



Against the Grain

“Linking Publishers, Vendors and Librarians”

ISSN: 1043-2094

Taking Charge: Preserving Our Digital Heritage Part II

by **Amy Kohrman** (Marketing Director, CLOCKSS and LOCKSS, 1450 Page Mill Road, Palo Alto, CA 94304) <akohrman@stanford.edu> <http://www.clockss.org> and <http://www.lockss.org>

Remembering 2008...

Only eight months ago in this same space, *Against the Grain* guest editor **Bruce Heterick**, (Portico’s Director of Library Relations) referred to the timeless essay, “The Tragedy of the Commons,” which describes what can happen when multiple individuals, acting independently in their own self-interest, ultimately destroy a shared, finite resource. **Heterick** invoked the “Tragedy of the Commons” to make a case for the importance of digital preservation. He could have not been more prophetic. Since then, in the blink of an eye, the tragedy became a reality. Banks collapsed, the stock market sank to 30 year lows, and the US economy was brought to its knees. Today, layoff announcements, budget cuts, and bailouts dominate the U.S. news media with global reverberations.

Given the dire state of things, could there be a worse time in modern history to promote

digital preservation? What institution can afford to add new costs in the face of Draconian cutbacks?

That’s one way of looking at the landscape. One could also ask, has there ever been a more critical time to be preserving our digital patrimony? It seems each new day brings with it another potential for irrevocable loss to our communal knowledge. Whether it is a newspaper, publisher, or bank, long-lived producers of information are going under at an alarming rate.

To be sure, our current model of working largely independently at digital preservation in academia is quickly becoming unsustainable. In 2008, the Windsor Group Declaration on Digital Archiving stated, “because scholarly publications are a collective resource, collective action is needed to ensure their preservation.” The pertinence of this argument is all the more palpable today.

Indispensible lessons about the fragility of monolithic structures and the positive outcomes of collective, collaborative preservation actions are presented in this volume. Not surprisingly, five of the seven extraordinary digital preservation projects featured are recipients of funding from **The Library of Congress’ National Digital Information Infrastructure and Preservation Program**. NDIIPP has long worked to build a national collaborative distributed preservation infrastructure.

What is needed today, more than ever before, is the deployment of concrete strategies. This issue illustrates that a number of publishers and libraries are not sitting back and waiting for the world to heed the call of digital preservation. They are embracing the Tragedy of the Commons challenge and preserving fragile assets, for our institutions, and for the benefit of future generations! 🌳

If Rumors Were Horses

What To Look For In This Issue:

<i>A Healthy Information Economy.....</i>	18
<i>Looking Back, Looking Forward.....</i>	22
<i>Social Media Simplified.....</i>	32
<i>Top Ten Innovations in Library History.....</i>	42
<i>Funds and Accounting Trees.....</i>	65
Interviews	
<i>Richard Charkin.....</i>	22
<i>Hazel Woodward.....</i>	46
Profiles Encouraged	
<i>George Burrows.....</i>	20
<i>David Levinson.....</i>	26
<i>Dave Pollard.....</i>	30
<i>Karen Christensen.....</i>	34
<i>Eric Calaluca.....</i>	38
<i>Plus more.....</i>	<i>See inside</i>



against the grain
people profile

Marketing Director, LOCKSS and CLOCKSS
1450 Page Mill Road, Palo Alto, CA 94304
Phone: (650) 725-1134 • Fax: (314) 584-3874
<akohrman@clockss.org> • www.lockss.org and www.clockss.org

A. Kohrman

BORN & LIVED: New York, Massachusetts, London, England, Kunming, Yunnan Province, China, California.

EARLY LIFE: Loved books and publishing.

FAMILY: Husband, two kids, one dog. 🐕

continued on page 16

Preserving Digital Public Television: Preparing for the Broadcast Afterlife

by **Nan Rubin** (Project Director, Preserving Digital Public Television, 450 W. 33rd Street, New York, NY 10001; Phone: 212-560-2925; Fax: 212-560-2833) <rubinn@thirteen.org> <http://www.ptvdigitalarchive.org> *Thirteen/WNET.org*

“Public Television is responsible for the production, broadcast and dissemination of programs which form the richest audiovisual source of cultural history in the United States.” — Librarian of Congress, 1997

New Preservation Practices for Television Archives

In less than a decade, television production, distribution and preservation has undergone a radical shift. Today, programs are nearly all shot, edited, and shared as digital files. Video recording and editing systems are now well within the means of most members of the public, and the ubiquity of media on the Internet, coupled with the mass deployment of hand-held devices, have transformed not only the medium of television but the entire environment for creating and watching moving images.

Distribution and transmission have been equally transformed, as tape-based submissions to the **Public Broadcasting Service (PBS)** and other national program services are being replaced by digital file transfers. On-demand viewing is growing just as on-air signals become all-digital, when every analog transmitter is turned off in 2009.

What do these changes mean for television archives? Practices to conserve and protect videotape recordings are well established, and the cost for maintaining and storing physical media are easily calculated. However, in an age of digital files, the requirements for preserving television programs are far different from storing videotape. It isn't enough to close a digital file and put it on a virtual shelf. For video in particular, acceptable practices to save and access very large files, manage ever-changing file formats, and maintain rich metadata are just now emerging.

Preserving Digital Public Television, a project funded by the **National Digital Information and Infrastructure Program of the Library of Congress (NDIIPP)**¹ set out to solve some of these difficult problems by designing a model repository for public television. In the process, the project also determined standards for metadata, explored rights issues relating to video archives, analyzed operating costs, and brought a new consciousness about the importance of digital preservation to the public television system.

Bringing Digital Preservation to Public Television

In the **Public Broadcasting Act** of 1967, Congress authorized the **Corporation for Public Broadcasting (CPB)** “to establish and maintain, or contribute to, a library and archives of noncommercial educational and cultural radio and television programs and

related materials.” However, CPB never allocated any funds to support this charge, and no demand for system-wide preservation was implemented. Consequently, only a few stations have established formal archiving activities to preserve their own materials.

