D-Lib IST-2001-32587
Digital Library Competence Center

Web Journal Preservation Testbed Report

D4.1.1

Deliverable Type: REPORT
Number: D4.1.1
Nature: Public
Contractual Date of Delivery: month 22
Actual Date of Delivery :month 22
Task WP4.1

Name of responsible: Costantino Thanos, ISTI-CNR

Authors: Vicotria Ann Reich, David Rosenthal, Stanford University, US
Stefania Biagioni, Francesca Borri, Carlo Carlesi, ISTI-CNR

contact info
stefania.biagioni@isti.cnr.it

Abstract

This report details the organization of the course on the Web Journal Preservation Environment and the manuals written for and distributed at the course itself. The complete documentation has been made available to the public through the D-Lib Center web site (http://dlibcenter.iei.pi.cnr.it/).

Executive Summary

LOCKSS: An affordable, cooperative, e- preservation and e-archiving program.
At the D-Lib Center took place a two day course on how to create and provide access to local collections of electronic journals; while simultaneously and affordably preserving and archiving this locally held materials for the very long term. The course presented the strategy, tactics, and technology of the LOCKSS Program (http://lockss.stanford.edu) which software is open-source and is freely available for download and use.

Currently, libraries lease access to web-published journals from the publisher's web sites. The LOCKSS Program provides librarians with affordable tools and methods to uphold their traditional role as society's custodian of scholarly materials. This course provided a philosophical and practical introduction to the techniques and methods for: (1) building and providing access to local e-journals collections; (2) leveraging an international technical and social international preservation and archiving system.

Outline of course

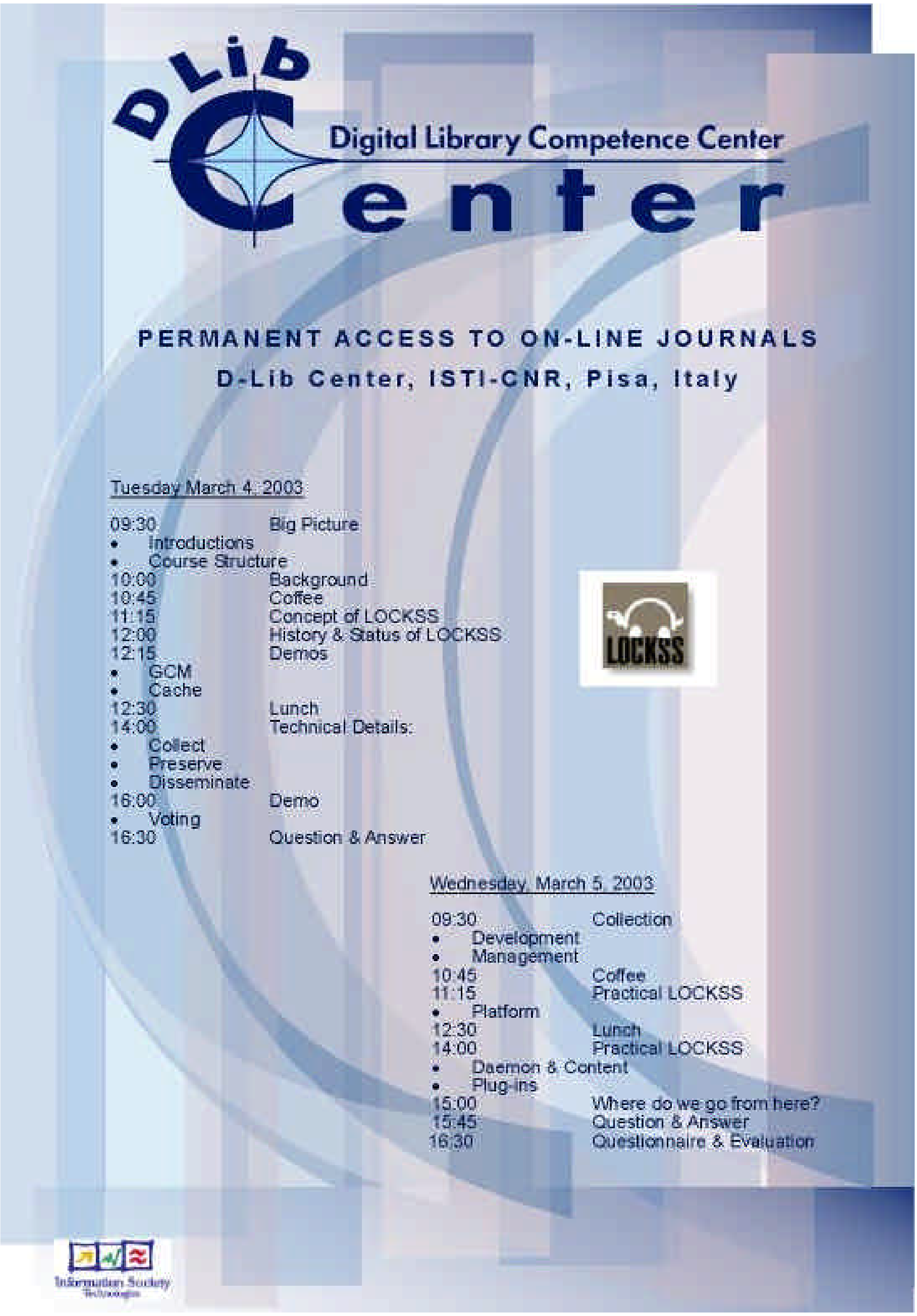
This two day course provided also a senior level introduction to:

- Strategic business and management issues
  - at the local level, what are the costs of building and maintaining local e-journal collections
  - at the international level, how will the LOCKSS program and technology be sustained once grant funding has ceased
- Collection issues
  - collection development (local collection development decision processes; engaging and leveraging publisher relationships, the importance of local consortia, the role of local collections for the international library community)
  - collection management (managing workflow, philosophy and use of preservation metadata for a decentralized archive)
  - collection access (integrating LOCSKS caches into institutional networks; reader access when the publisher's web site is unavailable or a subscription has been cancelled)
- Preservation and Archiving technology
  - Installing and configuring a LOCKSS cache to collect, preserve and archive e-journal content (this is a hands on activity)
  - Situating local LOCKSS caches within institutional networks so readers have seamless access to locally held content when that same content is no longer available from the publisher's web site
  - Customizing and applying the LOCKSS technology to apply to locally or nationally produced e-journals
  - Adapting the LOCKSS technology from e-journals to other web based genres

Teaching Staff:
Victoria Ann Reich, Director LOCKSS Program, Stanford University
David Rosenthal, Chief Scientist, LOCKSS Program, Stanford University
Technical Staff:
Stefania Biagioni, ISTI-CNR, Pisa, Italy
Carlo Carlesi, ISTI-CNR, Pisa, Italy
Francesca Borri, ISTI-CNR, Pisa, Italy

Table of Contents

Agenda



DLib Center
Digital Library Competence Center
LOCKSS
Information Society Technologies

PERMANENT ACCESS TO ON-LINE JOURNALS

D-Lib Center, ISTI-CNR, Pisa, Italy

Tuesday March 4, 2003

09:30	Big Picture
• Introductions	
• Course Structure	
10:00	Background
10:45	Coffee
11:15	Concept of LOCKSS
12:00	History & Status of LOCKSS
12:15	Demos
• GCM	
• Cache	
12:30	Lunch
14:00	Technical Details:
• Collect	
• Preserve	
• Disseminate	
16:00	Demo
• Voting	
16:30	Question & Answer

Wednesday, March 5, 2003

09:30	Collection
• Development	
• Management	
10:45	Coffee
11:15	Practical LOCKSS
• Platform	
12:30	Lunch
14:00	Practical LOCKSS
• Daemon & Content	
• Plug-ins	
15:00	Where do we go from here?
15:45	Question & Answer
16:30	Questionnaire & Evaluation

Letter to the Publisher

Building an Affordable E-journal Archive and Preservation System: LOCKSS

LOCKSS (http://lockss.stanford.edu), Lots of Copies Keep Stuff Safe, began in 1999 at Stanford Libraries with funding from NSF and Sun Microsystems. LOCKSS is an Internet "appliance", or "easy to use" software, designed to preserve access to authoritative versions of web-published journals. With Andrew W. Mellon foundation funding, we are building production quality software and a self sustaining support organization.

LOCKSS uses the caching technology of the web to collect pages of journals as they are published, allowing libraries to take physical custody of selected electronic titles they purchase. Unlike normal caches, however, pages in these caches are never flushed.

Through alpha and beta testing phases, the LOCKS S system is demonstrating that it is both easy and affordable to operate a purchase model for web journals. LOCKSS software is free and open-source and is designed to run on inexpensive hardware. Many publishers are supporting LOCKSS because the system protects both their content and their business models.

What is a Library Anymore, Anyway?

Michael A. Keller, Victoria A. Reich, Andrew C. Herkovic
Stanford University Libraries
direct comments to vreich "at" Stanford "dot" edu

Paper for submission to *First Monday* 18 February 2003

Abstract

Libraries in the future will undertake local control, especially for long-term preservation and accessibility of digital as well as analogue collections. Failure to embrace that role would cause libraries and librarians rapidly to lose relevance and value as Internet and other digital resources develop. Local control of collections is critical both to assure permanence and to provide a key degree of selectivity, which – contrary to the irrational exuberance of making everything available to everybody – is vital to providing service to communities of readers. Librarians need new tools, such as the LOCKSS system, to enable both persistence and selection of electronic information.

Introduction

We have observed a propensity for information technologists to predict with complete confidence the imminent demise of libraries. The seeds of this prognostication may date back to Vannevar Bush's seminal paper of 1945 (1), but the forest of such predictions has grown thick in the past decade. In our observation, the confidence with which such predictions are made is inversely proportional to the predictor's professional habitual use of published information. That is, the prediction that libraries are becoming obsolete or useless is a projection onto the world of the internal set of the speaker and reflects a lack of appreciation of libraries' deep, often hidden functions, especially in the realm of digital information resources. Long-term or intensive library users, on the other hand, rarely if ever, in our experience, foresee this demise, except as a distopian nightmare. And indeed, we often note that when some technologists talk of information, it is as an essentially quantitative abstraction, something to manage rather than use – in short, a commodity – not as the lifeblood and substance of scholarly (2) inquiry and endeavour.

So to some extent the question of what is a library anymore could be one about which, "if you have to ask the question, you wouldn't understand the answer." Our intention is to attempt the answer anyway. This intention is inspired, if not modelled, on the remarks of Gerhard Casper, then president of Stanford, at the 1999 rededication of the restored central campus library, "Who Needs a Library, Anyway?"(Casper 1999).

Discussion

As librarians, our task has always been and remains to use the tools of information technology to serve the needs of information workers or seekers, bearing in mind that all recorded information and its support infrastructure – books and shelves no less than online search engines – is information technology. Actually, understanding what is a librarian anymore may be easier than understanding what is a library. There can be, and increasingly may be, librarians without libraries (in the sense that they are not based at a single physical institution). But can there be libraries (or even the

Universal Virtual Library as some have imagined) without librarians? Or without tangible institutional existence (per the bricks and mortar cliché)? Obviously, many think so. Others, particularly those responsible for institutional budgets, may hope so. We think this is nonsense.

