use of digital publications? There are several reasons. One is that increasingly, the version of record is the digital one. Another is that, in many cases, there is only a digital version. Third, digital publications can be changed rapidly and without signal; knowing the original version could be crucial. Fourth, digital publications often have adjuncts, functional ones, like virtual models of molecular processing, or contextual ones like hyperlinks to cited literature. These characteristics and reasons are so different from those of traditional publishing, and digital publishing is becoming so rapidly the norm for scholarship, that to ignore the need for digital repositories is to perpetrate the creation of a digital dark age. Our society - mankind in all of our nations - needs the record of digital publication for our future.

This contribution to the IFLA Workshop on National Libraries – entitled “National libraries as access points: virtual libraries for virtual users” is about one mechanism, LOCKSS, that offers considerable promise for digital preservation, now applicable to a single genre, that of e-journal articles, and soon to another genre, that of government documents. Now in user testing stages, LOCKSS’ architecture will support the gathering, maintaining, and delivery of documents in other genres of publication beyond e-journal articles and e-government documents - but more about that later. What does that ridiculous acronym derive from? Lots of Copies Keep Stuff Safe.

This talk will cover the brief history of LOCKSS, a description of its functions, some notions of how it might be applied in library settings, including that of a of a national library, the present stage of development of LOCKSS and where it might be going.

By mid-1999, Stanford’s HighWire Press had grown to about 170 journals and had begun to hear from its publisher clients, as well as directly from institutional subscribers, of their need for a reliable, auditable, e-journal archiving system that would result in files residing in local institutional subscriber repositories. HighWire Press itself, of course, maintains its own digital archive, both on and off-line, but that solution and assurances about it were not fully satisfactory to librarians. Librarians wanted then, and want now, to exert custodial control over the intellectual assets to which they subscribe or which they buy, whether in printed or in digital form.

So, as the question of e-archives and digital repositories became more prevalent in our professional circles, Vicky Reich, then an assistant director of HighWire Press responsible for library relations and marketing support, and David Rosenthal, a senior systems engineer in the research lab at Sun Microsystems, took up the problem of e-journal archiving. Reich and Rosenthal conceived of a method by which e-journal source files could be gently and slowly visited by a special spider to retrieve e-journal articles to which the institution had subscribed and then those collected articles placed in a local digital repository or persistent cache. The spider would work only with the permission of the publisher of the e-journal, but would establish full sets of subscribed e-articles in the local caches. These caches and the slow, gentle spiders would work together, using a polling mechanism, to maintain the integrity of the institutional cached sets when the network went down or when the source files ceased to be available for whatever reason. In the event of the source files not being available, the local caches would become visible to the local readers and simultaneously, the many LOCKSS caches at different institutions would interact to assure consistency among the caches of any given e-journal. In order to protect the institutional caches from malicious or ill-conceived raids, there would be no central list of caches. Thus, no hacker, no government could decide to clear all of the LOCKSS caches of their contents. Another principle held by Stanford and the LOCKSS development team was and still is that of the software being freely available via open source licensing. By design, the software is easily operable on inexpensive personal computers. Also, the LOCKSS design assures publishers that their journals' editorial values and brands will be available only to authorized and authenticated subscribers; this is a critical matter for achieving acceptance of the system.

Publishers would like to use LOCKSS to audit distribution of their e-journals to assure themselves that only subscribers are developing local caches of their titles.

The LOCKSS development effort was originally supported by Stanford, by Sun Microsystems, and by the Andrew W. Mellon Foundation. Recently, funding has been provided by the National Science Foundation of the United States government. HighWire Press and LOCKSS both continue to be parts of the Stanford University Libraries and Academic Information Resources organization.

LOCKSS in its early incarnations, was tested by about 50 libraries s around the world. After the initial testing, the LOCKSS team decided to re-design the software to facilitate extensibility. The version of LOCKSS out for testing now is written so that “plug-ins” for any genre distributable via http could be prepared and used using the underlying LOCKSS demons and core software. Versions of the system are released as open source software on Source Forge once they are fully tested and debugged.

What does LOCKSS do? How would LOCKSS installations around the world help accomplish the tasks of digital repositories?