Without a preservation mandate, digitally produced programs in public television are at great risk of being lost. The rapid changes in digital technology are rendering recording and playback systems obsolete at breakneck speeds, at the same time adequate tools for managing large and complex video files are not yet perfected. This has left a very large gap in the preservation of America's public television legacy.

Public television stations **WNET** in New York and **WGBH** in Boston, which produce roughly 60% of the national prime time series including *Frontline* and *NOVA* at **WGBH**, and *American Masters* and *Great Performances* at **WNET**, recognized this challenge early. Because **WNET** and **WGBH** each maintain its own archives, the stations were already committed to long-term program preservation. Both knew that solving the demands of digital preservation would be costly and that no station could do it alone — it would take a collaborative effort.

The Preserving Digital Public Television Collaboration

When the **Library of Congress** invited proposals under **NDIIPP**, **WNET** and **WGBH** partnered with **PBS** to build a model preservation repository for “born-digital” public television programs. **PBS** operates the network that distributes public television programs to more than 300 stations, and because most national programs pass through **PBS** before they are aired, it is the principle *de facto* repository for these programs. (The **PBS** warehouse holds more than 150,000 videotapes of programs going back more than 40 years).

These institutions understood that public television had to take steps to protect its rapidly growing collection of digital assets. As broadcasters, however, they had little experience building a preservation repository. **New York University** provided the expertise that was lacking. The **NYU Digital Library** team had extensive experience designing repository systems specifically for large digital files wrapped in metadata. The project further benefited from a relationship with **NYU's Moving Image Archiving and Preservation Masters Degree Program**, whose students have provided excel-

lent research and whose graduates have become full-time project staff.

Together, **WNET**, **WGBH**, **PBS** and **NYU** organized **Preserving Digital Public Television (PDPTV)**² as a collaborative to introduce digital preservation issues and practices to the public television system. Understandably, the priorities of public broadcasting are program production and broadcast delivery, not saving program assets. Most program preservation is handled as an afterthought. To be successful, **PDPTV** had to demonstrate that *building* a repository was technically possible, and that *operating* a repository was functionally and economically feasible.

There were two major project goals:

1. Design a model preservation repository for large digital video files, and examine operating issues related to content selection, costs, and access.
2. Build system-wide support for digital preservation.

The project formally began in September 2004 and will be completed in 2009.

Building a Model Repository

The process for building the repository was initially conceived as a series of discrete technical tasks in a lab-type environment, with the approach that identifying commonly used file formats, determining appropriate metadata requirements, and adopting technical standards would be critical to repository functionality. The project naively assumed that commercial television networks and large collecting institutions such as the **Library of Congress** (completing its **Packard Campus of the National Audio-Visual Conservation Center**) were already making progress solving these same problems, and that public television could simply “tag along” with work underway.

The project quickly learned, however, that this was not the case. In reality, other video producers including the networks and the library itself were struggling with these technical issues and not making much progress.

Instead of following along, the project found itself in the unanticipated position of actually leading the effort to create a set of standards for preservation-based video file wrappers for the television industry. Likewise, little had been done to determine what metadata should accompany the video files, and the project was one of the first to adopt a set of metadata schema appropriate for long-term video preservation. Both of these outcomes were unexpected.

Collecting and Analyzing Metadata

The **NYU Digital Library** team based their repository design on **DSpace**, which they had

continued on page 18

used to build other moving image archives. Technical issues rested primarily on how best to organize files and metadata to create **Submission Information Packages (SIPs)** and **Archival Information Packages (AIPs)** using test digital program files.

The team used a sample of 35 hours of program files, all standard definition, drawn from *Nature*, *Frontline* and *Religion and Ethics Newsweekly*, plus a local selection from *New York Voices*. The test files originated from three sources — uncompressed program masters from **WNET** and from **WGBH**, and compressed distribution versions of the same programs from **PBS**. This provided a mix of both high and low resolution program file formats, with different flavor files from each source.

A fundamental requirement was to configure **AIPs** for long term storage by aggregating content plus metadata for each program without adding anything new. The **SIPs** therefore, had to contain comprehensive program-related and technical metadata along with the program files themselves.

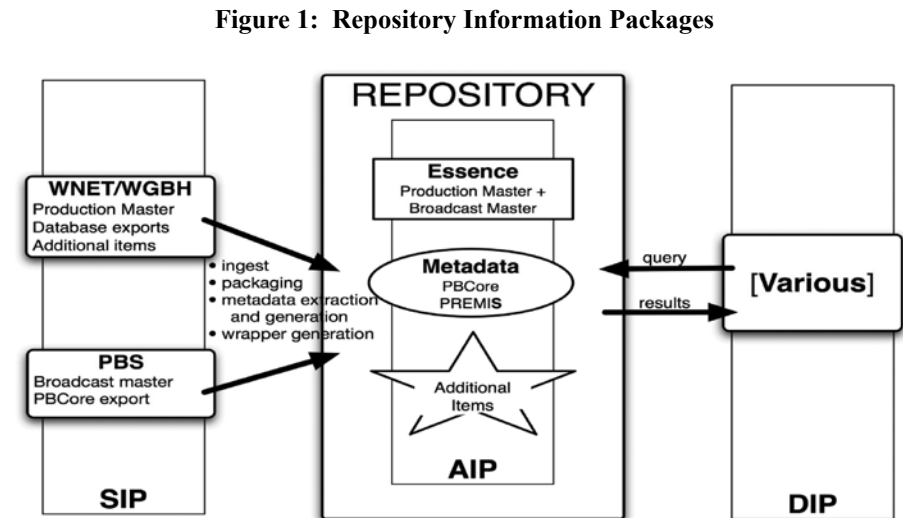


Figure 1: Repository Information Packages

Operating from the assumption that the repository should conform to the **OAIS** reference model for creating a trusted repository, the project examined a broad range of metadata schema used by libraries and archives. It also looked at standards emerging in commercial television, and assessed **PBCore**,³ a metadata

dictionary based on **Dublin Core**,⁴ designed specifically for public radio and television program files.

In practice, determining the appropriate sets of metadata fields was an intensive task. Individual program files were accompanied by a wide range of metadata, but because program information is not collected systematically even within **PBS**, it had to be gathered from multiple sources on a program-by-program basis. Also, because there are no uniform criteria, the quality of metadata associated with each program was idiosyncratic and inconsistent. To determine the components required for the **AIP**, the collected metadata had to be analyzed, particularly the extensive descriptive and rights metadata created by **PBS** for broadcast scheduling.