Ignoring the physical, technological underpinnings for now, we assert that the library is, at root, a collection of information selected for use of, and made useable for, a particular community. (3) That community may be large or small, physically proximate or not, present or future, homogeneous or not, but it is essential that it be identified and at least partially understood. That is, proverbially like politics, all collections are local.

The roles of collections in library services were enumerated by Michael Buckland in his 1992 manifesto (4) as preservation, dispensing, bibliographic and symbolic. He added, "If these are the four purposes of collections of library materials, we need to inquire how the change from paper to electronic media my change how we seek to effect these roles." He went on to discuss the "'owning version borrowing' trade-off" in the context of electronic vs. paper-based documents, while acknowledging that use, law, and publishing were in a state of flux. Since 1992, nearly everything about electronic information has changed, but we think many or all of the fundamental services and roles to which Buckland refers remain in principle unchanged. Our own enumeration of these functions is as follows: selection and acquisition for collections: provision of intellectual access to local and remote collections; interpretation and discovery information for patrons; distribution of information resources; and preservation of those resources. To these five, which obtain equally well for digital and analogue resources, we may add a sixth, the provision of clean, well-lit, relatively quiet places for books and readers, a physical function with some implications for virtual library spaces.

No collection is perfect or complete. Thus, there is a legacy, dating back at least for a millennium and a half, of interlibrary loan. We owe the existence of almost all ancient texts, with the exception of those preserved in cuneiform, to the practice of individual libraries or scriptoria borrowing and making local copies for local use of other libraries' manuscripts. They did not do so to save Western Civilization; they did so to have and to hold those texts for local purposes. So, today, if a scholar at Berkeley, to use a parochial example, needs a book not available at the University of California, she may use Stanford's copy either by visiting or simply by requesting it online. In an imagined future, she might expect the complete work to be available electronically and instantly. And while we are deeply engaged in creating electronic versions of existing texts and devising ways of making them more useful to readers and scholars, we know there are profound issues about the bland assumption that everything will just be there online someday.

The extension of this assumption is that, once content ceases to be a local concern, there will not be a need for Berkeley, or Stanford, or anyone else to hold collections. This brings us back to the notion that libraries are obsolete and fated for oblivion. For some years, some in the library field have been creeping in this direction under the rubric "access, not ownership." There is a trend toward redefining what constitutes a collection:

> Instead of describing collections as "those things owned", a better definition may be "information resources for which the library invests financial resources-- directly or indirectly--to manage, service, or preserve on behalf of library users, regardless of the location of content." "Collections" now include resources owned by the library and those accessed in remote locations; the norm is now an interdependent mix of ownership and access, with the location of the material increasingly irrelevant to users (5).

As recently stated by the ARL Collections and Access Issues Task Force, librarians are "Expanding the Definition of Collections":

> Libraries have expanded the traditional view and definition of collections so that the concept no longer equates with those materials that the library "owns". The boundaries have expanded far beyond the print collections on site or the electronic files mounted locally to include electronic materials licensed or managed by the library and materials available through consortia. Increasingly libraries are taking responsibility for born-digital collections

> (such as geo-spatial or numeric data sets, faculty or class Web sites) and developing tools for their management and use. In a growing number of cases, a library's collection also includes resources that reside outside the domain of the library but for which the library takes *some* responsibility for managing and servicing. (emphasis added) (6)

The centrality of access is not at issue here. Our concern is not with the library's proper role in providing such collections, nor is it with providing tools (read, web pages) for their use. We are concerned, however, that this masks an evolution from a model of professional guidance to a community of readers – which we heartily endorse and practice – into a larger vision of "libraries" abandoning responsibility for physical collections – a dysfunctional and possibly doomed vision, in our view. A web page with a set of links is not a library. If we understand libraries simply to be nodes in a global digital network of common resources, nobody retains any responsibility for either those materials or serving the needs of the community. In this environment, there would indeed be no meaningful role for libraries. We do not see this as serving the needs or interests of our readers, particularly if viewed over time. We have come to understand "access, not ownership" more as rationalization of constrained choices than as a functional understanding of libraries' service to their communities, other than purchasing and licensing agents.

It has been true for several years that the printed editions of some leading scientific journals are no longer complete or authoritative; the authentic versions of record are the electronic editions, which may have more articles, embedded or linked supplementary information (in electronic form) appended to the articles, or editorial content that cannot exist in print form (simulations, animated models, video or sound recordings, etc.). This is a wonderful development (7), and for this among many reasons we favour subscribing to the electronic edition, but it leads to the concern that we have no way to assure that this material will be available to our readers in one, two, or ten years' time (whether or not we attempt to continue subscribing to the journal). Whether we keep the printed edition or throw it away, our readers' future access to the *full* journal contents is in jeopardy.

At a meeting in Philadelphia in January 2003 (in conjunction with the American Library Association's midwinter conference), Theodore Bergstrom reportedly "got librarians' attention by asking a provocative question: just why are libraries involved in subscribing to e-journal site licenses when the e-journals aren't residing in the library, and are being used largely outside of the library (8)?" Subsequent discussion raised the question of boycotting site licenses, apparently for the purpose of redressing for-profit publishers' predatory pricing and bundling policies. Without commenting on the propriety of that strategy, we suggest a corollary motivation: the impermanence of what site licenses make momentarily accessible (9).

Somebody somewhere will do something, right? Lots of things have been tried, at least in principle, most of which have depended on: lead agencies acting ostensibly on behalf of all others indefinitely, and; publishers abdicating control of their proprietary content to those lead agencies.

Despite the best of intentions, such approaches have been largely stillborn. We are willing to assume that such approaches will eventually take hold at least selectively and indeed, we intend to take part in them. However, they will likely be highly selective, as to both content and beneficiary, and will be gambles on the part of librarians and leaps of faith on the part of publishers. Meanwhile, the vast majority of libraries are left as passive holders of others' promises. We think some professional scepticism is called for.

Publishers, may be inclined at times to "un-publish" articles and, in the case of the US Federal Government, even entire volumes. (The reader may wish to review the recent exchanges about Elsevier's removal of articles on the Liblicense -L discussion list archive at http://www.library.yale.edu/~llicense/index.shtml) If they control the only server on which electronic material exists, there is no recourse, a grotesque disservice to science, to scholarship, and to the public good. Our colleague Jim O'Donnell wrote the following to the Liblicense -L discussion:

> The discussion over publisher-removed articles is of course a discussion over the reliability of archives. We are accustomed to being able to go back to published material long after the

> fact and to find a stable and accurate record of what was said. Traditionally, libraries have been the guarantors of this process: preserving many copies, with no legal liability for the content (or at least less than the publisher might have), and with an institutional commitment to permanence and preservation. The "vanishing act" discussion highlights a feature of unreliability of e-archives that depends (1) on the physical malleability of the record and (2) on the slightly lower commitment to full preservation that a publisher might have. It is disturbing, because it is the tip of the iceberg, I think: if for fairly transient reasons, publishers will pull articles, when might not publishers prove unreliable for other reasons?
>
> But the question that follows on this discussion for me is this: If we were to ask that not publishers but authors be the guarantors of permanence, self-publishing or publishing in institutional repositories where the author retains control over the copyright and disposition of his/her material -- what protection do we then have to assure us that articles will remain archived, unchanged, in perpetuity? Are there articles I have written that I wouldn't mind disappearing? Actually, yes. Are there pieces of articles that I would quietly change if I could? Well, interesting thought, sure.
>
> Is it important that the record abide? Then should not all discussions of epublishing for scholarly purposes include a discussion of preservation that includes not only the physical vulnerability of the media but their psychosocial vulnerability? What guarantors other than libraries do we realistically have (10)?

Some publishers make one or another surrogate available to subscribers on tape or CDROM, with or without charge. This creates considerable challenges for maintenance and management, and may or may not feature the links and other functions that make the online edition distinctively useful. For example, Reed-Elsevier makes its journal content available to subscribing libraries on tape. Few libraries (11) in North American are known to load the Elsevier tapes, presumably because of both cost and storage management issues. So, almost everyone depends on continued vendor -based access to these materials. Tape delivery is not even an option for most publishers' journals or databases (or any other type of material). Is this an acceptable situation for libraries? Instinctively, we think not. As we stated earlier, a web page with a set of links (to publisher sites) is not a library, and a web page with a set of obsolete, denied or expired links is nothing at all.

For these reasons, Stanford is developing (12) the LOCKSS system and protocol, which allows libraries to create, manage, and maintain persistent caches of e-journal content to which it subscribes. We will not here go into detail of how LOCKSS works (13), but its salient features are that each library that wishes to physically possess and control ejournal content to which it has subscribed can do so, with minimal effort and cost, on behalf of its readers. When the publisher website goes down – be it for an hour or forever – the library can serve up that content to its users. The library is no longer reduced to being a web page with a set of links – a mere pass-through – vis à vis ejournals. The traditional virtue of libraries – to assure access to a community of readers by holding selected content of importance to those readers – is reasserted. Another virtue of libraries, that they have a high degree of collective redundancy, is also replicated in the e-journal sphere by the LOCKSS approach of multiple caches at different institutions. We emphasize that redundancy is a virtue, as structural engineers have learned. Any system or environment that rests on a single point of failure, whether it be the Alexandrine Library (ver. 1.X), a hard disk, a vendor's website, or even the Library of Congress, is subject to failure.

Our colleague Harold Billings at the University of Texas Austin raises these issues clearly:

> I have just suggested to my staff that we should "add" these publications [two guides to open-access publishing released online by The Budapest Open Access Initiative] to our collections, and "catalogue" them. But without exploring them too deeply, I then immediately wonder: how do we know these publications, digital in PDF, will continue to exist after we add and catalogue them? This is obviously the same kind of Question that one has to ask of all electronic publications, journal or otherwise. The most secure method would be to grab them and download them into our own servers. I guess this illustrates how

> important the ongoing existence of any publisher is. This is why I find individual archiving to be without any merit, while institutional archiving -- and subsequent networking – could have an important future if the archive is maintained by a trusted university, organization or commercial entity. This is where the concept of Stanford's LOCKSS comes into important play (14).

This is even more significant in the case of electronic-only journals, i.e., those that do not exist at all in print (and may be sitting unprotected on somebody’s desktop server). Ejournals constitute a single, albeit vital and costly, class of collection materials. There is simply no way to collect the rapidly growing corpora a of lost-cost or no-cost webpublished materials which are of incalculable value and interest, but that are unlikely to survive the years, to say nothing of the decades. Web sites have taken on the historical roles and research value of *samizdat*, avant-garde magazines, seditious literature, fringe political manifests, etc. The Internet houses the New Underground: highly specialized, highly controversial, sometimes dangerously political, and all extremely ephemeral. In reference to literary and critical journals not being collected (because they exist only online), our colleague William McPheron states, “If they were in paper, we’d be getting them.” At present, he cannot follow his curatorial instinct to collect and retain them.

Several libraries, working with the Council on Library and Information Resources, are currently conducting studies regarding the collection and retention of political web sites in various parts of the world; without such efforts, much of contemporary political evolution in unstable regions may be permanently untraceable, precisely because 1) the documents of interest were published only on the Internet, and 2) nobody – with the possible and notorious exception of the winners – will make the effort to collect and retain them (15).

