

First Monday, Volume 13 Number 1 - 7 January 2008

[HOME](#) [ABOUT](#) [LOG IN](#) [REGISTER](#) [SEARCH](#) [CURRENT](#) [ARCHIVES](#)

[SUBMISSIONS](#)

[Home](#) > [Volume 13 Number 1 - 7 January 2008](#) > [Seadle](#)



PEER-REVIEWED JOURNAL ON THE INTERNET

**In archiving we trust: Results from a
workshop at Humboldt University in
Berlin**
by Michael Seadle and Elke Greifeneder

.....

Abstract

If 25 specialists in preserving scholarly information had sat together in June of 1907 at the University of Berlin on Unter den Linden (Humboldt-Universität zu Berlin), they could likely have agreed that materials stored in the libraries of one of the world's great research universities in the capitol of the richest and most powerful state in Europe could reasonably be trusted to survive long term. One hundred years later, after the events of the twentieth century had assaulted the collections with fire, water, looters, and censorship, representatives of four digital archiving systems came together to discuss the strengths and weaknesses of their systems face-to-face in front of an audience of librarians, who would have to choose whether any of these systems could be trusted to overcome the unknown events of the twenty-first century. A key conclusion was the need for interoperability and to pool efforts. An alternative to collaboration may be to let archiving systems compete on price, performance and advertising, but then as customers in that market, libraries need to think about how we can test long-term archiving, so that we have real evidence to decide whether the claims of reliability make sense.

Contents

[Introduction](#)

[A portrait of the participants](#)

[Key questions for long-term archiving](#)

[Issues of trust](#)

[Issues of testing](#)

[Next steps](#)

[Interoperability testing: Testbed data](#)

[Interoperability testing: Goals](#)

[Conclusion](#)

Introduction

If 25 specialists in preserving scholarly information had sat together in June of 1907 in the palace-like main building of the University of Berlin on Unter den Linden, they could likely have agreed that materials stored in the libraries of one of the world's great research universities in the capitol of the richest and most powerful state in Europe could reasonably be trusted to survive long-term.

One hundred years later the twenty-five experts that the Deutsche Forschungsgemeinschaft (German Research Society or DFG) invited to discuss long-term digital archiving at the same university in the same building on Unter den Linden knew well how events of the twentieth century had assaulted the collections with fire, water, looters, and censorship. Trust is a troublesome concept when projected across long stretches of time.

Today a number of long-term digital archiving systems ask libraries and publishers to trust them to preserve their scholarly materials for the distant future. Many of these systems come with impressive credentials, including partnerships with corporations, foundations, government agencies, and research universities. They seem as safe as the University of

Berlin must have seemed one hundred years earlier. If we in libraries have learned something from the twentieth century, perhaps it is a certain wariness about how much to trust.

This article reports on the discussions at and surrounding the "Workshop on preservation networks and technologies" that took place 11–12 June 2007. The proposal for this workshop drew heavily on Seadle's (2006) article from a March 2005 workshop in Ann Arbor, Michigan, where Seadle represented LOCKSS and Portico took part as well. The authors of the present article are the Workshop's moderator and official note-taker. While every attempt has been made to present the discussion fairly and accurately, readers should realize that no such report can accurately reflect every opinion of every participant. We have chosen not merely to report on the official discussions that took place with all participants sitting around the meeting room table, but many of the side conversations at breaks, or meals, or discussions that continued after the official end. It is our belief that these are an integral and often the most valuable part of such meetings.

The structure of this article follows roughly the structure of the Workshop. We begin with a portrait of the participants and of the intellectual climate for the key discussion questions. The sections that follow recap two major areas of discussion. The first was trust with topics ranging from readability to commercial software. The second was testing, which all agreed was important but hard to do. The final three sections before the conclusion look at plans for interoperability testing that represent one of the major outcomes from the Workshop.



A portrait of the participants

The DFG invited four long-term archiving systems to the meeting to discuss technical solutions and potential collaborations. The systems were (in alphabetical order):

- KB e-Depot from the Koninklijke Bibliotheek, National Library of the Netherlands
- kopal from the German National Library and Göttingen University
- LOCKSS (Lots of Copies Keep Stuff Safe) from Stanford University
- Portico, from Ithaka/JSTOR

Each of these systems has received substantial financial support and recognition from their national governments and in the case of LOCKSS from both the U.S. and U.K. governments. They are by no means the only long-term archiving systems, but they certainly count internationally as four of the most serious attempts at solving the problems. These systems embody different assumptions about how to approach long-term archiving. The intent of the meeting was to look beneath the rhetoric of the standard presentations to try to understand how their technology works and where their strengths lie. Knowing their origins and business models helped to explain the context in which they made their technology decisions.

