OCLC RLG

Home: Publications: Newsletters: RLG DigiNews: Issue index: **Aug 15, 2006**

RLG DigiNews
RLG's online newsletter for digital imaging and preservation

Feature Article 2

Developing Electronic Records Capacity in the Small Collecting Repository: the Documenting Internet2 Project

Authors: Dharma Akmon - JSTOR (dakmon@jstor.org), Elisabeth Kaplan - University of Minnesota (kapla024@tc.umn.edu)

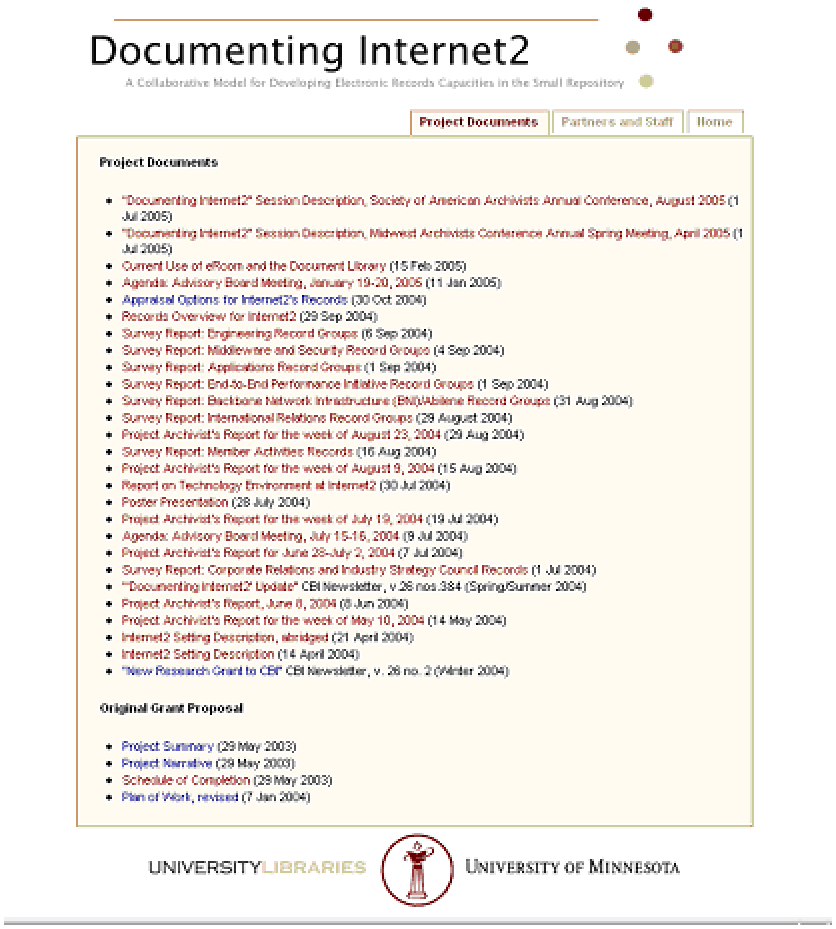
Introduction

What options are available to a small scale collecting repository when the core documentation in its primary subject area is no longer created in traditionally manageable formats? How well do traditional methods for appraising institutional records, which were developed in the context of stable, structured organizations, adapt to increasingly distributed, dynamic organizations whose records are primarily born-digital? For a collecting repository whose subject area is high technology, the problem feels particularly acute: the irony of trying to capture adequate documentation of developments in information technology in paper only is ever present.

The Project

These questions were at the core of a collaborative project, funded by the National Historical Publications and Records Commission and administered by the University of Minnesota's Charles Babbage Institute (CBI) Center for the History of Information Technology between 2003 and 2005.[1]

In this article, we describe a few of the methods, findings, and ideas for further exploration generated during "Documenting Internet2: A Collaborative Model for Developing Electronic Records Capacities in the Small Archival Repository."