Although **PBCore** is in the early stages of adoption, the repository chose to build its descriptive metadata requirements around it, which has encouraged others to use it as well. As a result, the most important source of metadata for national programming, **PBS's Program Offer Data Service (PODS)**, can now be exported directly into **PBCore**, making national program information much easier to access.

Incorporating technical metadata from the video files also proved to be a challenge. Because the program files were submitted to the repository in many formats (including such diverse wrappers and encoding formats as **MXF**, **Quicktime**, and various flavors of **MPEG** and **DVC Pro**), multiple tools were required to play the videos and extract information such as bitrate, file size, and frame size. Transforming this disorganized metadata into a standardized **AIP** was clearly a necessity.

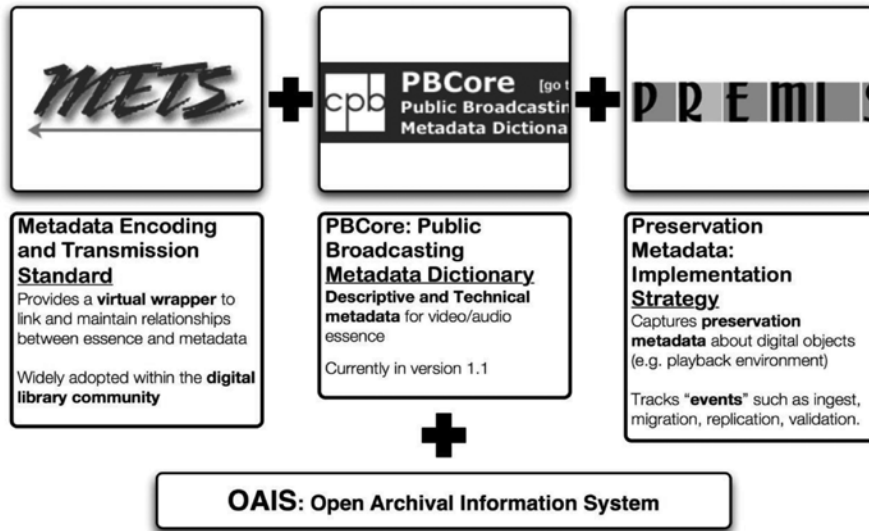
The solution was to select schema from several data dictionary standards that encompass descriptive and technical metadata, while maintaining information unique to public television programming. This **AIP** structure uses not only **PBCore**, but also **PREMIS (Preservation Metadata Implementation Strategies)**⁵ and **METSRights (Metadata Encoding and Transmission Standards Rights**

continued on page 22

continued on page 20

Declaration Extension Schema). Appropriate fields from these standards, along with virtual links to the program files themselves, are all contained within a METS wrapper.

Figure 2: Metadata Schema



Thinking it was desirable to capture some types of metadata during the production process, the project analyzed the workflows used to produce the sample programs. The intent was to identify points where key metadata and file types were created, and track them through the program lifecycle. However, in a complex production process that involves many stages, the need to plan for file preservation had little bearing on production deadlines. Consequently, there was no opportunity to test preservation practices in the production workflows.

The problems encountered testing these various file formats, combined with the time-consuming efforts to collect metadata, demonstrated the high priority for setting uniform metadata and technical standards for any future operation. Without them, automating the functions for extracting and managing the metadata and file integrity of large collections is simply not feasible.

Video File Wrappers

The use of a standardized video file wrapper is considered a requirement for successfully exchanging digital files, particularly to support future file migration. A number of so-called video wrapper "standards" exist, but despite vendor claims, the files do not all actually interoperate with many equipment configurations used by public broadcasters. To examine this issue, the project convened a "Wrapper Roundtable" of technologists, digital archivists and industry leaders.

The group was surprised to learn that the lack of consistent wrapper standards was also a major problem for the commercial networks. Any initiative to create technical standards for public television must dovetail with the needs of the commercial broadcasting industry,

because public television on its own does not carry enough economic clout to influence hardware vendors. The wrapper standard remains to be solved, but due in part to the "Wrapper Roundtable," the **Advanced Media Workflow Association** which represents vendors, has made a commitment to develop an appropriate standard that vendors will support.

Assessing Rights and Operating Costs

Television programs are multi-layered productions comprised of original and licensed elements from a myriad of sources and subject to a wide array of contract obligations, rights restrictions, and other encumbrances. Typically, rights to use this material in a non-commercial broadcast are granted for a finite period, for example five to ten years. When these rights expire, the system can no longer use the program without "re-upping" the rights by getting new permissions from each individual rights holder. It can be very expensive to find all the rights holders, renegotiate and pay for new use agreements. Consequently, unless a program is in great demand, rights are rarely renewed.

Specific authorization to preserve public television programs after the broadcast window expires is largely absent, and making older programs available for anything but the most narrow uses is fraught with risk of copyright infringement. There are a number of efforts in the U.S. and Europe working to improve the situation, but under current conditions, the **PDPTV** model repository is "dark" to the public until viewing and use rights become more favorable.

As existing digital repositories mature, operating costs are being documented by such institutions as **The National Science Foundation**, which commissioned the **Blue Ribbon Task Force on Sustainable Digital Preservation and Access (BRTF-SDPA)** in 2007 specifically to study cost models for large database repositories. The contribution of **PDPTV** has been to focus on the particular concerns of maintaining very large digital video files with

a manageable scale of operations. **PDPTV** is also closely monitoring the growing body of research being published on this topic.

Promoting System-wide Support

From the beginning, project partners promoted a position within public broadcasting that planning for digital preservation was no longer optional — it was a necessity. The explosion of online broadcast content, coupled with a constantly changing array of viewing devices, have created a fundamentally altered video environment which requires programming to be viewable on everything from the very smallest iPod screen to giant wall-size flat panels.