LOCKSS can collect any of the file formats distributed by http, hypertext transfer protocol. Among those formats are html, PDF, jpeg, mpeg, TIFF, and so forth. LOCKSS collects distributed versions of files, not underlying source files. Let me use the example of articles purveyed through HighWire Press. Sgml and xml are the underlying source code formats of HighWire articles; these are converted to html when called up by a reader. LOCKSS acts as a reader, so the LOCKSS caches would gather html from HighWire Press. In order to do this, the LOCKSS spiders must have authorization from the publishers. In other words, just as readers from any particular institution are authorized to gain access to content on the publisher’s servers, LOCKSS spiders act as though they are readers from that institution. In addition, publishers need to agree that local institutions may hold their content in LOCKSS caches for local use, including as local archives. LOCKSS spiders work relentlessly, but slowly, to gather all newly released content, including updated articles or other content previously gathered.

As you can see, we have adopted the image of a turtle to emphasize the slow, gentle nature of LOCKSS and the indestructible nature of the LOCKSS caches.

Once a local LOCKSS cache has been established, it constantly checks its holdings against many other, but not all other, LOCKSS caches of the same content. Should the source files disappear from the network, this process of constant comparison then results in a kind of polling so that each cache can be assured its content is uncorrupted.

At the same time, in the event of the publisher's or source files' disappearance from the network, the local LOCKSS cache provides local access to the local readership. In a sense, the LOCKSS caches constantly verify their completeness and correctness against the other caches.

And they repair gaps in one another's holdings automatically. Finally, the LOCKSS caches assure that the content is never unavailable due either to network problems or to problems at the source of the files.

Here is a diagram of the flow of information in a network of LOCKSS caches.

It is important to note that LOCKSS software works well on inexpensive personal computers, but requires large amounts of low-cost digital storage, commensurate with the amount of information being cached.

The software has been designed so that loading it and activating it takes no more technical expertise than loading a new application to one's personal computer.

How might LOCKSS assist national libraries in accomplishing their mission of collecting, maintaining, and providing access to the publications emanating from within their national boundaries?

In the most basic model, a national library would acquire the LOCKSS open source software, load and activate it on appropriate hardware with network connections, and then let LOCKSS do its work on national publishers' websites. That LOCKSS site would automatically interact with many other LOCKSS sites containing the same content to assure completeness and timely collection. Attention would have to be paid over time to the migration of contents of the cache as it grew to the maximum capacity of the local system supporting it. One can imagine eventual migration paths, e.g., to off-line tape or large-scale magnetic digital repositories. However as one can see from the previous chart, the cost of PC memory is relatively inexpensive and, by extrapolation, will be even more inexpensive and capacious in the coming few years.

LOCKSS holds great promise for use in collecting, maintaining, and providing access to digital government documents as well. Stanford librarians and the LOCKSS team have met under the terms of a grant from the National Science Foundation with officials of the United States Government Printing Office and with government documents librarians to devise specifications for desired functions of a LOCKSS cache of public documents. Two weeks ago the Superintendent of Documents and an entourage of about a dozen people from the GPO visited

Stanford for a day to discuss the possibilities of using LOCKSS caches in two innovative and important modes: United States Federal Depository Libraries may use LOCKSS to automatically gather public documents for public access and long term preservation, and the Government Printing Office may use a LOCKSS cache to gather public documents from the many arms of the United States federal government for re-distribution to the libraries in the Federal Deposit Library Program. Here is a diagram of how that scheme might work.

Why use LOCKSS for government documents? Consider the characteristics of U.S. government documents on the web. First, there are lots of them that have no print version – and this trend is accelerating. Second, web-based government documents are volatile; their half-life is four months, and they change a lot during their short lives. Third, government documents on the web comprise a wide range of formats, but all of them are accessible via http. U.S. government documents on the web are relatively opaque; that is, there are lots of layers in these documents that are not indexed by the common Internet search engines, like Google. Another way of expressing that notion is that web-based government documents have complex structures, with elements of many pages coming from a variety of source files when a reader calls them up. Among the concern all must have for government documents on the web is their authenticity; how easy is it to discern that the documents are not only official, but are seen as issued by the government agency. Federal Depository Libraries become part of the chain of custody of government documents and, in any web publishing or web distribution scheme, must have ways of authenticating documents coming into their hands and servers. We expect the U.S. Government Printing Office to begin their support of LOCKSS by re-writing their systems and DTDs for their serial publications to accommodate the LOCKSS spiders; government issued serials are much like other serials and the current LOCKSS plug-ins could suffice for them as a result of these minor changes.