LOCKSS is the oldest of these systems. Its initial external funding came from the U.S. National Science Foundation in 1999 as a special projects award under the Digital Library Initiative program. The beta version ran at 50 libraries throughout the world between 2000 and 2002, and the system went into production in 2004. Major funding came from a wide variety of sources, including the Andrew W. Mellon Foundation, Library of Congress (through the National Digital Information Infrastructure Preservation Program or NDIIPP), Sun Microsystems, HP Labs, Intel Research Berkeley, the computer science departments at both Harvard and Stanford, Stanford University Library, and now over 100 libraries who have become dues-paying members of the LOCKSS Alliance (the largest libraries pay US\$10,800 per year). LOCKSS is completely open source and has a long-term plan that relies on community-based development support on the model of Linux (LOCKSS, 2007).

KB e-Depot from the Koninklijke Bibliotheek began shortly after LOCKSS with a "call for tender" in 1999 that resulted in a contract with IBM to develop an OAIS-compliant system that has been in operation since 2003. The specifications came from a collaborative project of European national libraries called NEDLIB that developed a series of standards and guidelines. (Koninklijke Bibliotheek, 2007; Nieuwenburg, 2001) Funding for KB e-Depot comes from the Dutch Ministry of Education, Culture and Science via the Koninklijke Bibliotheek. Since IBM retains the rights to the DIAS (Digital Information Archiving System), the core of the KB e-Depot system is not open source.

Portico claims 2002 as its date of origin, though could also reasonably claim a longer history as an outgrowth of JSTOR. The Andrew W. Mellon Foundation, Ithaka, Library of Congress (also via NDIIPP), and JSTOR provided initial financial support for Portico, which also relies on fees from about 40 publishers and 369 libraries (Portico, 2007). The fees for libraries run up to US\$24,000 per year for the largest libraries. While Portico makes some use of commercial software, including the Oracle database and Documentum, it also has a strong commitment to open source tools. For normalizing a document, for example, it uses the Journal Archiving and Interchange Document Type Definition created by the National Center for Biotechnology Information of the National Library of Medicine (NLM) (Kirchhoff and Fenton, 2006).

kopal is the newest of the systems and is only just making the transition from being a development project to production system. Financial support so far has come from the German Federal Ministry for Education and Research (Bundesministerium für Bildung und Forschung or BMBF), but future support will likely need to come from other sources, presumably some combination of partner libraries. kopal's technical infrastructure resembles KB e-Depot, with which it cooperates closely. kopal uses the IBM DIAS software extensively, but has also build its own open source koLibRI (kopal Library for Retrieval and Ingest) software as an interface (kopal, 2007a). For future financing it offers three usage models: participant, client, and operator with a licensing fee of €96,000 to €385,000 and annual costs for the clients running up to €200,000.

More information about these systems can be found in "E-Journal Archiving Metes and Bounds: A Survey of the Landscape" (Kenney, *et al.*, 2006).

The definition of a long-term archiving system was not discussed at the Workshop but in subsequent discussions it became clear that the participants did not necessarily have a shared concept. The authors of this article recommend a pragmatic definition in which any system with broad recognition and broad financial support as a long-term archiving system is accepted as one. In a scholarly community definitions based on particular techniques or methods require test-based evidence before they can be accepted, and the testing for long-term archiving systems is only beginning.

All systems comment on each other occasionally, mainly verbally though occasionally in print [1]. As a practical matter, competition for customers exists among the systems, each of which have staff who depend on continuing funding. This competition makes collaboration harder, even though collaboration benefits the whole of the library world in the long run. While LOCKSS, Portico, and KB e-Depot have often appeared together at conferences, this is the first time that the joint appearance included kopal. The fact that leaders from these systems sat together in a plain-spoken discussion was a tribute to their ability to look beyond their own short-term interests.

Key questions for long-term archiving

In his role as moderator, Seadle posed the four following questions, which had sent to the presenters a week before the meeting. Will your system offer a reasonable probability that in 100 years a copy will be: a) available (*i.e.*, exist), b) unchanged (*i.e.*, have integrity)? c) be what it claims to be (*i.e.*, have authenticity) and d) be able to be read (*i.e.*, readability)?