Amid such extremely fluid technology, the project emphasized the importance of adopting standards for technical operations, plus consistently collecting critical metadata. Because these are the very same factors necessary for successful multi-platform digital distribution, the project was able to tie digital preservation to effective reuse of program content. The concept of digital preservation thus became highly relevant to stations, elevating it from a marginal concern to a major subject in the public television debate on how to make content available to reach more viewers.

Lessons Learned

Over the course of the project, a number of important lessons became evident.

- Technical problems will eventually be solved and standards will be adopted when private industry agrees to collaborate. But this is a slow and bumpy process.
- With producers beginning to use all-digital production workflows, now is the moment to introduce preservation compliant metadata requirements into the process. This should be done quickly or the opportunity might be lost.
- Prompted by the preservation message, stations around the country are actively exploring partnerships with other local cultural heritage institutions to share resources for preserving their respective digital collections.
- Despite a great deal of progress, a system-based commitment to preservation must be reinforced as an important national investment. Instead of being seen as overwhelming, costs need to be presented as feasible and manageable.
- Although there are some aggressive efforts to tackle the thicket of rights issues, especially for educational use, overall public television seems unwilling to push boundaries for wider access to archival content. Much more can be done in this area.

Maintaining Momentum

Since **Preserving Digital Public Television** began, broadcasting has shed its analog systems and moved completely into a digital universe. This project has been able to impress on the public television system the message that digital preservation is not an optional

continued on page 22

“add-on” cost, but a requirement for any future use of the materials. In this, the project has been instrumental in transforming an attitude of indifference to one that acknowledges the value of properly managing our collective archival holdings.

In a further indication of support, for the very first time CPB allocated preservation funding to pilot **The American Archive**. **The American Archive** will develop a repository for public radio and television, and PDPTV anticipates making a significant contribution to this initiative.

Viewers keep reminding us that public television programming is precious and has made an indelible imprint. What remains is to continue building commitment across the entire system, so the critical responsibility for saving this American media legacy will be shared, sustained and nurtured over time. 🌱

Endnotes

1. <http://digitalpreservation.gov>.
2. <http://www.ptvdigitalarchive.org>.
3. <http://www.pbc.org>.
4. From **Wikipedia**: The **Dublin Core** metadata element set is a standard for cross-domain information resource description. It provides a simple and standardized set of conventions for describing things online in ways that make them easier to find. **Dublin Core** is widely used to describe digital materials such as video, sound, image, text, and composite media like Web pages.
5. <http://www.oclc.org/research/projects/pmwg>.

against the grain people profile

Project Director, Preserving Digital Public Television
450 W. 33rd Street, New York City, NY 10001
Phone: (212) 560-2925 • Fax: (212) 560-2833
<rubinn@thirteen.org> • www.ptvdigitalarchive.org
Thirteen/WNET.Org

Nan Rubin

BORN AND LIVED: 1949, Newton, MA. Have lived in Ohio, Colorado, Washington DC, New York City.

EARLY LIFE: Hippie, Folkie, Lefty, Techie.

PROFESSIONAL CAREER AND ACTIVITIES: 30+ years building community radio stations and community media by special focus on facilities, technical planning and creating infrastructure. Put two community radio stations on the air, in Cincinnati and Denver. Long-time supporter of ethnic public media, particularly Native American projects. A founder of the **National Federation of Community Broadcasters** and the **World Association of Community Radio Broadcasters**. Producer of **The Hidden Jews of New Mexico** radio series, one of the most the most popular programs ever aired on NPR. Organizer of the **Highlander Media Justice Gathering**, which helped launch the modern media reform movement. A primary team member involved with restoring the **WNET** broadcast signal after all analog and digital transmitters were destroyed at the **World Trade Center**. Last five years, have been Project Director of **Library of Congress NDIIPP** project **Preserving Digital Public Television**.

IN MY SPARE TIME: I make Jewish papercuts in non-traditional designs [www.nanrubin.com]. I produce segments for a weekly radio program on progressive Jewish politics and culture aired on **WBAI**, the **NYC Pacifica** station. [www.beyondthepale.org]. I dabble in handwriting analysis.

FAVORITE BOOKS: *A Distant Mirror*, **Barbara Tuchman**. *The Lymond Chronicles*, **Dorothy Dunn**. *Mass Communications and the American Empire*, **Herbert Schiller**. *The Rabbi's Cat*, **Joann Sfar**.

PET PEEVES: Lima beans.

PHILOSOPHY: “You don’t need to be Jewish to love Levy’s.” (real Jewish Rye bread...). “Any sufficiently advanced technology is indistinguishable from magic” – **Arthur C. Clarke**.

MOST MEMORABLE CAREER ACHIEVEMENT: Signing my first community radio station on the air in Cincinnati. Being invited on the **Martha Stewart Show** to demonstrate papercutting techniques. Being invited to make presentations on digital preservation to the **Blue Ribbon Task Force for Digital Sustainability**, and to the **National Library of Medicine**.

GOAL I HOPE TO ACHIEVE FIVE YEARS FROM NOW: Help community media win significant access to the digital spectrum. Help bust open the copyright stranglehold on access to archival video. Develop activities to train volunteers to take on distributed cataloging. Create a new **Foundation for the Preservation of Television**, alongside the existing **Foundation for Film Preservation** and **Foundation for Preservation of Recorded Sound** chartered by Congress ten years ago.

HOW/WHERE DO I SEE THE INDUSTRY IN FIVE YEARS: I’d like to see a better understanding of the importance of digital preservation, an ongoing commitment from the system to support preservation and access, improved open source tools to manage files, and more robust outlets for archival video. But given the very uncertain future of funding for public media, at this moment it’s very hard to know what direction things will actually take. 🌱



From Dark Archive to Open Access: CLOCKSS Trigger Event Lessons

by **Victoria Reich** (Director, LOCKSS Program, Stanford University Libraries, 1450 Page Mill Road, Palo Alto, CA, 94304; Phone: 650-725-1134) <vreich@stanford.edu> <http://www.lockss.org>, <http://www.clockss.org>

The **CLOCKSS¹ Archive** is singular among digital Archives. Its governing board is comprised equally of librarian directors and publisher directors. Every institution supporting the archive has a seat on either the board or the Advisory Council. The archive is global, with archive nodes — libraries that are holding, and actively preserving, the physical digital bits on a server called a “**CLOCKSS box**.” These nodes are situated across four continents: in the United States and Canada; Scotland; Hong Kong and Japan; and Australia.