It is worth noting that a group of government documents librarians are exploring the possibility of adapting the LOCKSS system to build and maintain collections of US and local government documents (16). Others are exploring applying LOCKSS to medial gray literature (mostly evidence-based) and other corpora of literature, such as university theses and dissertations. What we (as interlocutors for our readers) need is not indiscriminate warehousing of ephemeral web sites, but selective, intelligent, targeted collection development – exactly what librarians have always done and, by and large, done well in service to their local communities. In clear distinction to the prevailing dynamics of the Internet – “Fast, Cheap, and Out of Control,” to borrow a documentary film title (17) – we need to be deliberate about what we gather, control carefully what we do gather, and even more deliberate about discarding information. (By no coincidence, the logo for the LOCKSS system (18) features a tortoise as a metaphor for being slow, but long-lived. This metaphor appears in a number of cultural contexts, ranging from Aesop to Native American legend and art.

New tools are needed to find, evaluate, select, and preserve such content. Whether they will resemble LOCKSS, how labour intensive they will be, how affordable they will be are all unknown. However, we assert that such tools will provide value and utility only to the extent they allow individual libraries to make their own collection decisions, to control file content locally, and to serve their own communities through their own collections.

Conclusion

Whether or not consciously, libraries and librarians have long been prominent among the few kinds of social agencies that have preserved continuity of cultural heritages. By serving as custodians of local collections, they have incidentally served a larger common good. Whatever other public benefits they provide, publishers and Internet promoters do not, and cannot be expected to, fulfil this custodianship role.

That libraries may be becoming obsolete is, to some degree, plausible, not because they are losing some kind of competition with the Internet for eyeballs or compellingly superior content, but rather because libraries may be in the process of abandoning their role as collection builders and managers. We do not suggest that anybody is to blame in this as a practical matter; we haven't yet figured out how to retain that role in an increasingly (if never exclusively) digital information universe. The economics and limitations of technology (with some enormous intellectual property issues on the side) are serious, if temporary, impedimenta. But if we don't assert the importance of that role, of the centrality of selecting, acquiring, retaining, preserving and building means of access to collections, we will inevitably fade away.

Acknowledgements

The authors thank Don Waters of the Andrew W. Mellon Foundation; The National Science Foundation, and Sun Microsystems for their continued support of the LOCKSS Program. We also thank the many libraries and publishers who are participating in the beta test and who are giving of their time and resources to make this program viable.

Notes

1. Vannevar Bush, "As We May Think" *The Atlantic Monthly*; July, 1945; Volume 176, No. 1; pages 101-108. http://www.theatlantic.com/unbound/flashbks/computer/bushf.htm
2. Scholarly in the inclusive sense of informal as well as formal inquiry
3. Although a good library is as much a service organization as a collection, we focus for now on the collection aspect.
4. "Chapter 6: Collections Reconsidered" Michael Buckland, *Redesigning Library Services: A Manifesto* . Chicago : American Library Association. 1992. Internet edition 1997 at http://sunsite.berkeley.edu/Literature/Library/Redesigning/html.html
5. Collections & Access for the 21st-Century Scholar: Changing Roles of Research Libraries, A Report from the ARL Collections & Access Issues Task Force. *ARL Bimonthly Report 225* December 2002 http://www.arl.org/newsltr/225/main.html
6. ibid.
7. Encouraged and implemented by Stanford's HighWire Press, among others
8. "Economics Lesson Leads to Inspiration: Is a Site License Boycott by Libraries Possible?" Library Journal Academic News Wire, January 30, 2003.
9. This begs the obvious question of what demand might such an action invoke, and we think there are several possible answers, ranging from the very vague, e.g., some tangible guarantee that subscribers at least will never lose access to content for which a site license was ever bought, to the very specific, e.g., active cooperation with the LOCKSS system.
10. e-mail from Jim O'Donnell to Liblicense-L Wed. 29 January 2003. Subject "Re: vanishing act" http://www.library.yale.edu/~llicense/ListArchives/0301/msg00118.html
11. Among them Los Alamos National Laboratory, the University of Toronto, Ohio Link; The National Library of the Netherlands has announced it is archiving this material as well
12. with generous grants from the Andrew W. Mellon Foundation, the National Science Foundation, and Sun Microsystems
13. which is available at http://lockss.stanford.edu
14. e-mail From Harold Billings to fox-forum@topica.com Thursday, January 30, 2003, subject "Re: two important guides to open-access publishing"
15."Political Communications Web Archiving : A Proposal to the Andrew W. Mellon Foundation," Center for Research Libraries, July 26, 2002. The Mellon Foundation has funded this project.
16. Funded by a Small Grant for Exploratory Research from the National Science Foundation
17. http://www.sonypictures.com/classics/fastcheap/
18. visible at the LOCKSS home page http://lockss.stanford.edu/

References

- ARL Collections & Access Issues Task Force, 2002, Collections & Access for the 21st-Century Scholar: Changing Roles of Research Libraries, A Report. *ARL Bimonthly Report 225* December 2002, http://www.arl.org/newsltr/225/main.html
- H. Billings, 2003, email to fox-forum@topica.com Thursday, January 30, 2003, subject "Re: two important guides to open-access publishing"
- M. Buckland , 1997, "Chapter 6: Collections Reconsidered" , *Redesigning Library Services: A Manifesto* . Chicago: American Library Association. 1992. Internet edition 1997 at http://sunsite.berkeley.edu/Literature/Library/Redesigning/html.html
- V. Bush, July, 1945; "As We May Think" *The Atlantic Monthly*; Volume 176, No. 1; pages 101-108. http://www.theatlantic.com/unbound/flashbks/computer/bushf.htm
- G. Caspar "Who Needs a Library, Anyway?" [This is the text of President Gerhard Casper's remarks to the Stanford community at the dedication of the Bing Wing of the Cecil H. Green Library on October 12, 1999]. Stanford University Libraries, 2000. http://www.stanford.edu/dept/news/report/news/october13/libtext-1013.html
- Library Journal, 2003, "Economics Lesson Leads to Inspiration: Is a Site License Boycott by Libraries Possible?" Library Journal Academic News Wire, January 30, 2003.
- J. O'Donnell, 2003, email to Liblicense -L Wed. 29 January 2003. Subject "Re: vanishing act" http://www.library.yale.edu/~ llicense/ListArchives/0301/msg00118.html
- CRL 2002, "Political Communications Web Archiving : A Proposal to the Andrew W. Mellon Foundation" the Center for Research Libraries, July 26, 2002 http://www.library.cornell.edu/iris/research/WebPolCom.pdf

About the Authors

At Stanford, **Michael A. Keller** is the Ida M. Green University Librarian, Director of Academic Information Resources, Publisher of HighWire Press, and Publisher of the Stanford University Press. These titles touch on his major professional preoccupations: commitment to support of research, teaching and learning; effective deployment of information technology hand-in-hand with materials; active involvement in the evolution and growth of scholarly communication. He may be best known at present for his distinctively entrepreneurial style of librarianship. As University Librarian, he endeavours to champion deep collecting of traditional library materials (especially of manuscript and archival materials) concurrent with full engagement in emerging information technologies.

Keller was educated at Hamilton College (B.A. Biology, Music 1967), SUNY Buffalo (M.A., Musicology, 1970)), SUNY Geneseo (M.L.S., 1971), and SUNY Buffalo (a.b.d. Ph.D., Musicology). From 1973 to 1981, he served as Music Librarian and Sr. Lecturer in Musicology at Cornell University and then in a similar capacity at UC Berkeley. While at Berkeley, he also taught musicology at Stanford University and began the complete revision of the definitive Music Research and Reference Materials, an annotated bibliography popularly known as "Duckles" in honour of its original compiler. Yale called him to the post of Associate University Librarian and Director of Collection Development in 1986. In 1993, he became the Ida M. Green Director of Libraries at Stanford. In 1994, he was named to his current position of University Librarian and Director of Academic Information Resources. In 1995, by establishing HighWire Press, he became its publisher, and in April 2000, he was assigned similar strategic duty for the Stanford University Press.

Vicky Reich works to facilitate the industry's transition from print to online publishing models. She is Director and co-founder of the LOCKSS Program, which allows libraries to retain local collection control of materials delivered through the web while preserving the functionality of the original web based content. (see http:// lockss.stanford.edu). Prior to the LOCKSS Program, she was, for eight years, the Assistant Director of HighWire Press, during which she led its Market Support Services to build international relationships with academic and corporate librarians. She has over 20 years of extensive library experience in both public and technical services and in both public and private institutions. She has held positions at the: Upjohn Company; University of Michigan; the Library of Congress; National Agricultural Library; and Stanford University. She earned her MLS from the University of Michigan.

Andrew Herkovic is responsible for foundation, corporate, and government relations at the Stanford University Libraries. His duties involve public relations, project management, speech and proposal writing, publishing, and liaison with partners in library initiatives. Past project interests include investigation of the scholarly use of e-journals; perceived crises in scholarly communications based on journal pricing and rights management; and leadership development for librarians. His has held positions in academia and business since his first professional incarnation as a library cataloguer at Cornell University, of which he is a graduate.

LOCKSS Alliance

The Alliance will be supported by publishers and librarians who wish to use LOCKSS technology to assure long term access to scholarly materials and to leverage Alliance resources for their institutions. Alliance leaders will oversee LOCKSS technology development and maintenance, and guide technology applications. They will have a collective investment to build services and community around this technology to support their local institution’s priorities.

Context

For centuries libraries and publishers have had stable roles : publishers produced information; libraries kept it safe for reader access. There is no fundamental reason for the online environment to force institutions to abandon these roles.

The LOCKSS model capitalizes on the traditional roles of libraries and publishers. LOCKSS creates low-cost, persistent digital "caches" of authoritative versions of http-delivered content. All file formats delivered through HTTP are included (html, jpg, gif, wav, pdf, etc.). The LOCKSS software enables institutions to locally collect, store, preserve, and archive authorized content thus safeguarding their community's access to that content. The LOCKSS model enforces the publisher’s access control systems and, for many publishers, does no harm to their business models. The current version of LOCKSS software is restricted to electronic journals.

Accuracy and completeness of LOCKSS caches are assured through a peer-to-peer polling and reputation system (operated through LCAP, LOCKSS' communication protocol), which is both robust and secure. LOCKSS replicas cooperate to detect and repair preservation failures. LOCKSS is designed to run on inexpensive hardware and to require almost no technical administration. The software has been under development since 1999 and is distributed as open source. See http://lockss.stanford.edu/projectdescbrief.htm#Medium for a fuller description of the LOCKSS software and protocol.

Mission

The LOCKSS Program has as its mission to build tools and to provide support to:

o Libraries, so they can easily and affordably create, preserve, and archive local electronic collections

- o Own rather than lease electronic information
- o Retain traditional custodial role of scholarly information
- o Provide continuing and perpetual access to their local community

o Publishers, so they can easily and affordably provide content to the libraries for preservation and archiving

- o With minimal risk to their business model or to their publishing platforms
- o Ensure perpetual access to their materials.
- o Fulfil librarians’ requirements that publishers guarantee both continuing (day to day) and perpetual (the very long-term) access to content sold.