He also asked four research questions during the discussion: Do we have evidence for the research community that our systems work as promised? Can we build tools that will test our systems? Are there areas where we can share techniques, research, and resources? And are there common tools that we need to research and develop?

The first set of questions especially grew out of a library literature that goes back at least as far as the 1992 and 1993 reports by Anne Kenney and Lynne Personius (1992; 1993) on the CLASS (College Library Access and Storage System) project that (among other things) considered whether digital images could substitute for microfilm. The Council on Library and Information Resources as successor to the Commission on Preservation and Access has taken a lead in publishing important texts, including works such as *Authenticity in a digital environment* (Council on Library and Information Resources, 2000) in which Peter Hirtle, Clifford Lynch and others discuss key issues of authenticity, integrity, and trust. Another major source is *RLG DigiNews*, which has published articles on digitization and preservation for the past 11 years. Nancy McGovern's (2007) article on "A Digital Decade: Where Have We Been and Where Are We Going in Digital Preservation?" offers a good summary of the state of the discourse.

Since the Workshop took place in Germany, the German-language literature on long term archiving shaped the expectations of the German participants. The NESTOR (Network of Expertise in Long-Term Storage of Digital Resources) Web site maintains a list of relevant German and European articles (NESTOR, 2007). An important article by Ute Schwens, Director of the German National Library in Frankfurt, and Hans Liegmann helps to understand the German emphasis on readability:

Substanzerhaltung ist nur eine der Voraussetzungen, um die Verfügbarkeit und Benutzbarkeit digitaler Ressourcen in Zukunft zu gewährleisten. "Erhaltung der Benutzbarkeit" digitaler Ressourcen ist eine um ein Vielfaches komplexere Aufgabenstellung als die Erhaltung der Datensubstanz. Folgen wir dem Szenario eines "Depotsystems für digitale Objekte", in dem Datenströme sicher gespeichert und über die Veränderungen der technischen Umgebung hinweg aufbewahrt werden,

so steht der Benutzer/die Benutzerin der Zukunft gleichwohl vor einem Problem. Er oder sie ist ohne weitere Unterstützung nicht in der Lage, den archivierten Datenstrom zu interpretieren, da die erforderlichen technischen Nutzungsumgebungen (Betriebssysteme, Anwendungsprogramme) längst nicht mehr verfügbar sind. (Schwens and Liegmann, 2004, p. 2)

Preserving the substance is only one requirement to ensure the access and usability of digital resources in the future. "Preserving the usability" of digital resources is a much more complex task than just the data. If we look at a scenario for "storing a digital object" in which the data stream is saved and protected from changes in the technical environment, then the future user will still have a problem. Without support he or she is not in a position to interpret the archived data stream, because the required technical usage environment (operating systems and applications programs) are no longer available. [our translation]

A good bit of research on migration and emulation exists, both in the library and the computing literature, where emulation has been used with varying success since the early 1960s with the advent of IBM System 360 architecture. Ensuring readability seems to be a particularly high priority within German discussions on digital archiving. In these discussions a key and perhaps defining issue is the ability to interpret a given bitstream using specific technical metadata that is stored apart from the digital objects. For them mere data integrity within a sustainable storage system leaves out a key part of the definition of long-term archiving.

kopal uses migration via its open source koLibRi software to ensure readability software and does not do migration at ingest. In the long-term kopal anticipates using migration primarily for static documents and emulation for dynamic works. Portico also emphasizes migration using the JAI DTD. KB e-Depot is doing research on both migration and emulation, which, like kopal, they see as important for different kinds of works. LOCKSS has done a system-wide test of format migration (Rosenthal, *et al.*, 2005) and, like KB e-Depot, recognizes that emulation may well be necessary for the future usability of interactive works like computer games. LOCKSS believes that migration during ingest adds significantly to costs with no guarantee that the chosen formats will solve future problems.