In 2008, two **SAGE Publication** titles, *Graft* and *Auto/Biography* were triggered from the archive and made available to everyone on the Web, at no charge.

What is a Trigger Event?

The **CLOCKSS Archive** formally defines a trigger event as follows:

A Trigger Event occurs when either the owner of all rights to the content gives unconditional consent to the release of such content to the general public, or the content is determined in good faith by the board to be unavailable from any publisher for at least six consecutive months and there are no successor interests or reversions or transfers of rights known to the board at the time of the determination.

Trigger Events include, but are not limited to, situations of non-availability of archived content in which:

1. **Publisher No Longer in Business.** Publisher is no longer in business or is no longer in the business of publishing Content or providing access to previously published Content and there are no successor interests or reversions or transfers of rights.
2. **Title No Longer Offered.** Publisher has stopped publishing and is no longer providing access to the Content and there are no interests or reversion or transfer of rights.
3. **Back Issues No Longer Available.** Publisher has stopped offering or providing access to some or all of the back issues of the Content and there are no successor interests or reversion or transfer of rights; or
4. **Catastrophic Failure.** While still publishing Content, Publisher is not able to provide access to the Content electronically due to technical or similar catastrophic and permanent failure.

The First Trigger Event

Three volumes of *Graft* are preserved in the **CLOCKSS Archive** (from 2001 to 2003).

When **SAGE** announced this title would no longer be available from the **HighWire** hosting platform, the board voted to approve the trigger event and agreed to make the content freely available. (See <http://www.clockss.org/clockss/Graft>).

The Second Trigger Event

SAGE ceased to publish *Auto/Biography* in 2006; however, **IngentaConnect** continued to host the title until 2008. When **SAGE** announced **Ingenta** would be taking *Auto/Biography* off-line, **CLOCKSS** experienced its second trigger event. Again, with board approval, the content was made available to anyone with a Web browser for free. (See <http://www.clockss.org/clockss/Auto/Biography>).

Carol Richman, **SAGE**'s Director of Licensing, said, “As these titles did not have a viable subscription base, **SAGE** thought it a good opportunity to offer the community real trigger event experiences.” Indeed, the trigger events were an excellent learning experience. They validated the **CLOCKSS** policy of making triggered content available open access; and are providing data about how readers use triggered content.

Why Open Access?

Content is most likely to be triggered when:

- it is not garnering enough subscription revenue or ad revenue to continue to earn money for the publisher, or
- a catastrophic disaster has occurred.

Content without a viable subscription base for the publisher is unlikely to have a viable subscription base for an archive, therefore the **CLOCKSS** board agreed to make content triggered from **CLOCKSS** freely available.

If a catastrophic disaster befalls one or more publishers, it's likely that this disaster will have other wide reaching consequences. The **CLOCKSS** board agreed in times of catastrophic disaster, generosity was an appropriate response and again agreed the content would be available for free.

Titles triggered from the **CLOCKSS Archive**, are assigned a **Creative Commons** license. For *Graft* and *Auto/Biography*, the content is copyright **SAGE** and licensed under a **Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 United States License**. The **Creative Commons** license clarifies how people can use this content. This particular **Creative Commons** license permits users to share (i.e., to copy, distribute and transmit the work) under the following conditions:

- Attribution. You must attribute the work in the manner specified by the author or

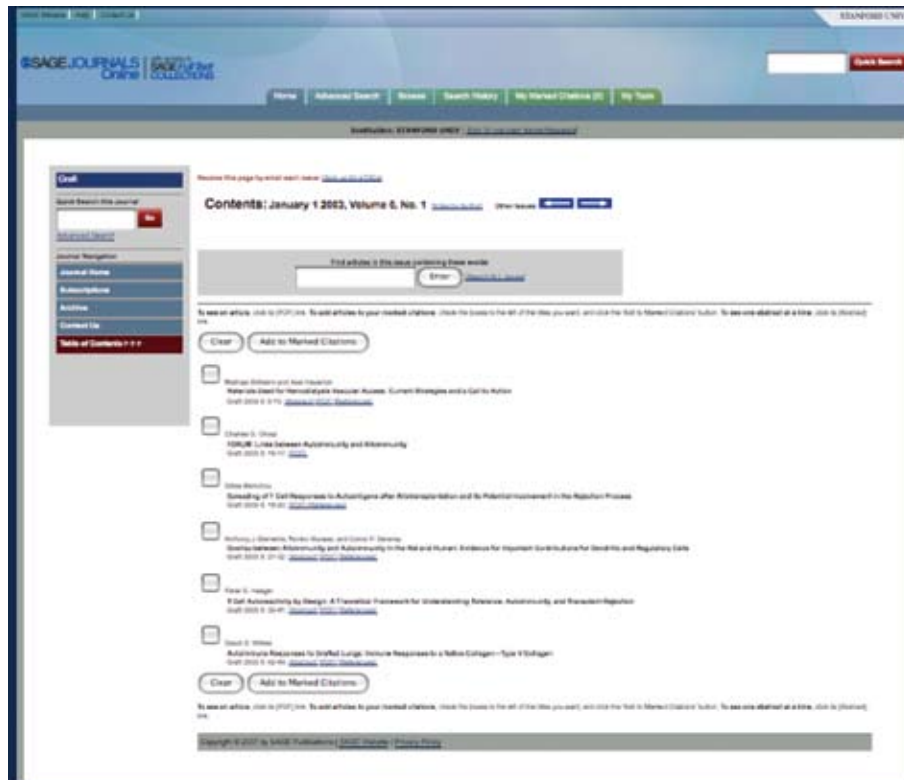


Figure 1: Screenshot of **CLOCKSS** triggered content, *Graft*.

continued on page 26

licensor (but not in any way that suggests that they endorse you or your use of the work).

- Non-commercial. You may not use this work for commercial purposes.
- No Derivative Works. You may not alter, transform, or build upon this work.

The **Creative Commons** license is important for archive interoperability, and the continued preservation of this content. It permits those archives that use a preservation method that preserves the original content — those, which do not alter the content upon ingest — to take this content into that archive with no further negotiation or contact with the copyright holder.