The mission of the LOCKSS Alliance is to assure the viability, dissemination, and utility of the Program for librarians and publishers.

Key Services

The LOCKSS Alliance will provide services to libraries and to publishers in three broad areas:

I. Technology *(extensive support to Alliance supporters; limited support in the way of access to the source code and documentation to larger community)*

- Software Development: assure the continued viability and growing utility of the LOCKSS software
- Support and Service: Remain alert to, and respond to yet unseen ways the software will need to be maintained including:
    - o extending the plug-in module and architecture to accommodate new publishing styles and formats
    - o reporting, tracking, and repairing bugs
    - o tracking and responding to security issues
    - o assistance with local installation, maintenance, and policy development
    - o monitoring network traffic loads and adjusting plug in specifications as needed
- Transfer of Expertise: Build an open source technical community so that LOCKSS software expertise is transferred to the community (see Attachment 2).

II. Collections *(extensive support to Alliance members; no support to larger community)*

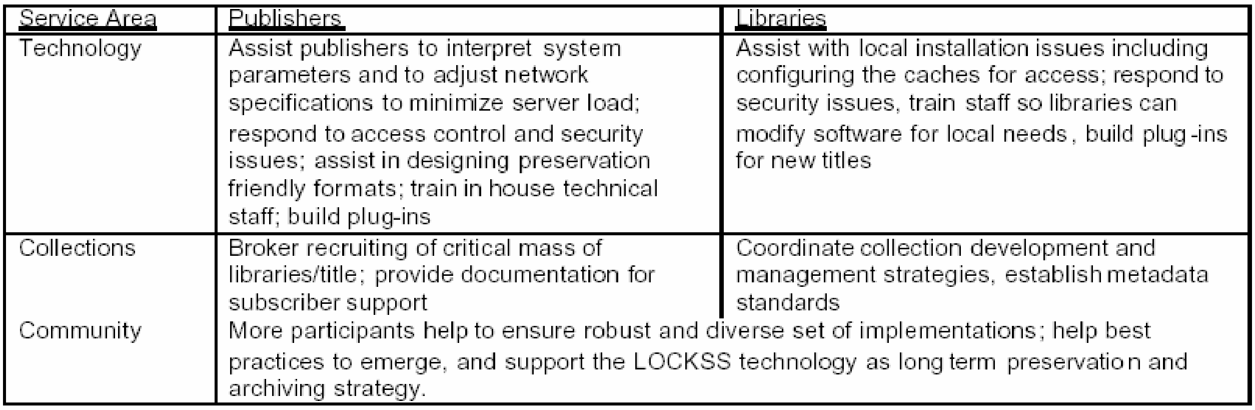
- Broker relationships between libraries and publishers
    - o Ensure a sufficient number of caches per title
    - o Share plug-in applications for different publishing platforms
    - o Disclose (tell community which publishers permit LOCKSS) publisher agreements
- Influence metadata standards and implementation
- Share and refine implementation strategies
- Build expertise and best practices (see Attachment 2)
- Expand beyond “traditionally formatted” e-journals

III. Community

- Recruit supporters
- Communicate and respond to the needs of current and prospective supporters
- Build a shared understanding of intellectual property issues
- Develop the LOCKSS brand

In short, the Alliance will assist libraries to build, preserve and provide access to local collections of web-based materials. The Alliance will assist publishers to provide an important service (preservation and archiving) to a core customer base and to potentially increase their electronic journal market. Several larger publishers have, for example, expressed concern that librarians will choose to cache only the “more important” titles on their list, and not preserve the “smaller titles”. While the librarians claim they are interested in preserving all subscribed to titles, this is exactly the kind of brokering the Alliance is designed to facilitate.

Examples of specific support activities



Service Area	Publishers	Libraries
Technology	Assist publishers to interpret system parameters and to adjust network specifications to minimize server load; respond to access control and security issues; assist in designing preservation friendly formats; train in house technical staff; build plug-ins	Assist with local installation issues including configuring the caches for access; respond to security issues, train staff so libraries can modify software for local needs, build plug-ins for new titles
Collections	Broker recruiting of critical mass of libraries/title; provide documentation for subscriber support	Coordinate collection development and management strategies, establish metadata standards
Community	More participants help to ensure robust and diverse set of implementations; help best practices to emerge, and support the LOCKSS technology as long term preservation and archiving strategy.	

Organization

The LOCKSS Alliance will be a non-profit services organization, based within the Stanford University Libraries, for affiliated institutions . A Board of Advisors (approximately 12 members) will broadly oversee the community's needs and represent major scholarly community constituents. The Board's role is to provide broad experience and counsel in support of the LOCKSS Program's mission.

The Alliance is intended to provide "medium term" transitional support for international decentralized "very long term" preservation and archiving strategy. We do not anticipate how long support will be needed – nor what will be the circumstances under which the Alliance should be dissolved. However, the Director, the University Librarian, and the Board of Advisors are charged with ensuring the Alliance structure is open to change and that there is an internal mechanism for structural review.

Fees

Our goal is to fund the LOCKSS Alliance solely through fees, beginning in 2006. The Board will advice the Director and the University Librarian on appropriate annual fees and if, and/or how, "contributions in kind" will be accepted.

I. Libraries - We suggest library fees be based on institutional type as outlined below. This is an industry-standard model, used by JSTOR, BioOne, and American Physical Society for e-journal access.

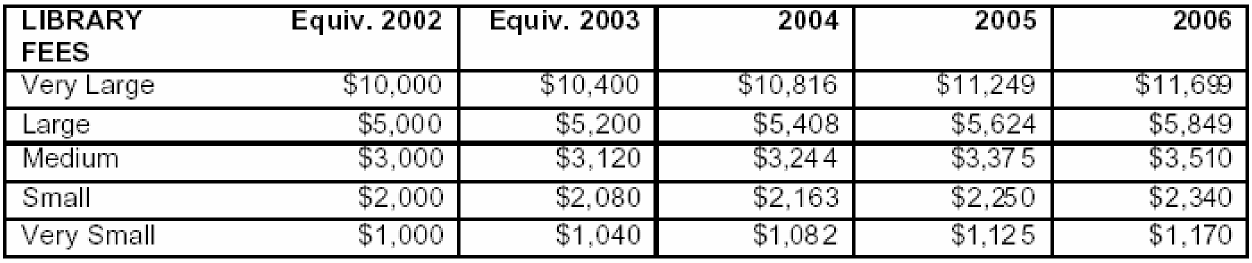
These categories may not translate gracefully to non-U.S. libraries; foreign libraries will be urged to contact the Alliance to determine fair fees. The categories also below do not apply to public libraries, consortia, for-profit, or large government agencies; fee structures for such libraries will be determined as necessary. Special rates will be available for libraries in developing countries.

Rates for U.S. academic libraries:

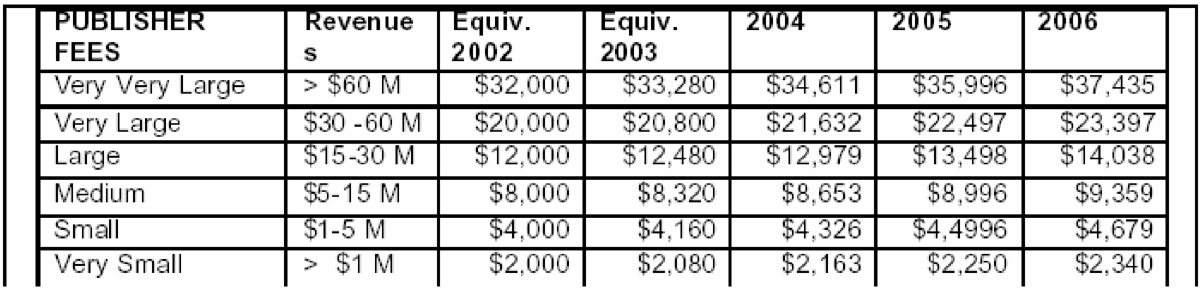
- Very Large
    - o All institutions classified by Carnegie as Research I
- Large
    - o All institutions classified by Carnegie as Research II or Doctoral I
- Medium
    - o All Doctoral II and Masters I institutions with FTE enrolments above 2,500
- Small
    - o Doctoral II and Masters I institutions with enrolments below 2,500
    - o Masters II and Bachelors I colleges with FTE enrolments of 1,000 or more

- Very Small
  - o Doctoral II and Masters I institutions with FTE enrolments below 1,000
  - o Masters II and Bachelors I colleges with FTE enrolments below 1,000
  - o Bachelors II institutions
  - o Centers for advanced study and foundation libraries

II. Publishers - We suggest basing publisher fees on gross total journal revenue. It is an open question whether ad revenue would be included in this calculation. The model below combines elements of the AAP and the ASLSP membership pricing schemes.



LIBRARY FEES	Equiv. 2002	Equiv. 2003	2004	2005	2006
Very Large	$10,000	$10,400	$10,816	$11,249	$11,699
Large	$5,000	$5,200	$5,408	$5,624	$5,849
Medium	$3,000	$3,120	$3,244	$3,375	$3,510
Small	$2,000	$2,080	$2,163	$2,250	$2,340
Very Small	$1,000	$1,040	$1,082	$1,125	$1,170



PUBLISHER FEES	Revenues	Equiv. 2002	Equiv. 2003	2004	2005	2006
Very Very Large	> $60 M	$32,000	$33,280	$34,611	$35,996	$37,435
Very Large	$30 -60 M	$20,000	$20,800	$21,632	$22,497	$23,397
Large	$15-30 M	$12,000	$12,480	$12,979	$13,498	$14,038
Medium	$5-15 M	$8,000	$8,320	$8,653	$8,996	$9,359
Small	$1-5 M	$4,000	$4,160	$4,326	$4,4996	$4,679
Very Small	> $1 M	$2,000	$2,080	$2,163	$2,250	$2,340