Documenting Internet2

A Collaborative Model for Developing Electronic Records Capacities in the Small Repository

Project Documents

- "Documenting Internet2" Session Description, Society of American Archivists Annual Conference, August 2005 (1 Jul 2005)
- "Documenting Internet2" Session Description, Midwest Archivists Conference Annual Spring Meeting, April 2005 (1 Jul 2005)
- Current Use of eRoom and the Document Library (15 Feb 2005)
- Agenda: Advisory Board Meeting, January 19-20, 2005 (11 Jan 2005)
- Appraisal Options for Internet2's Records (30 Oct 2004)
- Records Overview for Internet2 (29 Sep 2004)
- Survey Report: Engineering Record Groups (6 Sep 2004)
- Survey Report: Middleware and Security Record Groups (4 Sep 2004)
- Survey Report: Applications Record Groups (1 Sep 2004)
- Survey Report: End-to-End Performance Initiative Record Groups (1 Sep 2004)
- Survey Report: Backbone Network Infrastructure (BNI)/Abilene Record Groups (31 Aug 2004)
- Survey Report: International Relations Record Groups (29 August 2004)
- Project Archivist's Report for the week of August 23, 2004 (29 Aug 2004)
- Survey Report: Member Activities Records (16 Aug 2004)
- Project Archivist's Report for the week of August 9, 2004 (15 Aug 2004)
- Report on Technology Environment at Internet2 (30 Jul 2004)
- Poster Presentation (28 July 2004)
- Project Archivist's Report for the week of July 19, 2004 (19 Jul 2004)
- Agenda: Advisory Board Meeting, July 15-16, 2004 (9 Jul 2004)
- Project Archivist's Report for June 28–July 2, 2004 (7 Jul 2004)
- Survey Report: Corporate Relations and Industry Strategy Council Records (1 Jul 2004)
- "Documenting Internet2 Update" CBI Newsletter, v.26 nos.3&4 (Spring/Summer 2004)
- Project Archivist's Report, June 8, 2004 (8 Jun 2004)
- Project Archivist's Report for the week of May 10, 2004 (14 May 2004)
- Internet2 Setting Description, abridged (21 April 2004)
- Internet2 Setting Description (14 April 2004)
- "New Research Grant to CBI" CBI Newsletter, v. 26 no. 2 (Winter 2004)

Original Grant Proposal

- Project Summary (29 May 2003)
- Project Narrative (29 May 2003)
- Schedule of Completion (29 May 2003)
- Plan of Work, revised (7 Jan 2004)

The immediate context of the project was an increasing level of concern about the future of archival collecting at CBI. Established in 1979, CBI is arguably the world's pre-eminent repository of research material on the history of information technology. CBI's collection strengths follow the development of computing, from the calculators of the Burroughs Corporation in the nineteenth-century, through the first electronic digital computers after World War II, the emergence of software products distinct from hardware, the growth of the computer and software industries, and the advent of networking. The collections include personal papers, institutional records of various kinds, oral histories, product literature, and rare publications. By far the largest collections are corporate records and records of professional organizations, all of which are in traditional formats. Over the years, CBI archivists had, with some regret, declined offers of digital materials. These decisions were by and large not too painful: most of the materials offered were of uncertain research value, or raised proprietary issues, or were in obsolete formats that would have required prohibitively expensive reformatting before we could even begin to address the larger issues. Other collections offered to CBI were from defunct companies or organizations that could offer no assistance to help us understand the content of inadequately labeled tapes and discs. Finally, CBI simply didn't have the infrastructure to support basic preservation of electronic records, much less to commit to providing long-term access. This state had become untenable, though, by the late 1990s, as the documentation produced in CBI's subject area was increasingly created in digital forms with no paper equivalent and as collections of serious research interest became available.

> ... CBI simply didn't have the infrastructure to support basic preservation of electronic records, much less to commit to providing long-term access.

When project planning began in late 2002, few institutions had implemented successful electronic records preservation programs, and those that had were either large corporations or governmental entities with institutional archives and records programs. These programs were not scaleable to a collecting repository such as CBI. Another key difference was that, as a collecting repository, CBI can only inherit records, not influence their production, as is often suggested in the archival literature. Finally, archivists had yet to become seriously involved in the fledgling institutional repository movement. As a result there were few models for smaller collecting repositories seeking to expand their scope into the digital realm.