How is it Accessed?

The triggered content was copied from the archive and posted by **CLOCKSS** host institutions, the **University of Edinburgh EDINA** data centre and **Stanford University**. The content is not heavily used. For the month of November 2008, for example, the **Stanford** server for *Graft* delivered about eight URLs per hour. Excluding obvious search engine crawlers, readers downloaded 64 PDFs to 36 distinct IP addresses, of which 31 had domain names that could be found by reverse DNS lookup. Nine were identifiably academic. 28 of the IP addresses found the content via **Google**. Seven of them found the content via the **CLOCKSS** Website. In March 2008, a few months after the initial release of the content, about 7% of the access came via OpenURL resolvers. In November no accesses were recorded.

The *Auto/Biography* content is even less used. In the same month, only two PDFs were downloaded, both by a crawler.

At the time of the trigger event the **CrossRef** DOI resolver could map a DOI to only a single URL. **Portico** claimed the DOI, so it pointed (and still points) users to the **Portico**

against the grain people profile

Director, LOCKSS Program, Stanford University Libraries
1450 Page Mill Road, Palo Alto, CA 94304
Phone: (650) 725-1134 • <vreich@stanford.edu>
www.lockss.org • www.clockss.org
http://www.lockss.org/lockss/Vicky_Reich

Victoria Reich

PROFESSIONAL CAREER AND ACTIVITIES: I've been privileged to work in great libraries with outstanding professionals. I learned how to do reference at the **University of Michigan** and further honed these skills at the **Library of Congress' Science Division**. I was Head of Serials and Acquisitions, first at the **National Agricultural Library** and then at **Stanford University**. My introduction to digital materials, and digital preservation came in 1982. I worked in the Office of the **Librarian of Congress** and one of my many assigned projects was to participate in the "**LC Optical Disk Pilot Project**." It would take twenty years for digital preservation to become a general concern. In the 90s I helped to start **Stanford University Libraries' HighWire Press**, the digital imprint of some of the best STM journals in the world. In 1998, with the help of a brilliant engineer, a small grant from the **NSF**, and the support of **Stanford University** Librarian, **Michael Keller**, we launched the **LOCKSS Program**. Today libraries are using the technology to preserve a broader spectrum of digital materials than we could ever have imagined. Most recently, I joined a two-year pilot called **Controlled LOCKSS**, or **CLOCKSS**. Working alongside some of the brightest and boldest publishers and librarians, we successfully launched **CLOCKSS** as a non-profit organization in October 2008. **CLOCKSS** is the only archive governed and owned by publishers and librarians.

BOOK I AM READING NOW: *Here Comes Everybody: The Power of Organizing Without Organizations* by Clay Shirky

HOW/WHERE DO I SEE THE INDUSTRY IN FIVE YEARS: If you're interested in my thoughts, please email me (vreich@stanford.edu) for a copy of an article I co-authored with **Michael Keller**, and **Dr. David Rosenthal**, "Just in Time" in *Difficult Times: Lessons to be Learned*, to be presented at the Spring 2009 **CNI Spring Task Force** meeting (April 6-9).

copy of the PDF, which is available only to **Portico** subscribers.

The **CLOCKSS** experience led to the discovery that the DOIs for triggered content should be owned and managed by a community

organization, not by a single Archive.

"The availability of *Graft* content in **CLOCKSS** prompted **CrossRef** to create an implementation of **CrossRef Multiple Resolution** since the content was available in more than one archive. The end result is that different archive URLs can be registered with the *Graft* DOIs so that users can easily find all the options available for the content. **CLOCKSS**, **Portico**, the **KB** and **CrossRef** have worked together closely to put a solution in place" — **Ed Pentz**, Executive Director, **CrossRef**.

Who Uses it?

The statistics above, reflecting one month's usage of one of the two servers concerned, show that triggered content gets little use, and that the majority of the use (75% in this case) is not identifiably academic. This is not surprising; the reason the content was triggered was that it was not generating enough use from academic subscribers to justify the costs of making it accessible.

Experience thus validates the decision by the **CLOCKSS** board to make triggered content open access, and the use of the **Creative Commons** license to do so. Charging for users to access the content would likely reduce us-



Figure 2: Screenshot of CLOCKSS triggered content, *Graft*.

continued on page 28

From Dark Archive to Open Access ...
from page 26

age considerably. It would probably eliminate most of the accesses via Google, which are consistently the vast majority.

What Does it Look Like?

Graft was hosted on HighWire Press, and the content was ingested into the archive directly from HighWire Press. SAGE deposited into the CLOCKSS Archive exactly what was published. Hence, the preserved copy is what the readers saw in 2008, the look and feel, the publisher branding is preserved. (See figures 1 thru 3, on pages 24, 26, 28.)

The *Auto/Biography* files ingested into the CLOCKSS Archive were the “pre-publication” files (sometimes called “source files”). The content was not available to the publisher’s Website. These pre-publication files are preserved in the CLOCKSS Archive. To prepare the volumes for the hosting platforms, the content had to be published. The look and feel for this title is not preserved.

What has CLOCKSS Learned?

The *Graft* and *Auto/Biography* trigger events validated the CLOCKSS board decision make triggered content Open Access,

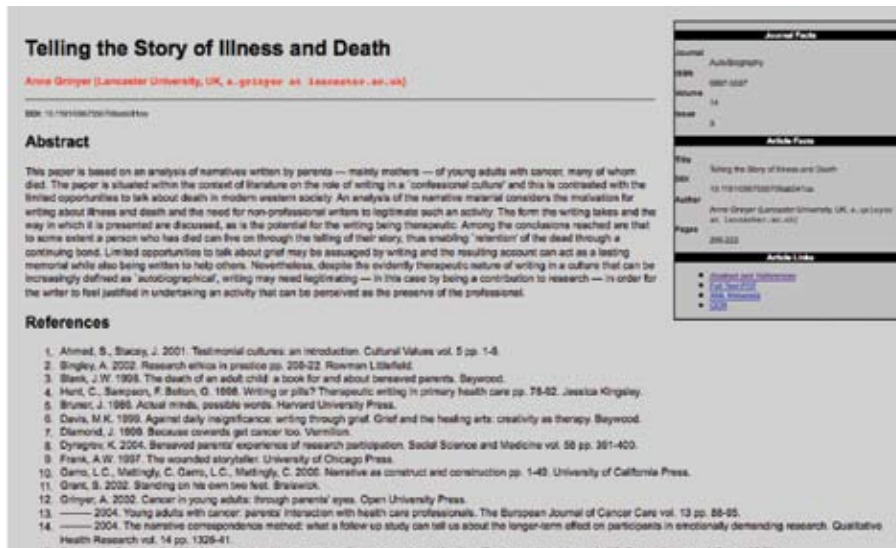


Figure 3. Screenshot of CLOCKSS triggered content, *Auto/Biography*.