What is the current state of LOCKSS? There are a number of activities underway. The LOCKSS software is about to be released in a major new version, one designed so that the underlying code, the middle level demons, and the top layer plug-ins each play their role. A new beta test version has been distributed recently to participating libraries. With this three layer design, new plug-ins can be written, not just by the LOCKSS programmers themselves, but by others in the field. Such plug-ins will be needed for specific publishers and for genres other than e-articles. Testing will begin on that new version by many of the original 50 libraries that tested the earlier versions. As we get feedback on the new version, we will improve it. Librarians have been working at Emory and Indiana Universities as well as at New York Public Library to devise an interface for collection development officers to easily specify to their local LOCKSS installations what e-journals to collect for their local caches. I have mentioned the work underway on government documents. We are at the end of the needs assessment phase and ready to go to the development of specifications for use of LOCKSS to cache government documents. We are now seeking outside funding, probably a federal grant, to write and test the LOCKSS plug-ins for government documents.

Finally, we are considering and have been making limited inquiries about how the LOCKSS effort might become self-sustaining. LOCKSS has depended for development on support from Stanford, Sun Microsystems, the Mellon Foundation, and the National Science Foundation. Now we are asking ourselves and others whether libraries might form an alliance and support such an alliance in exchange for training in the use of LOCKSS, help in the operation of LOCKSS, writing new plug-ins, upgrades and bug fixes, and to make LOCKSS known to many libraries around the world. We are considering a proportional scheme in which smaller libraries pay less and larger libraries pay more. Publishers too would pay for annual membership in the LOCKSS alliance, for many of them can see already that LOCKSS is an effective way to address the demand for locally controlled digital archives.

A central LOCKSS team, staffed for maintenance, training, and user support for a mature, system, one changing slowly, could be supported painlessly by an alliance of fewer than 100 libraries, national bodies, and publishers.

In the next year or 18 months, LOCKSS will be widely used to preserve e-articles from scholarly journals. However, that software contains the basis for plug-ins appropriate to retrieve and maintain caches of government documents. And so, in a relatively short time, LOCKSS may help the U.S. Government Printing Office gather and then re-distribute automatically government documents to numerous libraries in the depository program. Obviously, it could be deployed for other national governments' needs similarly. We can see LOCKSS serving non-government organizational documents and similar public policy web sites.

It could even work on e-books and other genres too. If LOCKSS is to have a long-term future and grow with changing technologies, it will need to become financially self-sustaining. Assuming its future depends on a membership model alliance, governance issues will need to be sorted out; member libraries and publishers should and must have a voice in the operations and policies of the LOCKSS alliance.

Another consideration is how LOCKSS fits into the developing pattern of different types of digital repositories. Clearly there will emerge a relatively few, highly managed, very large digital repositories, many, though not all, operated by national and major regional libraries. In all likelihood, there will a network of those repositories. Could LOCKSS be a feeder mechanism to such digital repositories, whether at a national library, a university, or a trusted third party? Some of us at Stanford think that some sort of relationship between the caches of LOCKSS and our own Digital Repository might be quite effective. However, we need to grow the Stanford Digital Repository from its present prototype to an operational phase before we investigate the complexity of adding LOCKSS to the picture.

To conclude, let me suggest that the accumulated expertise of each national library and a few of the large research libraries, along with societies' expectations that each of those entities are effective and on-going cultural custodians might be applied in a roughly coordinated manner to actively collect and preserve the digital documents that comprise the new, if incomplete, record of man. Those great collecting institutions, the national libraries and the few great research libraries, need to and are expected to reflect in their holdings not just the traditional print and archival records of our national heritages, but as well the new media, the digital records too. The collection development policies of these great libraries already have the elements of coordination and in many cases, specifications of format, media, and genre have already been expanded to included the digital ones. LOCKSS may be one of the mechanisms suitable to employ to realize the continuing mandate we have to care for cultural assets, even digital ones. I commend LOCKSS to your careful consideration.

Further information about LOCKSS may be found at: http://lockss.stanford.edu .
Thank you for your patient attention.