Overarching project goals were twofold:

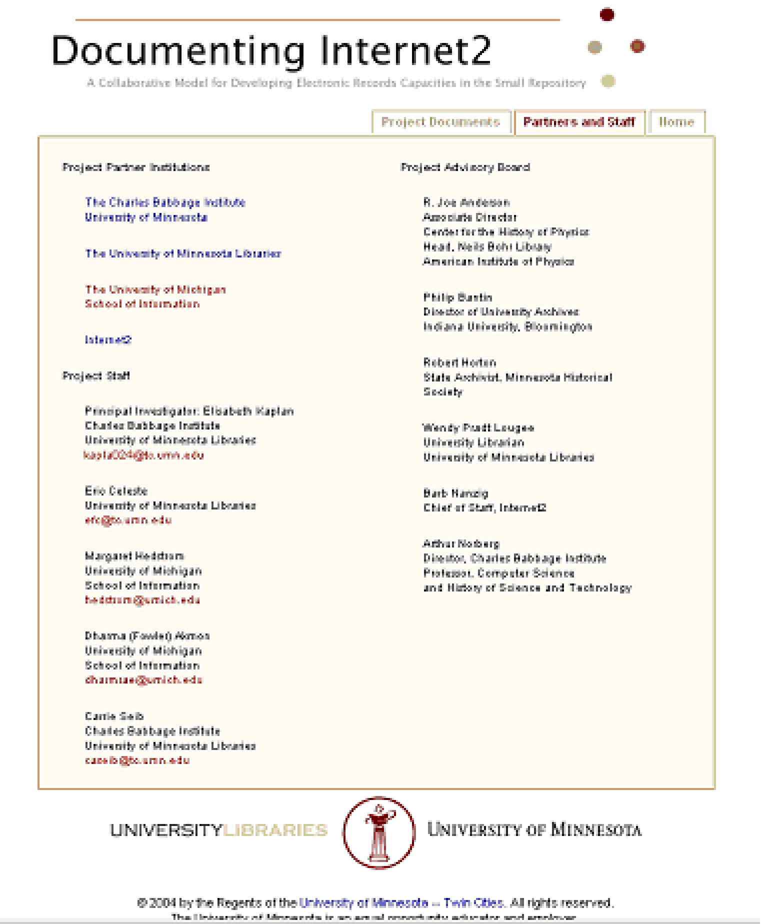
- to explore the application of traditional archival appraisal methods in a digital context through hands-on experimentation with techniques and tools
- to begin an assessment of institutional capacity for building a sustainable digital component into the CBI archives program, within the context of the University of Minnesota Libraries

The project team agreed on some basic principles from the start. First, this was a planning, not an implementation grant, meaning that we would explore issues and conduct experiments, but we didn't expect to have accessioned a digital collection by the project's conclusion. Second, our orientation was fundamentally practical, not theoretical: we were aiming for adequacy, not perfection, and we wanted to get our hands dirty. Finally, for the archivists on the project team, this was an opportunity to learn about techniques and approaches from other fields and evaluate their application in an archival setting. We were prepared to be flexible in our approach.

We hoped the project would lead to some practical application at CBI, as well as to guidelines or best practices that could be useful in other settings.

The Partners

We found the ideal object of our documentation in Internet2. Headquartered in Ann Arbor, Michigan, Internet2 is a research and development consortium of universities working in partnership with government and industry to develop and deploy advanced network applications and technologies. As an organization at once developing and deploying new technology, it fit squarely within CBI's collecting scope. As a collaborative and distributed organization whose modes of communication were primarily born digital, it provided a promising vehicle for our exploration. Other factors contributed to the appeal of working with Internet2. As a publicly supported organization, concerns over access to proprietary records would be minimal. As a living, growing organization, it would offer us the opportunity to interact directly with records creators to better understand their culture and practices.The timing was good, too: Internet2 was approaching its tenth anniversary, and Director Doug Van Houweling and Chief of Staff Barbara Nanzig were beginning to think about how best to capture and manage their organizational records. Nanzig generously provided us access to records, access to key staff members, on-site workspace, and even took the time to travel to multi-day advisory board meetings and archival professional conferences.



Documenting Internet2

A Collaborative Model for Developing Electronic Records Capacities in the Small Repository

Project Partner Institutions

The Charles Babbage Institute
University of Minnesota

The University of Minnesota Libraries

The University of Michigan
School of Information

Internet2

Project Staff

Principal Investigator: Elisabeth Kaplan
Charles Babbage Institute
University of Minnesota Libraries
kapla024@tc.umn.edu

Eric Celeste
University of Minnesota Libraries
efc@tc.umn.edu

Margaret Hedstrom
University of Michigan
School of Information
hedstrom@umich.edu

Dharma (Fowler) Akmon
University of Michigan
School of Information
dharmrae@umich.edu