accompanied by a Creative Commons license. The Creative Commons license clearly states how this content may and may not be used. As expected, use of this content is relatively low.

<p>Endnotes</p> <p>1. CLOCKSS stands for Controlled LOCKSS (Lots of Copies Keep Stuff Safe).</p>
--

Federal Depository Library Program: Services and Collections

by **James A. Jacobs** (Data Services Librarian, Emeritus, University of California San Diego, 7050 Condon Drive, San Diego, CA 92122; Phone: 858-452-9704) <jajacobs@ucsd.edu> <http://www.freegovinfo.info>

In the age of digital information, libraries and librarians are struggling to define their proper roles. In a time of financial uncertainty and economic crisis, many libraries are facing decisions that will have long-term implications and consequences. More than ever, it is particularly important that we have a clear vision of a sustainable role for libraries.

The issues libraries face can be seen very clearly in a proposal by the **Depository Library Council**, which advises on matters related to the **Federal Depository Library Program (FDLP)**. It has recommended that the **Government Printing Office (GPO)** should “prepare depository libraries for a digital **Federal Depository Library** system that is not centered on collections.” The **Council** is suggesting that government depository libraries should focus on services *instead of* collections.

With this recommendation, the **Council** has reached its own implicit conclusions about the roles of librarians and libraries in society. The **Council** is saying that the role of *librarians* is to provide information services and the role of *libraries* (collections) should be in the hands of **GPO**, the **National Archives (NARA)**, and individual government agencies.

There are at least two reasons that this decision is a troubling one in these tumultuous times. First, it seems counter-intuitive to claim that the best future we can imagine for libraries in the digital age is “libraries without collections.” Second, it is not clear that government agencies have or should have the role that the **Council** wants for them.

The Role of Librarians

An emphasis on service at the expense of collections comes mostly from a view that users are overwhelmed by an information glut and need information professionals to help them navigate a bewildering array of choices. Although this view is a bit paternalistic, implying that librarians know better than users what they need, it is at least based on an understanding of the complex and difficult job of finding the right information on the Web today. In this view, librarianship would be about helping people navigate a complex, networked maze of shifting, changing information. There is nothing wrong with the view that libraries should provide information services and there is in fact much to recommend it, but this service-only model misses a key role for libraries. It is a view of librarians without libraries.

This view assumes an unorganized, undifferentiated Web of information controlled by information providers (e.g., government agencies, commercial vendors, information aggregators, publishers), visible only through the information silos and portals created by those providers. It accepts that libraries will not build digital collections to fit the needs of their users but will simply provide services for information over which librarians have no control.

Librarians, in this view, are valuable precisely because they have no control over information.

This view also accepts that information will be tightly controlled by producers and distributors. What is available, who can use it, under what conditions it may be used, and when it becomes unavailable will be controlled by government agencies, commercial vendors and other “content” distributors.

To me, this suggests that librarians will be analogous to travel agents who, because they deal every day with the complex, difficult, disparate, unconnected systems, are better able than the traveler to navigate these systems and

continued on page 30

find the best flight at the best price. So librarians, in this view, will help casual information users navigate a variety of complex, difficult, disparate, unconnected, public-freely-available and proprietary-and-licensed information systems. Just as travel agents have no control over what flights or trips are available or what they cost or what restrictions are placed on them, so librarians will have no control over what information is available or what it costs or what restrictions are placed on its use.

In this view, librarians will not manage collections but will license the right to read from those who control information. Whether the license comes in the form of designation as an **FDLP** library, or from a contractual “partnership” with **GPO** (which **GPO** is promoting as a substitute for **FDLP** deposit), or from payment to a commercial vendor for a license to access information, or by the granting by the **Google Books** legal department of permission (and restrictions) on use, the result is the same. A recent article in *Library Hi Tech* summarized this view succinctly: “In [the] future, librarians will no longer manage media, they will manage rights.”¹

This view reshapes the role of librarians from information providers to information gatekeepers; from information curators to business officers who sign contracts and pay bills and police contracts for publishers. It is not clear that such a role is either desirable or that it requires a librarian.

The Role of Libraries

Those who believe libraries need not have digital collections apparently assume that, because there is information available on the web, there is no need to duplicate it locally. Librarians should be the first to understand that current availability of any given piece of information does not guarantee its availability or usability in the future. Librarians who understand the difficulty of finding information on the Web today should look to building digital collections to solve these problems rather than playing a never-ending game of catch-up with shifting information and then hoping that users will recognize them as indispensable service providers.

There are many organizations, institutions, and vendors that have information on the Web that they will give or sell to you. But, the word “library” does not mean “I have some information.” If it did, bookstores would be libraries and publishers would be librarians. We need libraries in addition to publishers and bookstores and information vendors and government agencies that distribute information as a by-product of their primary mission. Scholars, journalists, economists, historians, lawyers, physicians, engineers, and citizens of all kinds require a continuing, complete record of information, not just a temporary flow of contemporary information. Who will ensure long-term, free access to the information they need if libraries do not?

The issue we face is not simply understand-

against the grain people profile

Data Services Librarian, Emeritus
University of California, San Diego
7050 Condon Drive, San Diego, CA 92122
Phone: (858) 452-9704 • <jajacobs@ucsd.edu>
freegovinfo.info

James A. Jacobs

PROFESSIONAL CAREER AND ACTIVITIES: Specialist in government information and social science data services.