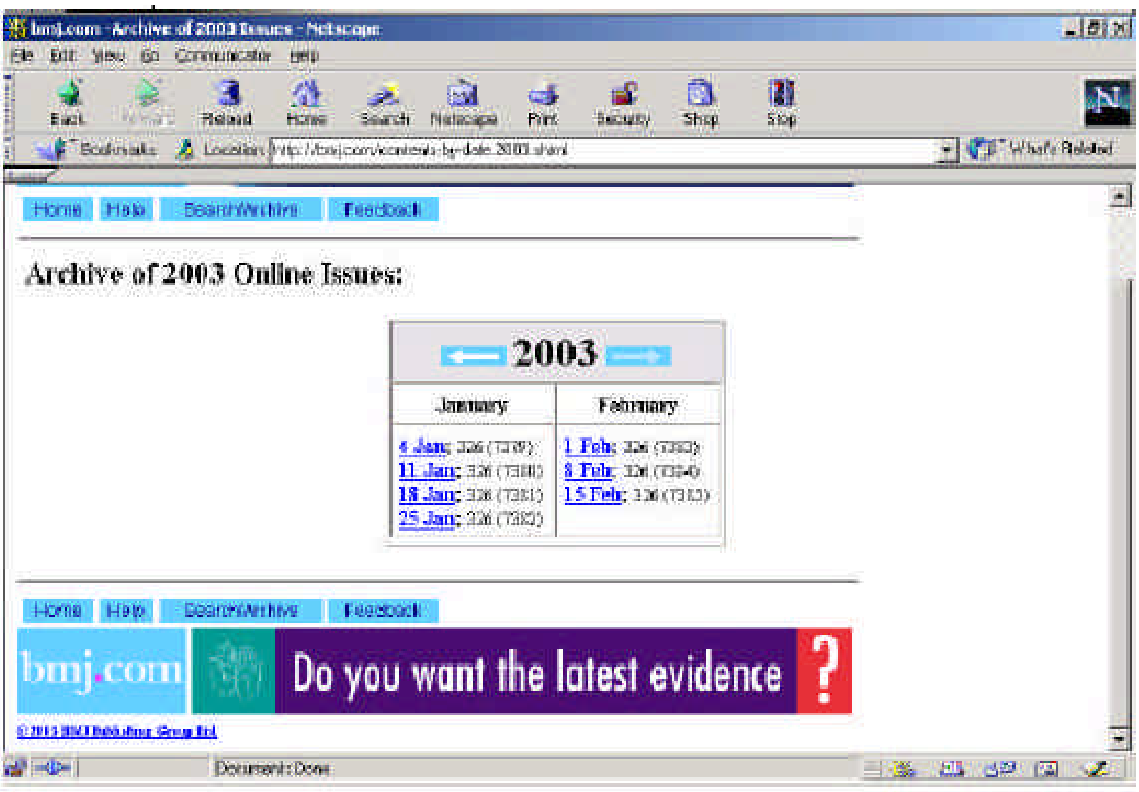
These fees for libraries and publishers project a 4% annual increase based on 2002 dollars.

LOCKSS Publisher Manifest

Draft February 19, 2003

A. The publisher manifest is used for:
1. A starting point for the collection of an Archival Unit (journal volume). The publisher manifest will be the first page our web crawler collects. It needs to link, either directly or indirectly, with every other page that is a part of that AU.
2. Verification that the publisher has given permission for LOCKSS to cache its content. The publisher manifest also demonstrates that the publisher is aware that LOCKSS is being used to cache its content, and that they give permission for this. Ideally, there would be wording to the affect on the manifest, or at least the word LOCKSS in the url that points to the manifest.
3. Containing AU-level metadata. LOCKSS caches preserve the content that they collect, including the publisher manifest page. This is where metadata about the archival unit as a whole should go. The LOCKSS caches will know how to interpret metadata from a publisher manifest.

Most journals already have a volume index page, which fulfils # 1 above, for example:



http://bmj.com/contents-by-date.2003.shtml

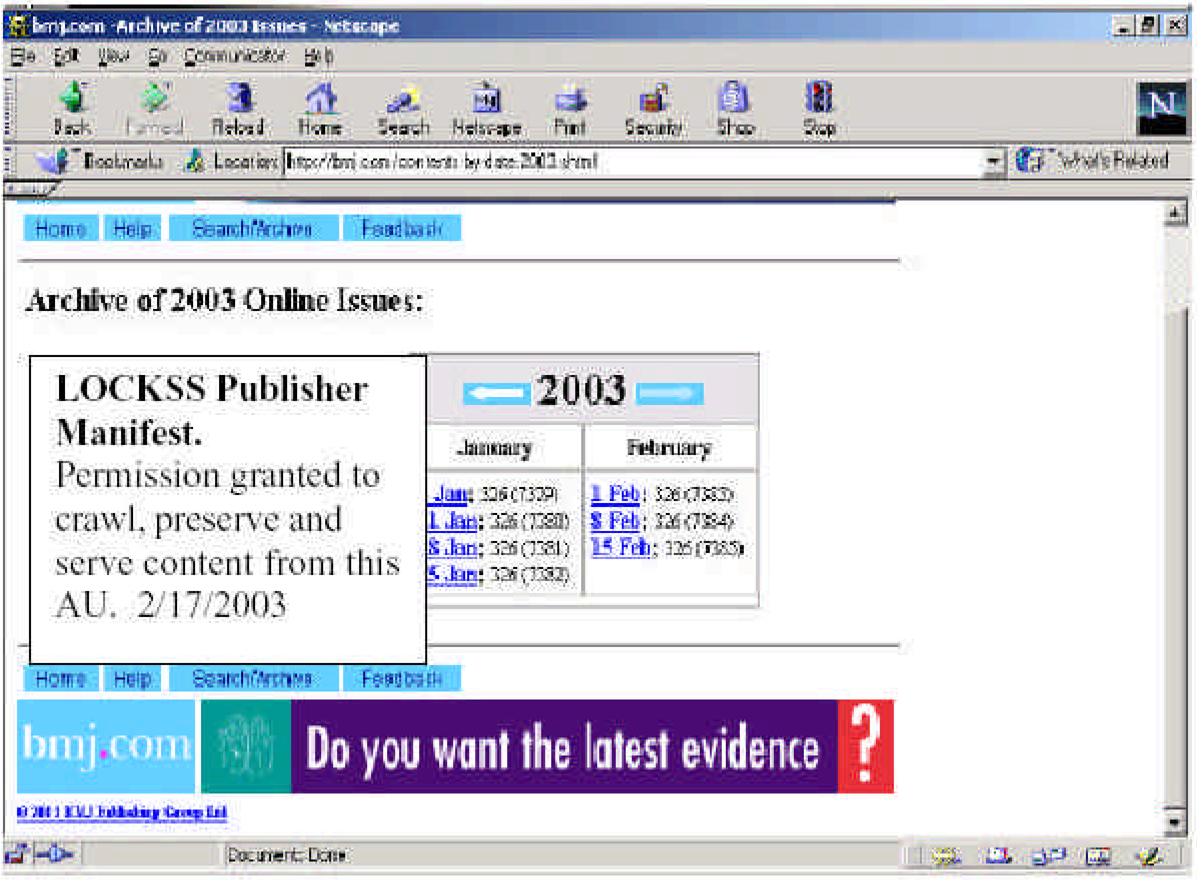
Archive of 2003 Online Issues:

2003

January	February
4 Jan: 326 (7379)	1 Feb: 326 (7383)
11 Jan: 326 (7380)	8 Feb: 326 (7384)
18 Jan: 326 (7381)	15 Feb: 326 (7385)
25 Jan: 326 (7382)	

© 2003 BMJ Publishing Group Ltd

#2 above can be achieved by creating a version of this index page with a statement that contains the word, “LOCKSS”. For example:


bmj.com - Archive of 2003 Issues - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: http://bmj.com/contents-by-date.2003.shtml
What's Related
Home Help Search/Archive Feedback
bmj.com
Do you want the latest evidence ?
Document: Done

Archive of 2003 Online Issues:

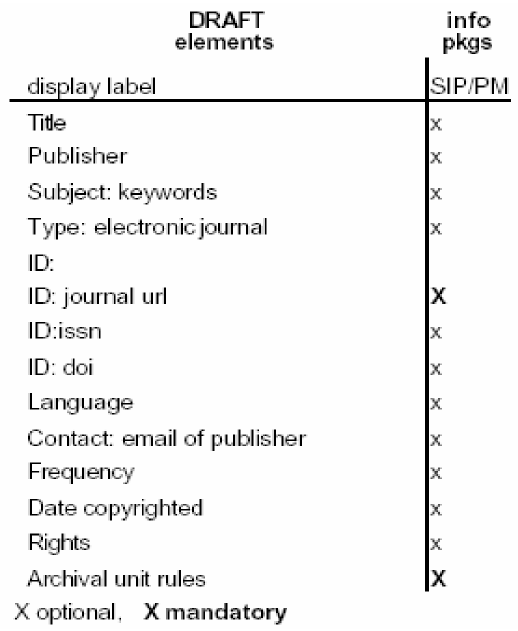
LOCKSS Publisher Manifest.
Permission granted to crawl, preserve and serve content from this AU. 2/17/2003

2003

January	February
Jan: 326 (7379) 1 Jan: 326 (7380) 8 Jan: 326 (7381) 5 Jan: 326 (7382)	1 Feb: 326 (7383) 8 Feb: 326 (7384) 15 Feb: 326 (7385)

© 2003 BMJ Publishing Group Ltd

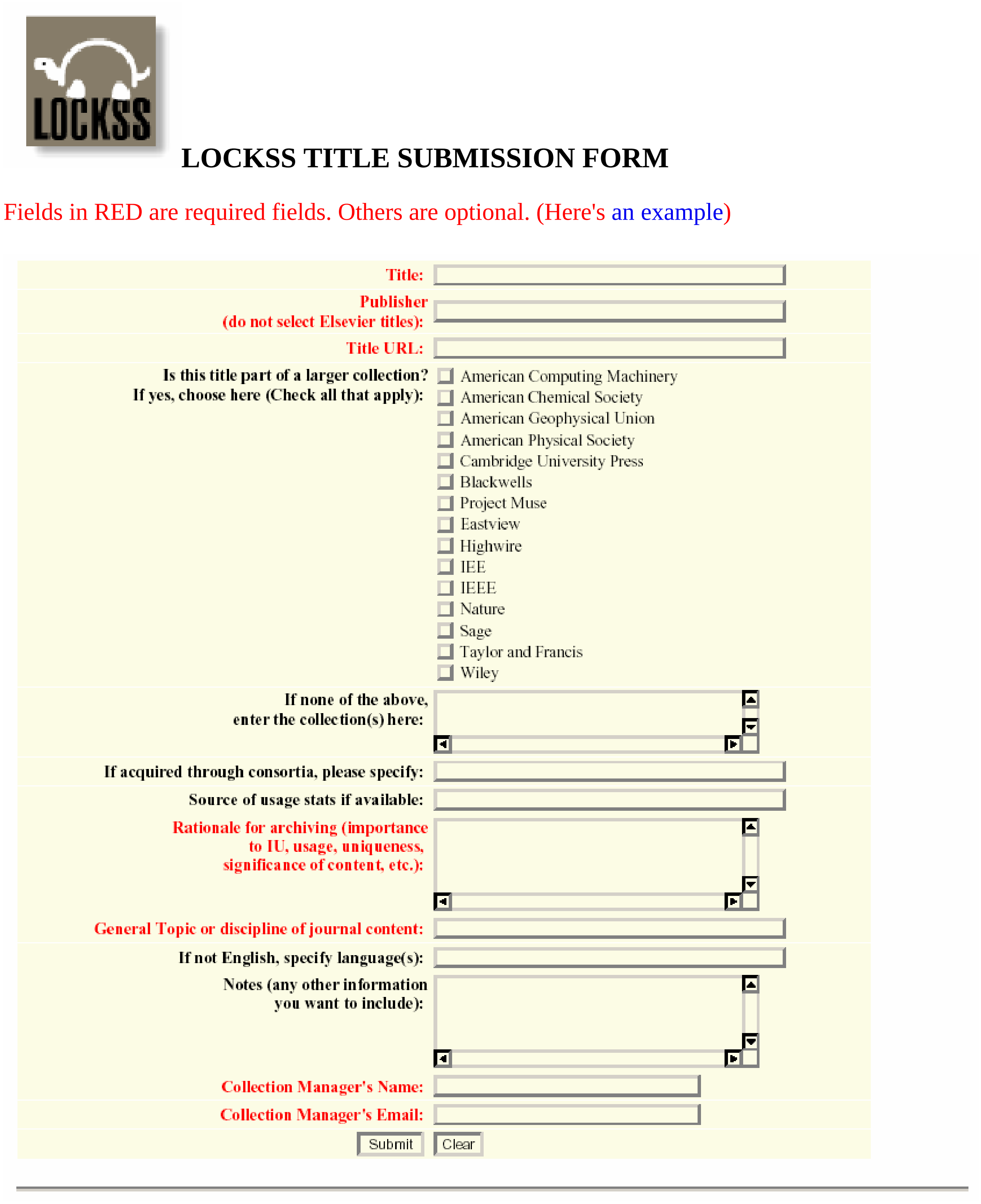
B. The Publisher Manifest will also provide an optional area for metadata associated with the content to be preserved for the long term. For example:



DRAFT elements	info pkgs
display label	SIP/PM
Title	x
Publisher	x
Subject: keywords	x
Type: electronic journal	x
ID:	
ID: journal url	**X**
ID:issn	x
ID: doi	x
Language	x
Contact: email of publisher	x
Frequency	x
Date copyrighted	x
Rights	x
Archival unit rules	**X**

X optional, **X mandatory**

Data should be formatted as Dublin Core Meta Data Tags
<META NAME="DC.Language" CONTENT="en-uk">



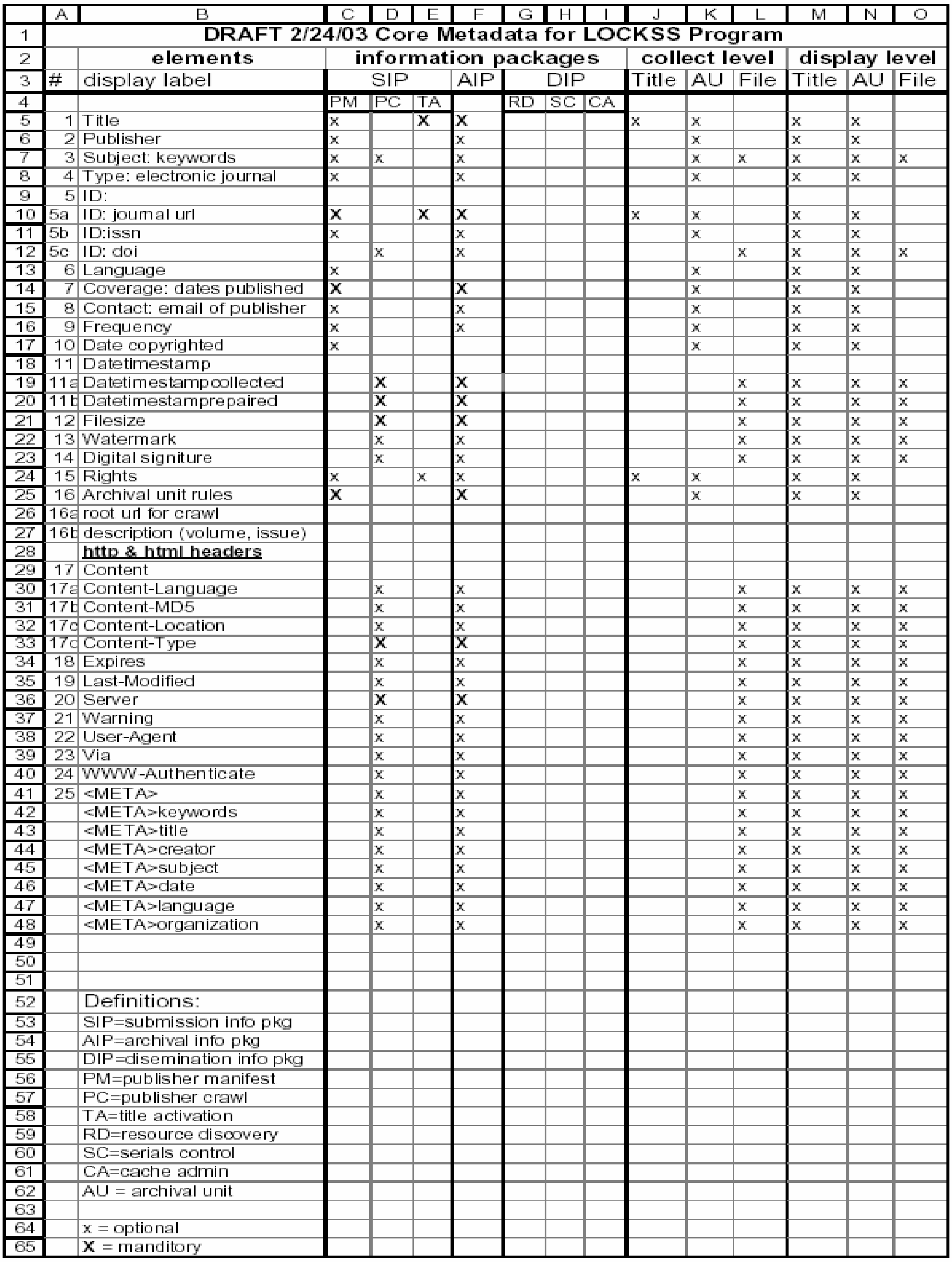
LOCKSS TITLE SUBMISSION FORM

Fields in RED are required fields. Others are optional. (Here's an example)

Title:	
Publisher (do not select Elsevier titles):	
Title URL:	
Is this title part of a larger collection? If yes, choose here (Check all that apply):	□ American Computing Machinery □ American Chemical Society □ American Geophysical Union □ American Physical Society □ Cambridge University Press □ Blackwells □ Project Muse □ Eastview □ Highwire □ IEE □ IEEE □ Nature □ Sage □ Taylor and Francis □ Wiley
If none of the above, enter the collection(s) here:	
If acquired through consortia, please specify:	
Source of usage stats if available:	
Rationale for archiving (importance to IU, usage, uniqueness, significance of content, etc.):	
General Topic or discipline of journal content:	
If not English, specify language(s):	
Notes (any other information you want to include):	
Collection Manager's Name:	
Collection Manager's Email:	
Submit	Clear

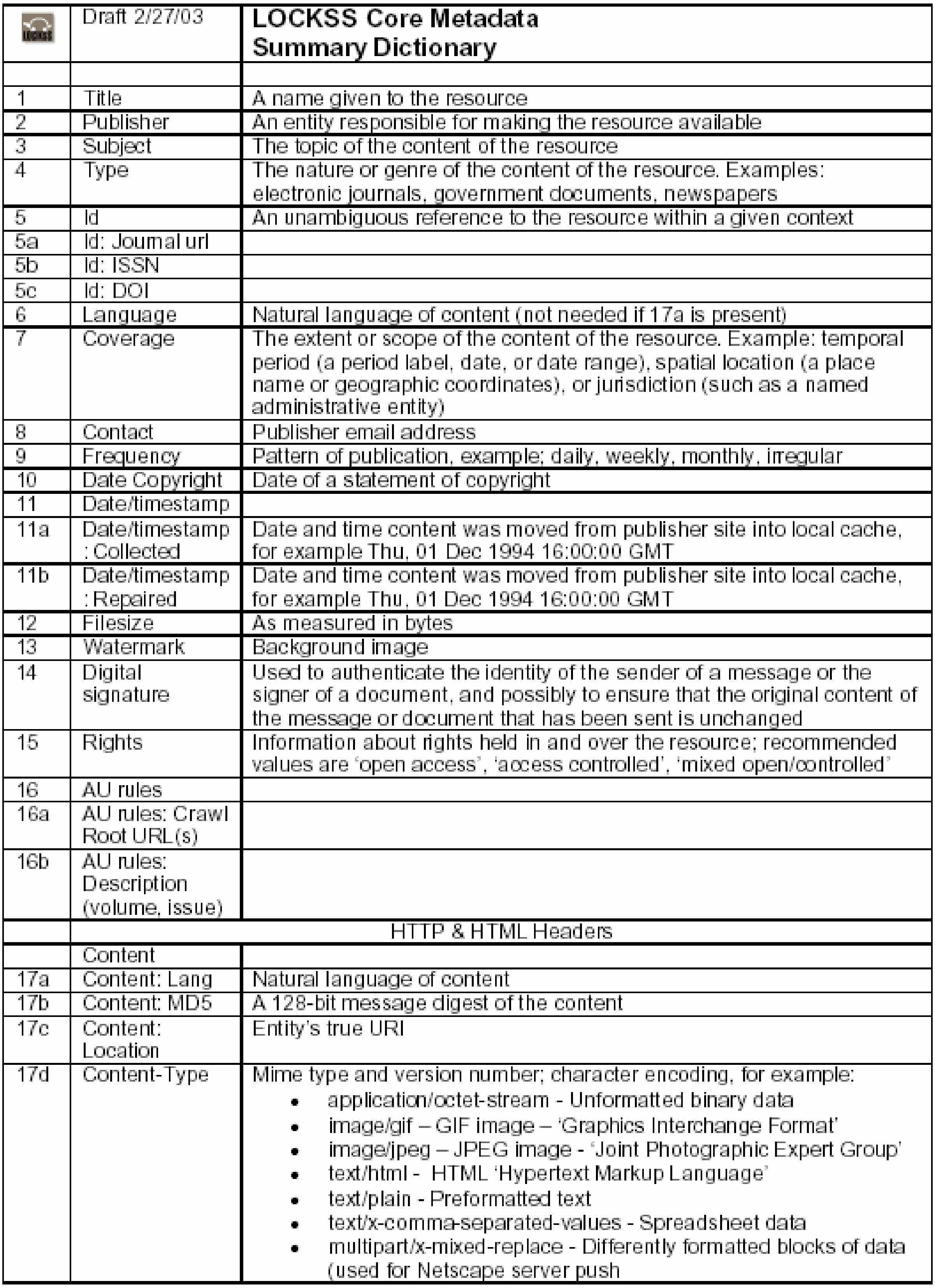
Relevant links
Indiana University Libraries Partner in Archiving Project To Ensure Long-term Access to E-Journals
LOCKSS site