The second issue Schwens and Liegmann (2004) discuss is the establishment of trusted repositories:

Es werden deshalb Anstrengungen unternommen, allgemein akzeptierte Leistungskriterien für vertrauenswürdige digitale Archive aufzustellen, die bis zur Entwicklung eines Zertifizierungsprogramms reichen. Die Konformität zum OAIS-Referenzmodell spielt dabei ebenso eine wichtige Rolle wie die Beständigkeit der institutionellen Struktur, von der das Archiv betrieben wird. (Schwens and Liegmann, 2004, p. 3)

Efforts are therefore being undertaken to establish generally accepted performance criteria for trusted digital archives through the development of a certification program. Conformity to the OAIS-Reference model plays as large a role as the reliability of the institutional structure that manages the archive. [our translation]

This goal to establish trusted repositories for managing the long-term archiving process is not unique to Germany. The Center for Research Libraries in Chicago is doing similar work on repository certification.

The discussion in Berlin can be summarized in terms of three topics: trust, testing, and next steps. The attempt here is not merely to summarize the official discussions of the formal meetings, but the even more important dialogues that occurred on breaks and at meals.



Issues of trust

Trust is the one inescapable issue in archiving of any sort. Archiving materials for the next 100 years represents a bet on a future well beyond our own lifetime. Much of the current digital archiving effort stems from a lack of trust that publishers will archive materials in a way that will make them accessible to future scholars. The reasons for the distrust come partly from the fact that commercial publishers lack a strong history of continuity. They buy one another or go out of business, especially the smaller firms. Even those that survive generally do not manage their inventory from 100 years ago well and they have little economic incentive to do so.

In contrast libraries and universities are among the most stable of institutions, but they too have vulnerabilities. A number of trust issues arose during the discussion.

Certification

The certification of libraries and other repositories is popular in both the U.S. and Germany because it offers a measure of trust. Certification was discussed extensively at the Frankfurt conference "The Challenge: Long-Term Preservation — Strategies and Practices of European Partnerships" (DNB, 2007), which some of the Workshop participants attended. Any repository for long-term digital archiving should ideally meet certain auditable standards, but audits are far from foolproof, even in the business world where failing an audit could have fatal financial consequences. A significant scholarly literature exists about the flaws in even the most professional audit checklists and operations. Gendron (2004) for example, raises questions about whether "generic" checklist questions are necessarily the relevant ones. The flaws in a system may be far from ordinary and obvious. Trusted employees can, for example, become a risk: any type of employee of any age might in fact cause accidental or deliberate damage (Keeny, 2005). Even a newly certified repository can have a hidden flaw that betrays its trust and risks its contents.

Funding

Business models matter in any trust calculation. An archiving system that goes out of business could hardly be called reliable. At best bankruptcy freezes development and forces customers to shift platforms. At worst it puts the content at risk. Today's funding models for long term archiving vary widely. Funding for kopal and e-Depot comes at present from the national government. At one time a central funding model might have seemed sensible and stable, but even the richest governments balk at spending heavily to archive knowledge from other countries, even when future scholars need that information. In times of fiscal pressure the political will to fund archiving may weaken. LOCKSS and Portico have focused their business models on the specific problem of scholarly journals, although both have the ability to archive a much broader range of materials. Their business models involve a combination of library and publisher support with implicit assumptions about decreasing government funding. This shared funding responsibility shares the risk if any single source changes its priorities. It also spreads the risk and responsibility across national boundaries. Its problem is that maintaining the broad funding base costs time and effort and the explicit competition for paying supporters (*i.e.*, customers) fractures a limited financial base.

Readability

All participants agreed that archived files must exist to be read and that readability over time matters, but they differ philosophically about how and when to trust that a file would be readable. Portico migrates files to standardized formats as part of the archiving process in the hope that these migrations will make future readability more likely. LOCKSS puts its emphasis on maintaining the integrity of a file on the principle that the existence of an undamaged original is a prerequisite to future readability, and does not try to anticipate future standards or future tools for reading. It has, however, tested its ability to do on-the-fly format migration and published the results of the experiment. kopal's developers are openly skeptical about trusting readability to future tools and it puts substantial efforts into testing and ensuring readability on the principle that an unreadable file is not useful. LOCKSS, kopal and e-Depot treat all emulation and migration as possible models for future readability.

If future readability is the chief criterion for long-term archiving, then it is unsettling to trust in the ability of future systems to decode obsolete file formats. Standardization is highly attractive and all systems agree that it is important, but our record in predicting standards 100 years in the future is not encouraging in areas of rapid technological change, and the library community has little power to set standards outside of its own small sphere.

Commercial software

The extent to which archiving systems should rely on commercial software also divided the systems. Both e-Depot and kopal rely on IBM's DIAS software for the archiving and file maintenance. Portico uses some commercial software, including Documentum and the Oracle database, but minimizes dependence on it. LOCKSS is entirely open source.