Carrie Seib
Charles Babbage Institute
University of Minnesota Libraries
caseib@tc.umn.edu

Project Advisory Board

R. Joe Anderson
Associate Director
Center for the History of Physics
Head, Neils Bohr Library
American Institute of Physics

Philip Bantin
Director of University Archives
Indiana University, Bloomington

Robert Horton
State Archivist, Minnesota Historical Society

Wendy Pradt Lougee
University Librarian
University of Minnesota Libraries

Barb Hanzig
Chief of Staff, Internet2

Arthur Norberg
Director, Charles Babbage Institute
Professor, Computer Science
and History of Science and Technology

© 2004 by the Regents of the University of Minnesota -- Twin Cities. All rights reserved.
The University of Minnesota is an equal opportunity educator and employer.

The project team and board of advisors, like Internet2 members, were geographically distributed. One valuable lesson from previous NHPRC projects was that a collaborative approach and a wide range of expertise would be required. We deliberately sought partners from a variety of fields, with expertise in such areas as digital library development, electronic records program development, library information technology, historical research, documentation strategies and functional analysis, institutional archives, organizational culture and communication, and records management. Beth Kaplan and Carrie Seib led the project from the CBI end; Margaret Hedstrom and Dharma Akmon of the University of Michigan School of Information led the on-site work and experiments. Eric Celeste at the University of Minnesota Libraries worked with Akmon and Hedstrom on the experiments from the Minnesota end. The advisory board provided input at two two-day meetings during the course of the project, and each member contributed substantially.[2]

Project Methods and Activities

Archival appraisal is complex and not uncontroversial. Several theories of institutional archival appraisal underlie its practice. Reduced to its simplest level, one school of thought would have the archivist develop a thorough understanding of the organization and its history before even thinking about collecting activities. A related approach would begin with an intensive analysis of the organization's core functions, before identifying the ways in which documentation is produced in the carrying out of those functions. Other methods would advise the archivist to jump right in with a records survey and learn about the organization later. Still another would be to let collecting be guided by the organization's hierarchical structure. While we had our biases, we didn't want to rule out any possibilities.

Akmon began the appraisal process with intensive research on Internet2's history, structure, and functions. She prepared a detailed setting description that would provide much of the contextual information we needed in order to move forward and to which we would return regularly throughout the project.[3] In May of 2004, Akmon began full-time, on-site work at Internet2, with the goal of compiling information about records and record-keeping practices at the organization. Akmon did conduct a traditional records survey, but by far the most valuable source of information about records came from interviews with key staff members. The interviews provided a wealth of

The value of this information suggests that "soft skills," in information on the organization’s core functions and how those translated into records, as well as insights into records creators’ record-keeping needs, practices, assumptions, and attitudes. The value of this information suggests that “soft skills,” in this case the ability to interact with records creators on their terms, may be increasingly important for archivists working with modern organizations.

this case the ability to interact with records creators on their terms, may be increasingly important for archivists working with modern organizations.

One widely held assumption on the part of Internet2 leadership was that the organization’s core records were routinely captured and made accessible through two central electronic document management systems: Documentum’s eRoom and the Internet2 Document Library. A significant turning point in the project occurred when we realized this was not the case. Despite the availability, ease of use, and apparent convenience of these tools, and despite a high level of institutional commitment and technological savvy among staff members, most people relied on personal computers and the Internet2 Web space for keeping and sharing the documents they viewed as important. As we reviewed information from the surveys and interviews, we realized that in reality, the Internet2 public website functioned as the only centralized, shared, repository of core organizational content.

Experiments and Findings

Why were the document management systems in place at Internet2 underutilized? The Document Library had been designed specifically for use at Internet2 to serve as “an authoritative source of documents and other deliverables produced by Internet2 Working Groups, Advisory Groups, or Projects.”[4]

The Document Library is publicly accessible online via the Internet2 website. We inventoried its contents in February 2005 and conducted a detailed review of its features including its structure and use: the submissions process, existing guidelines, metadata requirements, and other features that could impact submissions.

Our first experiment addressed the theme of records creators’ behaviors and tested the submission model of collecting. In a perfect world, high-level mandates and compliance with records management would result in the routine capture of appropriate documentation. Given the opportunity to work with records creators at Internet2, we wanted to test staff responses to record-keeping guidelines and to learn more about why the Document Library and the eRoom did not turn out to comprise a cache of critical documentation as had been assumed.