Core Metadata for LOCKSS Program

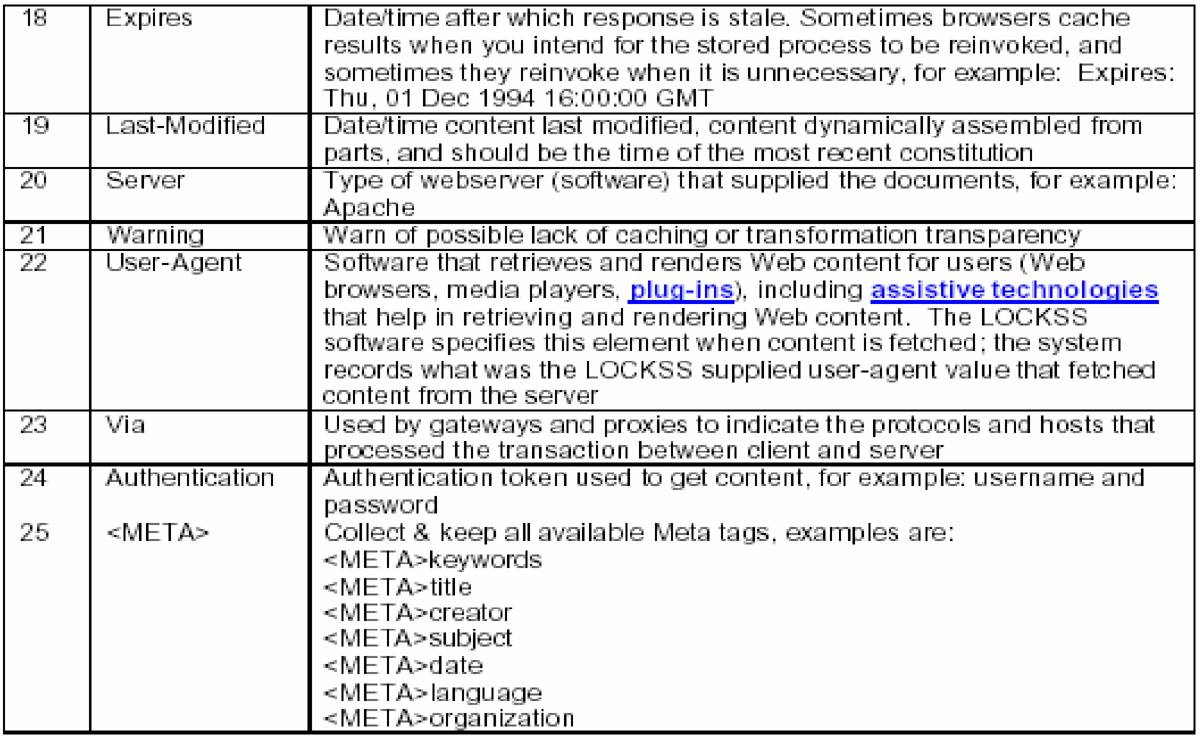


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	DRAFT 2/24/03 Core Metadata for LOCKSS Program														
2		elements	information packages							collect level			display level		
3	#	display label	SIP			AIP	DIP			Title	AU	File	Title	AU	File
4			PM	PC	TA		RD	SC	CA						
5	1	Title	x		X	X				x	x		x	x	
6	2	Publisher	x			x					x		x	x	
7	3	Subject: keywords	x	x		x					x	x	x	x	x
8	4	Type: electronic journal	x			x					x		x	x	
9	5	ID:													
10	5a	ID: journal url	X		X	X				x	x		x	x	
11	5b	ID:issn	x			x					x		x	x	
12	5c	ID: doi		x		x						x	x	x	x
13	6	Language	x								x		x	x	
14	7	Coverage: dates published	X			X					x		x	x	
15	8	Contact: email of publisher	x			x					x		x	x	
16	9	Frequency	x			x					x		x	x	
17	10	Date copyrighted	x								x		x	x	
18	11	Datetimestamp													
19	11a	Datetimestampcollected		X		X						x	x	x	x
20	11b	Datetimestampprepaired		X		X						x	x	x	x
21	12	Filesize		X		X						x	x	x	x
22	13	Watermark		x		x						x	x	x	x
23	14	Digital signiture		x		x						x	x	x	x
24	15	Rights	x		x	x				x	x		x	x	
25	16	Archival unit rules	X			X					x		x	x	
26	16a	root url for crawl													
27	16b	description (volume, issue)													
28		**http & html headers**													
29	17	Content													
30	17a	Content-Language		x		x						x	x	x	x
31	17b	Content-MD5		x		x						x	x	x	x
32	17c	Content-Location		x		x						x	x	x	x
33	17c	Content-Type		X		X						x	x	x	x
34	18	Expires		x		x						x	x	x	x
35	19	Last-Modified		x		x						x	x	x	x
36	20	Server		X		X						x	x	x	x
37	21	Warning		x		x						x	x	x	x
38	22	User-Agent		x		x						x	x	x	x
39	23	Via		x		x						x	x	x	x
40	24	WWW-Authenticate		x		x						x	x	x	x
41	25	<META>		x		x						x	x	x	x
42		<META>keywords		x		x						x	x	x	x
43		<META>title		x		x						x	x	x	x
44		<META>creator		x		x						x	x	x	x
45		<META>subject		x		x						x	x	x	x
46		<META>date		x		x						x	x	x	x
47		<META>language		x		x						x	x	x	x
48		<META>organization		x		x						x	x	x	x
49															
50															
51															
52		Definitions:													
53		SIP=submission info pkg													
54		AIP=archival info pkg													
55		DIP=disemination info pkg													
56		PM=publisher manifest													
57		PC=publisher crawl													
58		TA=title activation													
59		RD=resource discovery													
60		SC=serials control													
61		CA=cache admin													
62		AU = archival unit													
63															
64		x = optional													
65		**X** = manditory													

LOCKSS Core Metadata – Summary Dictionary



	Draft 2/27/03	**LOCKSS Core Metadata** **Summary Dictionary**
1	Title	A name given to the resource
2	Publisher	An entity responsible for making the resource available
3	Subject	The topic of the content of the resource
4	Type	The nature or genre of the content of the resource. Examples: electronic journals, government documents, newspapers
5	Id	An unambiguous reference to the resource within a given context
5a	Id: Journal url	
5b	Id: ISSN	
5c	Id: DOI	
6	Language	Natural language of content (not needed if 17a is present)
7	Coverage	The extent or scope of the content of the resource. Example: temporal period (a period label, date, or date range), spatial location (a place name or geographic coordinates), or jurisdiction (such as a named administrative entity)
8	Contact	Publisher email address
9	Frequency	Pattern of publication, example; daily, weekly, monthly, irregular
10	Date Copyright	Date of a statement of copyright
11	Date/timestamp	
11a	Date/timestamp : Collected	Date and time content was moved from publisher site into local cache, for example Thu, 01 Dec 1994 16:00:00 GMT
11b	Date/timestamp : Repaired	Date and time content was moved from publisher site into local cache, for example Thu, 01 Dec 1994 16:00:00 GMT
12	Filesize	As measured in bytes
13	Watermark	Background image
14	Digital signature	Used to authenticate the identity of the sender of a message or the signer of a document, and possibly to ensure that the original content of the message or document that has been sent is unchanged
15	Rights	Information about rights held in and over the resource; recommended values are 'open access', 'access controlled', 'mixed open/controlled'
16	AU rules	
16a	AU rules: Crawl Root URL(s)	
16b	AU rules: Description (volume, issue)	
	HTTP & HTML Headers	
	Content	
17a	Content: Lang	Natural language of content
17b	Content: MD5	A 128-bit message digest of the content
17c	Content: Location	Entity's true URI
17d	Content-Type	Mime type and version number; character encoding, for example: • application/octet-stream - Unformatted binary data • image/gif – GIF image – 'Graphics Interchange Format' • image/jpeg – JPEG image - 'Joint Photographic Expert Group' • text/html - HTML 'Hypertext Markup Language' • text/plain - Preformatted text • text/x-comma-separated-values - Spreadsheet data • multipart/x-mixed-replace - Differently formatted blocks of data (used for Netscape server push



18	Expires	Date/time after which response is stale. Sometimes browsers cache results when you intend for the stored process to be reinvoked, and sometimes they reinvoke when it is unnecessary, for example: Expires: Thu, 01 Dec 1994 16:00:00 GMT
19	Last-Modified	Date/time content last modified, content dynamically assembled from parts, and should be the time of the most recent constitution
20	Server	Type of webserver (software) that supplied the documents, for example: Apache
21	Warning	Warn of possible lack of caching or transformation transparency
22	User-Agent	Software that retrieves and renders Web content for users (Web browsers, media players, plug-ins), including assistive technologies that help in retrieving and rendering Web content. The LOCKSS software specifies this element when content is fetched; the system records what was the LOCKSS supplied user-agent value that fetched content from the server
23	Via	Used by gateways and proxies to indicate the protocols and hosts that processed the transaction between client and server
24	Authentication	Authentication token used to get content, for example: username and password
25	<META>	Collect & keep all available Meta tags, examples are: <META>keywords <META>title <META>creator <META>subject <META>date <META>language <META>organization

Publisher’s Frequently Asked Questions

Branding and Display

- What will my journal look like to readers who retrieve my content from a LOCKSS cache? Will LOCKSS maintain my journal’s brand and image?

Many readers, especially non-US readers, already access Web journals through caches run by their ISPs. These caches normally use the technology on which LOCKSS is based. For example, all participants in the UK’s JANET academic network access Web journals published outside the UK via such a cache. As with all cached materials, the article content will look precisely as you designed.

- What about the ads?

Only some, or none, of a web sites ad will be stored in a LOCKSS cache. If the journal article contains a 'request' or a 'command' to fetch content from an external ad server AND that ad server is online - a recent ad will be fetched.

Retaining Subscribers

- Will LOCKSS decrease my ability to retain subscribers?

LOCKSS permits libraries to cache content they can access. If a library cancels a subscription and has not cached the content, they can not get access to that content in the future. If a library caches content and then cancels their subscription, they continue to have access to the content they cached.

User Statistics

- Will I loose contact with my end users? I want to retain access to reader usage data and have access to the record of the reader's interactions with my journal web site.

Because the reader is supplied preferentially from the publisher, with the cache only as a fallback, the publisher sees the same interactions they would have seen with any caching software. If a local reader access journal content from a LOCKSS cache and the publisher’s web site is available, LOCKSS will send "clickstream" data from the local caches back to the publisher's site. If, for any reason, the publisher’s site is not available the reader will see the content from the local cache. In this case the publisher will not see the clickstream data, but in the absence of LOCKSS there would have been no access and thus no clickstream.

- Will LOCKSS accesses to the journals show up in user reports?

If the publisher's system is up, and content is delivered from a LOCKSS cache, the publisher will get an "access report" from the LOCKSS cache. We recommend that libraries configure their LOCKSS caches to serve content from the publisher whenever the publisher is up and only serve content from the LOCKSS cache when the publisher's site is not available.

However in some cases it makes sense for an institution to use a LOCKSS cache as a proxy (for example where institutions pay for each bit of bandwidth used). In these cases, most requests for content will be served from the local LOCKSS cache. These uses will be reflected in the access statistics. The IP address of the LOCKSS cache would accurately reflect the” institutional subscription". We are working to also deliver to publishers’ details of individual user's computer IP addresses; for some of you it’s possible you would have more accurate user data than you do now!

- Can the LOCKSS system tell the publishers, for each journal, which institutions are using LOCKSS?

You should be able to tell which accesses are from a LOCKSS crawler; which accesses are from a LOCKSS cache and the original source of the request.

Content Integrity

- What happens if (when) the content on the publisher's site changes? Publishers sometimes need to change an article's content (replace an image; correct an author's name, etc.) Will these changes propagate into the LOCKSS caches and be saved for posterity?