Participants agreed that neither commercial nor open source software has a quality

advantage in terms of the code. IBM has also done its best to provide information about how its DIAS system works internally, so that it is not a commercial black box. The differences lie in choices about development versus licensing costs, system support, and long-term commitment. By using DIAS, kopal and e-Depot have paid license fees and continue to pay ongoing maintenance costs (which in theory could be stopped). Short-term this could be a reasonable deal. The choice may seem less attractive after 100 years of fees, but for now they avoid the substantial up-front costs for system design and programming expertise, and the ongoing in-house support costs of open source software. IBM has an excellent reputation for software support. The key issue with commercial software is long-term commitment: if IBM were to decide that archiving were not a commercially worthwhile product line to maintain, could the data safely migrate elsewhere? The same could be asked, of course, of open source software, if the community supporting it lost interest, but ideally we are that community.

Issues of testing

One of the most active and creative discussions during the Workshop took place when people were asked what they wished we could all collaborate on. These were often areas where participants such as Tobias Steinke from kopal and Evan Owens from Portico called for making testing part of the research agenda. David Rosenthal from LOCKSS, who also strongly favored testing, cautioned about the gap between the extent to which we can actually test our systems and the requirements people placing on them.

The library community has not had a strong experimental culture. The focus tends to be a businesslike search for solutions to specific problems and a tendency to buy rather than build. The community has a lot of experience with establishing standards and relatively little with creating and testing tools. That, however, has changed in recent years as library science and information science have drawn more on concepts and experiences from both computer science and engineering. Participants at the Workshop came from engineering and computing as well as from libraries, and all agreed about the value of testing.

Virtually every decision that the archiving systems have made involves some level of trust — essentially a bet on future outcomes. It became clear in the course of the discussions at the meeting that testing the various decisions about the best way to approach long-term archiving would help libraries and publishers decide what their options are, and would help the archiving systems themselves address potential weaknesses in their plans. Testing is, however, never easy. The discussion touched on the following points.

Routine testing

Some testing has already been done or goes on regularly. Portico offered a good example when Evan Owens described their testing on a PDF file that met all the standard measures for being well-formed and valid, but that blew up when opened because of a damaged font in the TeX file that was used to generate the PDF. KB e-Depot conducts routine random tests of documents within its system to be sure of their integrity. It also conducts tests as part of its versioning management. The LOCKSS servers perform ongoing checks of the content on other LOCKSS servers to ensure that nothing has changed.

Use is one of the best tests for archived materials. KB e-Depot makes materials in its archive available to anyone at the Library. This does not guarantee that every file will be looked at, but it provides a relatively random check on content. LOCKSS materials also can be used any time the original site goes down. While this happens relatively rarely, it does occur. One small publisher in fact lost its whole Web site and had to rely on the archived copy in LOCKSS to restore it. Portico has a dark archive, though a copy is constantly available for audit and verification viewing by participants. and kopal has no official content to expose yet, since it is still in the development phase.

Disaster testing

Disaster testing is a component of the NESTOR certification for trusted repositories, but actually carrying out a disaster test is time-consuming and disruptive, and very few disaster tests go well the first time. LOCKSS has been in operation long enough that it has faced some real crises, such as the SSH CRC-32 Compensation vulnerability in Linux that occurred early in LOCKSS testing. It was discovered on New Year's eve in 2001 and had to be patched at once to protect the LOCKSS servers from exposure. Portico's ability to recover from a disaster is enhanced by its relationship with JSTOR and its long experience with 24/7 content delivery operations, but formal disaster testing has not yet begun. KB e-Depot has done limited disaster testing and kopal is, as noted above, not yet in production.

Requiring testing

Certification came up in the discussion as one way to require testing, but requiring testing is only a first step. The challenge is to design tests that can expose the vulnerabilities of

long-term archiving systems. Serious testing is in the interests of every institution that wants reasonable assurance that archived materials will be around in 100 years, but testing is not necessarily in the interests of the business models of archiving systems, which must convince their customers and funding agencies that they offer the best, most cost-effective, most complete, and most especially the reliable method.

Potential unintended consequences

An archiving system that failed a test may well be able to address the problem quickly and reliably and be a better system as a result, but any system that appears to perform poorly in a public test risks losing the confidence of its financial supporters. And, ironically, any threat to the financial viability of an archiving system represents one of the most serious risks to long-term archiving. Even an open source system like LOCKSS needs to maintain funding for the development team until that work can be fully distributed among members of the community.