We discovered several reasons for this. Interviews with staff revealed that the Document Library is in fact a hyperlinked citation list. Entries in the Document Library serve as pointers to documents in staff’s personal Internet2 server space. Staff members who do submit material to the library also have the ability to move or remove the documents without warning and without leaving a documentary trail. Because of this, and despite its potential, the Internet2 Document Library clearly was not living up to its intended purpose.

We wanted to know more about the depth of records creators’ willingness (or reluctance) to add another step to their daily routines and what factors might affect that level of involvement.

But what if upper management mandated submission of content to the Document Library? We wanted to know more about the depth of records creators’ willingness (or reluctance) to add another step to their daily routines and what factors might affect that level of involvement. We wanted to see how employees would respond to a very clear management directive to submit documents to the Document Library, accompanied by the rationale that, essentially, it was for the greater good of the institution. We asked a senior staff member to send a message to managers in two areas of the organization reiterating the purpose of the Document Library and encouraging staff members to submit the documents and deliverables of their areas and working groups “so that the Document Library can become the single, authoritative source for working papers, technical reports, proposals and recommendations, and any other important project documentation.”

We monitored additions to the Document Library for approximately a month after the message was sent, and during that time not a single additional document was submitted. In fact, since its implementation in fall 2003, only twenty documents had been submitted to the Document Library, and interviews with staff revealed an important reason. Submitting content to the Document Library involves filling a fairly extensive list of metadata fields based upon Dublin Core guidelines, a time consuming task with no perceived direct value to the records creators. In other words, there was no real incentive for compliance.

With this less than promising response to the “submission” model, we determined to focus next on a "capture" model that would entail little or no extra work on the part of the records creators. We knew from the interviews that staff members gravitated toward the Internet2 website for finding and sharing documents. Our next pilot project was a crawl of the website in order to capture those documents.[5] We conducted three crawls over a period of two months.

Once the crawl was processed, we had an online mirror of the Internet2 website and one that could be searched using existing desktop software.[6] Although the number of documents captured varied from one area in the organization to another (as would be the case in the paper world, as well), a surprising volume of important documentation was captured. We were able to create our own archived instance of the site that could be surfed and searched as though it were still live. The crawl stored each page as we found it while also preserving the connections inherent in Web content (the links and hierarchy), or, in archival terms, the "original order." The idea of original order is that a document's meaning is related to and to some extent derived from the context in which it was created or filed by the creator. Archivists conducting records surveys in the paper world are careful to capture even the most mundane of contextual information (specific location, labels, post-its, and notes on drawers) because of the meaning these bits of information can convey.[7] The ability to capture the documents in their original order while also enhancing search capabilities was one of the most exciting aspects of the project.

> ...only twenty documents had been submitted to the Document Library, and interviews with staff revealed an important reason. Submitting content to the Document Library involves filling a fairly extensive list of metadata fields based upon Dublin Core guidelines, a time consuming task with no perceived direct value to the records creators.

Some notable shortcomings were also immediately apparent. The crawl results, for example, do not provide the date of file creation or last modification, only the date the material was crawled; a significant drawback. As well, certain uses of scripts, we found, can render a Web page uncrawlable.[8]

While one particularly appealing aspect of the crawl method is that it requires minimal disruption of staff workflow and no time wasted trying to enforce record-keeping requirements, there was a learning curve, and the crawl required significant intervention. Our first crawl got hung up on Web-based calendars and its parameters had to be redefined. Further, on top of time spent on our side to administer the crawls and process the results, the richness of our crawl results was a function of Internet2's willingness to work with us. This enabled us to capture significant documentation missing from our first crawl.

> We believe our results would have been significantly enhanced had we been able to capture data from staff members' personal computers. But capturing personal computer data would have raised a host of problems: many staff members interviewed acknowledged that they keep personal, non-work-related documents on their work computers, making privacy and

We were able to conduct the Web crawl in part because so much of the Internet2 website is publicly accessible. We believe our results would have been significantly enhanced had we been able to capture data from staff members' personal computers. But capturing personal computer data would have raised a host of problems: many staff members interviewed acknowledged that they keep personal, non-work-related documents on their work computers, making privacy and the relevance of documents captured a concern. Further, personal computers are not only used as document storage systems, but also as work spaces, meaning that documents captured are not necessarily complete. In the context of this project, we decided that arranging to capture PC data would not have been worth the effort.