We know publishers rarely change content, but that changes do occur. We've measured these changes in real journals and the probability of change to one issue falls off rapidly to very low values after a subsequent issue or two has been published. Changes past this point tends to be corrected though erratum. LOCKSS only collects content slowly; these early changes tend to happen before LOCKSS collects the content. Changes after that may not be collected in the initial pass. We have a mechanism to identify later changes and bring them to the attention of the system administrators, but testing it is not planned until a later stage of the beta test. We don't currently plan to handle this situation automatically; it is rare and easy to confuse with, say malicious damage to a publisher's site.

- What happens if a publisher removes content from their site? Does it automatically disappear from the LOCKSS caches as well?

In order for LOCKSS to function as an archive there must be no automated way to remove or alter data once collected and agreed upon. Any such change must require human intervention. An automated change or removal mechanism would provide precisely the tool the bad guys would need.

For example, in print, libellous information is never "with drawn". The publisher prints a correction. The publisher does not ask for the paper issue to be returned or destroyed. And even if they do, most librarians would not comply. It's probably not that uncommon for material in library to be libellous. The library is not responsible - the author and the publisher are. We would like the LOCKSS caches to work the same way.

If a LOCKSS cache ever obtains content from the publisher, from a subsequent crawl or as a repair, that doesn't match the majority's version, it will alert the human administrators of that cache. They can inspect the differences and decide what to do. In the absence of signed and time-stamped content from the publisher's website we have no way to determine the authenticity of a change.

Content Leaks and System Security

See http://lockss.stanford.edu/lockssecurity.html for a technical discussion.

- Does LOCKSS permit free and open access? Will LOCKSS allow journal content to be illegally replicated, or leaked, on a massive scale once copies are in the custody of others?

LOCKSS is not an "open access" system. If a library wants to preserve access to content they must establish a local cache. All publisher access control mechanisms are enforced. Content is not leaked to unauthorized users. LOCKSS does not provide Librarians with any additional or new methods or tools for turning on or authorizing whole new user communities.. One cache may supply content to repair damage to another cache, but only if the receiving cache is known to have had that content in the past. Because content is provided to other caches only to repair damage to content they previously held, no new leakage paths are introduced. LOCKSS will not allow Libraries to get protected material from sites that they are no longer subscribed to.

LOCKSS does not subvert access control, so once an institution has unsubscribed from a journal they cannot get any of the new content (unless the publisher later makes that content free). The only power that LOCKSS removes from the publisher is their ability to revoke the rights to back content.

- Can a publisher participate in LOCKSS and thereby "audit" the LOCKSS system?

Yes, publishers can run LOCKSS caches for their own journals and, by doing so, over time they could 'audit' other caches holding these titles. Any cache with ill gotten content runs the risk of being caught in such an audit, because the damage detection and repair protocol requires caches to announce their holdings to other caches. Let's say there is a non-subscriber cache on the LOCKSS system. It would eventually reveal itself by taking part in the protocol. The mere possibility of detection should deter non-subscribers from taking part in LOCKSS. There is no way for caches to be sure none of the other caches belongs to the publisher.

- What would stop a subscribing institution from sharing the cached content with people who shouldn't have it?

Any subscriber who wants to violate copyright by mass sharing of journal content can easily do so now; they don't need LOCKSS. LOCKSS doesn't provide any new opportunities for copyright violations. A loose analogy is LOCKSS caches are to electronic publishers as paper copies are to paper publishers. Both LOCKSS caches and paper copies of a journal provide access, but they cannot be easily used for repurposing (doing new things with the content).

- Does LOCKSS software in itself have access control functionality?

LOCKSS caches use and are dependent upon the publisher's access control mechanism. So far, we have only implemented IP address access control, since that is used by the vast majority of journals. Other mechanisms could be implemented.

Increasing the stringency and therefore the obtrusiveness of access control is not likely to solve the abuses problem. When access to online journals is controlled by institutional subscriptions, the publishers are dependent on these subscribing institutions to delegate the subscription to individuals at that institution. They are dependant on those individuals to prevent abuse of whatever authentication token is supplied to them. LOCKSS is in the same position as the publisher in this regard - a cache has to depend on its administrator (i.e. the subscribing institution) to tell it which authentication tokens are valid for which content.

As we move toward production, we've begun to talk a few librarians about integrating LOCKSS into campus networks. The caches will be configured to serve content to the local community in the event that the publisher's site goes down. All configurations involve all or some portion of campus network traffic being directed through the LOCKSS caches. The LOCKSS caches could provide a central monitoring point and potentially limit access to the journal content they are caching. The caches keep logs of all content access that goes through them. These logs could be mined to see if there are anomalies that might signal automated content theft. It would eventually be possible to have the caches block offending IP addresses or limit the rate at which anyone can get content through them. There is a pretty stable pattern to the access of e-journal content by a human. Deviations from this pattern could certainly be flagged and/or the offending party blocked.

- Since the software is open source can potential hackers exploit weaknesses relatively easily, at least more easily than systems where the source code is unavailable?

We believe that open source software is more secure, because it is much easier for a large number of people to audit than software. Bugs are found more quickly, as are security holes. Propriety software is very difficult to audit. Open source software cannot rely on "security by obscurity" or being secure simply by virtue of no one understanding how a system works.

LOCKSS is engineered for the long term - the system has to be open, transparent, and understandable. People who are concerned about the security of the LOCKSS system are welcome to examine the source code and make their own decisions.

DOIs

- How will the LOCKSS software work with publisher DOIs?

Many journals provide DOIs for their articles. These are the only names for the articles that they commit to be stable. We expect the DOIs will play two roles in LOCKSS: If and when the publisher restructures their web sites, the DOIs will be used to connect the old and new locations of articles, preventing the relocated old content being treated as new. In addition to providing the ability to find articles in the local cache by keyword search, we will support finding it via DOI, even after access to the original publisher ceases.

As LOCKSS collects content via the current URL, it will also preserve the DOIs. If a restructuring is anticipated, the plug-in for the journal will know the mapping between old and new URLs for the same DOI, and the relocated old content will not be crawled. If it is not anticipated, the relocated old content will be crawled but the articles will be recognized as old because their DOIs are already known.

The exact treatment of relocated or republished content will be under the control of the plug-in for the journal in question. We expect that most plug-ins will attempt to preserve the most recent version of a given DOI that they encounter. Thus the old URLs would end up pointing to the latest version of the article named by the DOI.

<u>Preserving Access</u>

How is access to preserved content provided in the following scenarios?

- When an institution has not continued to renew its subscriptions and has only a dated copy (that does not contain the most recent content)?"

Such an institution would not be able to collect the republished content. Their cache would preserve the dated content by communicating with other caches that had dated content.

- When formats evolve, for example a move from HTML to XML/MathML, and the "republishing" of everything on a publisher's platform?

LOCKSS collects and preserves content in whatever format would have been supplied to a user's browser (e.g., HTML, XML, GIF, etc.). If the content is fetched from a LOCKSS cache, it will be supplied in the same format in which it was received.

For most purposes, LOCKSS doesn't care about the format. However, when hashing content in order to compare its copy with that of other caches, LOCKSS must filter out those parts of the content that may change each time it's fetched, for example, advertisements. If it didn't do this, it might appear that two caches disagree on the content, when the only difference is the ads that were supplied when the cache fetched its copy. This requires LOCKSS to understand enough about those content formats that might contain variable content in order to perform this filtering. As the exact details of the filtering are often journal-dependent, this function is performed by a journal-specific plug-in. We have written filters for HTML, and we expect that a filter for XML would be similar; this would have to be examined in the context of the specific journals.

- When published material is no longer available to the institution, for example because the subscription was cancelled?

When published material is no longer available to the institution, for example because the subscription was cancelled, the material that was obtained whilst the subscription was in effect remains in the local LOCKSS cache. Because the material is still there, the cache continues to be able to prove its right to have that material, and to obtain repairs to it when damage is detected. The material continues to be available to local readers as described above. All that happens is that newly published content is not available to readers at the cancelling institution, nor to the institution's LOCKSS caches.

- When the publisher changes the URL structure of the journal?

If the publisher changes the URL structure of the journal web site, LOCKSS will perceive this as two related events: 1. Material that used to be available under a set of URLs is no longer available, and if the reader requests it, it must be supplied from the cache. 2. New material has become available and must be preserved.

LOCKSS thus preserves the new and the old versions separately. Access via URL will return the specified version; access via full-text or bibliographic search will return both versions. With the right information LOCKSS system would be able to make the connection between the new and the old material. It could flag the old material as obsolete and present only the new material, for example by redirecting from the old to the new URLs.

- When some or all published material is no longer available to anyone from the publisher, for example because the publisher went bankrupt?

When some or all published material is no longer available to anyone from the publisher, for example because the publisher went bankrupt, the effect at each preserving institution is exactly the same as if the institution had cancelled its subscription. Old content remains available via the LOCKSS cache, which continues to be able to repair it from other LOCKSS caches of the same content. New content is not available because it is no longer being published.

- When browsers no longer interpret HTML, PDF, etc?

Over time each of the formats (HTML, PDF, JPEG, etc.) in which content is currently being published for the web will become obsolete and be replaced. Because of the immense volume of material that has been published in the web formats, we can predict some things about the process that will take place: 1. Tools will be available to convert the old to the new format. Unless they are, the new format will not succeed in the marketplace. 2. Browsers will continue to support both the old and the new format for an extended period of time, long enough for the remaining content published only in the old format to become unimportant.

During the overlap there will be time to upgrade the software running each of the LOCKSS caches to support the new format by either: 1. performing a one-time format conversion of the cached data. 2. performing on-the-fly format conversion of requested data, retaining the stored content in the old format but exporting the new format.

- When HTTP is no longer supported as a protocol?

Similarly, when HTTP is rendered obsolete as the transport protocol by which browsers and LOCKSS caches communicate with the publisher’s servers, there will be a long period of overlap. During this period servers will export content using both the old and the new transport protocol. During this time the LOCKSS caches can be upgraded to support both old and new transport protocols.

Our plans for both kinds of migration depend on the well-known adage that there can be “no flag days in the Internet”. The huge numbers of independently administered systems mean that there can be no rapid changes in the infrastructure. The beta test has established both the effectiveness of LOCKSS’ software distribution mechanism and its vulnerability to certain possible forms of subversion. It is routine to gradually upgrade the system on the fly, with various beta sites running various versions of the software. We plan significant improvements to the software distribution mechanism to make it more robust and less vulnerable, but we remain confident that it will work in timescales of weeks as against the timescales of many years for protocols and formats to become obsolete.

The LOCKSS protocol itself has been engineered for evolution, and this evolution has been tested. Messages sent between caches carry protocol version numbers. Caches understand multiple versions of the protocol, and can be configured to send messages in old or new formats. We have changed protocol versions without interrupting the system. A version of the daemon is gradually rolled out that understands both an old and a new version of the protocol. Initially it is configured to send only the old version. When all the caches are able to understand both versions, a configuration change is gradually rolled out that switches the cache to send the new version. Eventually, all messages in the system are in the new format. At that stage, it would be possible to remove support for the old version. This is a consideration that will be explored and resolved over the course of time. It is not expected to arise during the proposed phase of the project.