Next steps

One of the key conclusions from the discussion was the importance of interoperability testing. That kind of testing serves in some sense as an audit that tests the integrity of the archived materials. Interoperability is also important because it gives libraries and publishers the flexibility to change their choices if circumstances change. Libraries or publishers could, for example, begin with one system and later switch to another system that provided better services for submission and extraction, or that they trusted more. Interoperability in effect facilitates the market for archiving systems.

Agreement on a single system for all countries, including financial and technical support and broad policy agreement, would make interoperability unnecessary, but this seems politically and practically unlikely. In a market with multiple systems, maintaining choice through interoperability mitigates the effects of wrong choices over time.

The discussion did not explore either the value (or danger) of having an open market for long-term archiving systems. Librarians do not generally turn by choice to market-based solutions. The tendency is to establish standards and to seek systematic national or international solutions rather than to facilitate an environment in which competition is explicitly encouraged. The fact is, however, that a market for archiving systems exists. LOCKSS and Portico compete directly for customers (financial supporters) in the North American market, and both have a presence in multiple countries in Europe. IBM is also a strong market competitor. The market competition is by no means bad. It helps the community to learn how to deploy all available preservation resources and technologies to achieve the best and most cost-effective coverage.

In this market development a parallel exists with library automation systems. These systems began as local solutions but quickly looked for broader markets. A few, like the venerable MELVYL System for the University of California, remained local. Other systems, like Innovative Interfaces Inc. (III), began not as fully integrated library automation systems, but as a vendor that solved a particular problem exceptionally well and cost effectively, and only slowly expanded to incorporate other functions. Today III is one of the most successful full-system vendors. Long-term archiving systems seem poised to follow a similar path. The likelihood of a single solution prevailing nationally or internationally seems small.

In this market context interoperability is important, but it is far from easy to achieve. All agreed on the need for common standards, well-defined interfaces and good documentation. Standards help especially when developed within an experimental context that shows whether they work. The following sections of this article look at how interoperability might be implemented.

Interoperability testing: Testbed data

One of the first requirements for interoperability testing is a body of materials that may be used. This is far from trivial, since the copyright permissions that enable archiving systems to ingest the works do not automatically allow their ingest into another system. Getting permission from rights holders is a slow and expensive process and not practical on any large scale.

One potential solution to this problem is to test the interoperability of metadata. It was clear from the discussion that most systems store metadata such as JHOVE (JSTOR/Harvard Object Validation Environment at <http://hul.harvard.edu/jhove/>) to describe the digital

objects that they store and that they regard this metadata as important to their concept of the archiving process. This metadata is less sensitive to copyright issues and is more easily shared. Exchanging bibliographic metadata is more or less a routine job, but exchanging objects within a metadata "shell" in a dedicated package like METS is a challenge. Portico and KB e-Depot use relatively similar metadata and share a view of its importance. They are about to set up an archive preservation agreement with each other. LOCKSS takes a different view of the role of metadata and considers the look and feel of the historical context a critical piece of intellectual content (Reich, 2007).

Nonetheless exchanging metadata is not quite the same as testing whether the original digital objects can readily move from one system to another. The Library of Congress has been working on an exchange project involving government documents now in digital formats and used to be exchanged in paper. Conversations on this project began 2005 at a meeting in Frankfurt between representatives of the Library of Congress, LOCKSS, German National Library, Staatsbibliothek zu Berlin, University of Regensburg, Göttingen University, and Humboldt University in Berlin, which also represented DINI (Deutsche Initiative für Netzwerkinformation, the German equivalent of the Coalition for Networked Information). Concrete plans using LOCKSS were discussed, but implementation was delayed for unrelated reasons.

The Library of Congress has recently allocated funds to pursue this exchange project through the National Digital Information Infrastructure Preservation Program and is interested in using the project as a vehicle for the interoperability testing. While the U.S. government documents arguably have copyright protection outside of the U.S., and the German government documents definitely have copyright protection, partners on both sides believe the necessary permissions can be resolved easily. This means both that a testbed for interoperability testing exists and that partial funding is in place. LOCKSS and kopal are the logical partners for this interoperability testing, since representatives of both have been involved in the preliminary discussions. The project can go forward once funding for the German side of the work is in place.