As the cost of storage diminishes, and the availability of increasingly sophisticated tools grows, we believe that the data capture model will be an important strategy for institutional archives.

Parting Thoughts

As organizations create and manage their records differently, to have any hope of preserving them or the "order" in which they are kept will take a very basic rethinking of how archivists actually get at the records they hope to collect. The theory and appraisal models discussed above provided an excellent starting point, but in the end, we needed to step outside of the box in order to actually get at the records. This also demonstrates the value of a diverse group of project the relevance of documents captured a concern.

participants and advisors. Had we not considered the suggestion of a librarian on the advisory board that we “save everything,” we wouldn’t have come to the realization that capturing “everything” through the crawl would not only gather content but preserve connections and context.[9]

Other, more conventional, organizations will have a much different record keeping landscape than that of Internet2, which might allow for appraisal at a more granular level. Some collections will be so large, distributed, and unmanaged that they can realistically be “processed” on only the highest level. Appraisal in this context may mean determining to do the Web crawl, but forego the email or the information on people’s personal computers. It may mean selecting which desktops to capture and which document file types are the most important in adequately capturing the documentary record of the organization. The project affirmed our bias that a thorough evaluation of the record keeping landscape in a particular organization is more critical than ever. The project also highlighted an unexpected level of similarity between decision making in the digital world and in the paper world. It may be that as archivists, we have just become comfortable with the spectrum of options when working with paper, and more willing to accept imperfection. Just as in the paper world, archivists need to work with the organization to identify options, determining what is possible and what is practical, and making decisions based on that information. In the digital context, these decisions are made more consciously and are more visible and open to scrutiny.

Had we not considered the suggestion of a librarian on the advisory board that we “save everything,” we wouldn’t have come to the realization that capturing “everything” through the crawl would not only gather content but preserve connections and context.

If we began the project hoping to create some best practices and guidelines that would help us to begin to build our program and contribute something to others, we concluded it with a deepened respect for the adage that “no one size fits all.” At the same time, our findings—the successes, the problems, the questions that arose—will be useful for other projects and at other institutions. Internet2 is perhaps an extreme example of a modern high tech organization, but in fact we are already drawing on our findings in other contexts.[10]

Notes:

[1] National Historical Publications and Records Commission grant number 2004-036. We are grateful to the NHPRC for making this project possible.
[2] Arthur Norberg, CBI; Wendy Lougee, University of Minnesota Libraries; Bob Horton, Minnesota Historical Society; Joe Anderson, American Institute of Physics; Phil Bantin, Indiana University; and Barbara Nanzig, Internet2.
[3] All project documents referenced here are available online at www.cbi.umn.edu/documentinginternet2/
[4] http://docs.internet2.edu/
[5] The primary technology used for the crawl was Heritrix (a Java-based crawler from the Internet Archive). Perl, JavaScript, PHP (along with a Web server that can serve the PHP), MySQL, and the assistance of several University of Michigan graduate students were used to process the results of the crawl. Detailed information is accessible at http://wiki.lib.umn.edu/DI2/HowToCrawl. The Heritrix crawler now also powers the Internet Archive’s ArchiveIt service, which simplifies focused crawling for institutions.
[6] We used Apple’s Spotlight tool for this task.
[7] For example, “the first three drawers of the tan file cabinets against east wall in the hallway contain background files used to support patent litigation for software product A.”
[8] We used a JavaScript URL rewriting method, and while this is very functional for simple Web pages, it can fail when the crawler encounters Web pages that use scripts to construct links on the fly.
[9] Of course, lurking below the surface of all of this (and outside the scope of our project) is the continuing problem of longterm preservation. Our crawl summary revealed thousands of PowerPoint, PNG graphics, XML, and PDF documents, hundreds of MS Word, RealAudio, ZIP files, and dozens of other movie formats—in all about two dozen file formats. We offer no new solution to these thorny issues, but we do have a significant body of content that we can experiment with. In the meantime, we just note here that we are committed to “bit preservation” and to capture of content in platform independent, open sources contexts whenever possible—without any real idea how to maintain access to these formats over time.
[10] The lessons learned from the Internet2 project have already been applied not only at CBI but in the University Archives and Minnesota’s University Digital Conservancy initiative, which serves as the digital arm of the University Archives as well as a repository for faculty research.