Interoperability testing: Goals

Exchanging documents between LOCKSS and kopal could present special challenges, because of the very different basis and assumptions build into the two systems. The advantage is, that if they can establish principles for exchange, it should be relatively easy to expand the exchange to other systems, including some that are not part of the Workshop, such as DAITSS (Dark Archive in the Sunshine State) (FCLA, 2007), PANDORA (PANDAS) from the National Library of Australia (NLA, 2007), or the Internet Archive (2007).

The overall goal for interoperability testing should ideally be a generalizable mechanism that archiving systems can use to exchange files, much as automated library systems in North America have used the MARC Communication Format to exchange bibliographic data, including whole databases. But as anyone knows who has migrated a bibliographic database from one vendor to another, the process requires considerable testing and adjustment even with relatively standardized bibliographic data.

A number of potential exchange formats already exist and the temptation to define yet another would not simplify the problem. In principle all the systems can ingest any type of digital file, regardless of its format or structure. The exchange could take place using the original files and file formats, or something like the ARC format (Burner, 1996) that compresses the files and is used by the Internet Archive, or something like METS (Metadata Encoding and Transmission Standard) (Library of Congress, 2007) that kopal uses. The actual choice of formats will need to be worked out between partners.


Authenticity will need to be a goal of interoperability testing. This could be particularly difficult because it means establishing original standards of trust for the authenticity as well as standards of trust between systems. Some models for authenticity testing over time exist in the physical world, where, for example, documentary evidence that Rembrandt created a painting enhances its value significantly over mere stylistic attribution. Whether such models can be imitated will need to be part of any test.

Conclusion

This Workshop represented an important stage in collaboration. Representatives from these four archiving systems discussed the strengths and weaknesses of their systems face-to-face in front of an audience of librarians. The meeting also mixed German, British, Dutch, and American styles of discussion, with the predictable result that some felt the

discussion was a bit too aggressive and others felt frustrated by what they saw as an unwillingness to confront issues. It is very much to the credit of all the participants that these potential frustrations were essentially invisible during the meetings.

This kind of discussion needs to continue if the library community wants to pool the efforts of the array of archiving systems that have appeared in the last several decades since libraries grew aware of the problem of digital archiving. An alternative may be to let archiving systems compete on price, performance and advertising, but then as customers in that market, libraries need even more to think about how we can test long-term archiving, so that we have real evidence to decide whether the claims of reliability make sense.

The bets we make now on archiving systems are ones that our successors in 100 years need to live with. Can they trust us to make informed choices? 

About the authors

Michael Seadle is director of the Institute for Library and Information Sciences at Humboldt University in Berlin, Germany, where he holds the Krupp Professorship for Digital Libraries. He is also editor of *Library Hi Tech*. He served as convener and moderator of the "Workshop on Preservation networks and technologies".

Elke Greifeneder is a lecturer at the Institute for Library and Information Sciences at Humboldt University in Berlin, Germany, and is assistant editor of *Library Hi Tech*. She participated in the Workshop and served as the official note-taker.

Acknowledgements

The authors would like to thank Reinhard Altenhöner, Eileen Fenton, Erik Oltmans, Evan Owens, Victoria Reich, David Rosenthal, Chris Rusbridge, and Vivian Petras for their invaluable help with corrections, comments, and notes from the workshop.

Note

¹. E.g., kopal, 2007b, p. 2.

References

Mike Burner and Brewster Kahle, 1996, "ARC file format," at <http://www.archive.org/web/researcher/ArcFileFormat.php>, accessed 21 August 2007.

Council on Library and Information Resources (CLIR), 2000. *Authenticity in a digital environment*. Washington, D.C.: Council on Library and Information Resources, at <http://www.clir.org/pubs/reports/pub92/contents.html>, accessed 21 August 2007.

Deutsche Nationalbibliothek (DNB), 2007. "The challenge: Long-term preservation — Strategies and practices of European partnerships," (April) Frankfurt, at <http://www.langzeitarchivierung.de/eu2007/>, accessed 17 August 2007.

Florida Center for Library Automation (FCLA), 2007. "Welcome to DAITSS," at <http://daitss.fcla.edu/>, accessed 21 August 2007.

Internet Archive, 2007. "About the Internet Archive," at <http://www.archive.org/about/about.php>, accessed 21 August 2007.

Anne Kenney and Lynne Personius, 1993. "A testbed for advancing the role of digital technologies for preservation and access: Final report," Washington, D.C.: Commission on Preservation and Access, at <http://eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED403910>, accessed 21 August 2007.

Anne Kenney and Lynne Personius, 1992. "Joint study in digital preservation," Washington, D.C.: Council on Library and Information Resources, at <http://www.clir.org/pubs/reports/joint/index.html>, accessed 21 August 2007.

Anne R. Kenney, Richard Entlich, Peter B. Hirtle, Nancy Y. McGovern, and Ellie L. Buckley, 2006. "E-journal archiving metes and bounds: A survey of the landscape," at <http://www.clir.org/pubs/abstract/pub138abst.html>, accessed 21 August 2007.

Amy Kirchoff and Eileen Fenton, 2006. "Archiving electronic journals: An overview of Portico's approach," at <http://www.portico.org/news/papers.html>, accessed 19 August 2007.

Koninklijke Bibliotheek (KB), 2007. "The e-Depot: An introduction," at <http://www.kb.nl/dnp/e-depot/dm/inleiding-en.html>, accessed 21 August 2007.

kopal, 2007a. "About kopal," at <http://kopal.langzeitarchivierung.de/index.php.en>, accessed 21 August 2007.

kopal, 2007b. "kopal: Ein Service für die Langzeitarchivierung digitaler Informationen," at http://kopal.langzeitarchivierung.de/downloads/kopal_Services_2007.pdf, accessed 21 August 2007.

Library of Congress (LC), 2007. "METS Metadata Encoding & Transmission Standard," at <http://www.loc.gov/standards/mets/>, accessed 21 August 2007.

LOCKSS, 2007. "Lots of Copies Keep Stuff Safe: About LOCKSS," at http://www.lockss.org/lockss/About_LOCKSS, accessed 21 August 2007.

Nancy Y. McGovern, 2007. "A digital decade: Where have we been and where are we going in digital preservation?" *RLG DigiNews*, volume 11, number 1 (April), at http://www.rlg.org/en/page.php?Page_ID=21033#article3, accessed 21 August 2007.

National Library of Australia (NLA), 2007. "About PANDORA," at <http://pandora.nla.gov.au/about.html>, accessed 21 August 2007.

NESTOR, 2007. "Kompetenznetwork Langzeitarchivierung: Informationsdatenbank," at http://nestor.sub.uni-goettingen.de/nestor_on/browse.php?show=2, accessed 9 September 2007.

Betty Nieuwenburg, 2001. "Solving long term access for electronic publications," *D-Lib Magazine*, volume 7, number 11 (November), at <http://www.dlib.org/dlib/november01/11inbrief.html>, accessed 21 August 2007.

Portico, 2007. "About Portico," at <http://www.portico.org/about/>, accessed 21 August 2007.

Victoria Reich, 2007. Private e-mail message to Michael Seadle (22 August).

David S.H. Rosenthal, Thomas Lipkis, Thomas S. Robertson, and Seth Morabito, 2005. "Transparent format migration of preserved Web content," *D-Lib Magazine*, volume 11, number 1 (January), at <http://www.dlib.org/dlib/january05/rosenthal/01rosenthal.html>, accessed 21 August 2007.

Ute Schwens and Hans Liegmann, 2004. "Langzeitarchivierung digitaler ressourcen," In: *Grundlagen der praktischen Information und Dokumentation*. München: K.G. Saur, pp. 567–570, at: <http://www.langzeitarchivierung.de/downloads/texte/kss-b20.pdf>, accessed 3 August 2007.

Michael Seadle, 2006, "A social model for archiving digital serials: LOCKSS," *Serials Review*, volume 32, number 2 (June), pp. 73–77, and at http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6W63-4K7X49X-6&_user=964000&_coverDate=06%2F30%2F2006&_rdoc=1&_fmt=&_orig=search&_sort=d&view=c&_acct=C000049508&_version=1&_urlVersion=0&_userid=964000&md5=56f6970311bbf3dbb3187446df3117de, accessed 21 August 2007.

Editorial history

Paper received 10 September 2007; accepted 5 January 2008.

Copyright © 2008, *First Monday*.

Copyright © 2008, Michael Seadle and Elke Greifeneder.

In archiving we trust: Results from a workshop at Humboldt University in Berlin by Michael Seadle and Elke Greifeneder

First Monday, Volume 13 Number 1 - 7 January 2008

<http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/viewArticle/2089/1923>