HOW/WHERE DO I SEE THE INDUSTRY IN FIVE YEARS: My optimistic view is that libraries will identify and act on their role in the life cycle of information by selecting, acquiring, organizing, and preserving information from digital and non-digital sources, and will provide access to and services for that information, each library acting for a specific user-community. Libraries will become the trusted source for users of all kinds of information in all formats. Users will bookmark and tag library copies of information because they know the information is valuable, usable, and will not move. 🐶



ing the role of libraries but also understanding the role of information creators and distributors. For us to assume that producers and distributors will have the same values and ethics and practices as librarians is to confuse the role of producers with the role of curators. In the life-cycle of information, the role of producers ends with users, but the role of libraries begins with users.

It's About Control

Let's be clear. Even in the paper and ink world, libraries and their collections were about wresting control of information from producers and distributors and granting control to local communities and information users. A publisher could take a book out of print, but a library could keep it available. A user could pay for a book or a magazine subscription, but could choose instead to use the information for free at the library. Libraries leveraged economies of scale for the benefit of the community, enabling every community member to have benefits of access to information that no individual could possibly afford.

The need for wresting control away from those who wish to control the access to and the use of information has not changed in the digital world. But the battle lines have changed and we need librarians in the fight to keep free, open, usable access.

“Content providers” want to replace copyright with license agreements. Distributors want to impose **DRM** technologies that tie content to particular technologies that make the information harder to preserve and difficult or even impossible to reuse or repurpose. Producers want to charge for every single use and dictate who can use information, under what conditions, and in what way. In addition, the proliferation of requirements to register to read or use information portends a world

in which people will not have the right of privacy when reading or even when searching or browsing. Governments are not immune to these realities. Governments want to be able to control information they create; they want to be able to alter and even withdraw information after it has been released. Governments increasingly want to view their information as a commodity, which they can use to generate income. And governments are constrained by laws and regulations that prohibit them from “competing” with the private sector, a fact that puts all government information at risk of being constrained by commercial interests.

It is ironic that, given technologies that enable almost unlimited use and re-use of information and that enable information to be distributed and used and re-used almost without cost, we face producers who want to limit access, charge for every use, restrict re-use, and look over our shoulders to see what we're reading. Librarians should be the first to recognize that the interests of readers and user-communities are different from the interests of information producers; libraries and library collections are a way to bridge the gap between the two.

The Optimistic View

Even if one takes an optimistic view and assumes the best intentions on the part of politicians and bureaucrats, it would be irresponsible to assume that government agencies will be able to provide long-term, free public access to information as well as libraries can.

Few government agencies have information access as a primary mission and even those that do face multiple barriers to assuring permanent, free access to usable digital information. The **National Archives** is a prime example. While **NARA** has an explicit mission of making re-

continued on page 32

records available “in perpetuity,” it is constrained by technology, budgets, and recalcitrant agencies. Put simply, it has too much to do and not enough funding to do it. In an honest attempt to deal with these realities, NARA is turning to the private sector to make information more readily available, effectively privatizing the public record. The **Government Printing Office** likes to claim that there has been “a paradigm shift in preservation of depository materials” but you will look in vain in the *GPO Access Act* of 1993 (107 Stat.112), on which it bases these claims, for the words “preservation” or “long-term” or “permanent”. There are good intentions, but no mandate; there are inadequate budgets and no guarantees. Even GPO recognized this in its early policies to implement this “paradigm shift” when it said it would maintain information online only “as long as usage warrants.”

Agencies that have information access as a secondary mission or provide information as a by-product of some other function will not have the inclination, ability, or budget to provide long-term access to their information. And, as the missions of agencies change or are split among new agencies, and as agencies are dissolved or subsumed by other agencies, information will be lost.

But even if one assumes that the government will eventually overcome these problems, there are still other problems. Chief among these is that no one can keep everything forever. Whether it is superseded information, out-of-date information, embarrassing information, expensive-to-keep information, or low-use information that no longer “warrants” keeping, *everyone* will weed something sometime. The question we should be asking is, “Who will be in charge of weeding?”

Society needs different libraries with different collections that respond to the needs of their user-communities (no longer necessarily geographically-based) when making decisions on the value of information. A society without digital libraries will be relying only on federal budget priorities and the market to decide what is worth keeping. Having different collections meeting the needs of different user-communities will better ensure preservation of the information that society as a whole needs. A law library will make different decisions than a medical library and both will make different choices than a library that caters to historians of science. This is a good thing. It builds robustness into preservation and access.

Finally, the e-government movement is reshaping government information policies to be more flexible and interactive. In practice, this means that government will value information transactions more than it values instantiating information in a preservable, re-usable form. Such changes will value current information, but will devalue “out-of-date” information. In such an environment, agencies will find it difficult, if not impossible, to justify preserving last year’s annual report, much less something from ten years or a hundred years ago.

against the grain people profile

Bernard F. Reilly

President, Center for Research Libraries
6050 South Kenwood Avenue, Chicago, IL 60637-2804
Phone: (773) 955-4545 x.334 • Fax: (773) 955-4337
<reilly@crl.edu> • <http://www.crl.edu>

BORN AND LIVED: Born in Philadelphia, PA, lived in Washington DC (1977-1997); Chicago, IL (1997-present).

EARLY LIFE: Curator, art historian.

PROFESSIONAL CAREER AND ACTIVITIES: Research libraries and museums throughout.

FAMILY: Yes.

IN MY SPARE TIME: What spare time?

FAVORITE BOOKS: Conrad, *Heart of Darkness*; Coetzee, *Disgrace*; Franzen, *The Corrections*.

PET PEEVES: Don’t get me started.

PHILOSOPHY: Cynic.

MOST MEMORABLE CAREER ACHIEVEMENT: Growing CRL.

GOAL I HOPE TO ACHIEVE FIVE YEARS FROM NOW: A global CRL.

HOW/WHERE DO I SEE THE INDUSTRY IN FIVE YEARS: Research libraries will still provide essential support to academic research and teaching, but will have a smaller brick and mortar footprint. 🌱

Conclusion

For those who believe that information should just remain in the possession and control of producers and for those who view the Web as a virtual “library,” the idea of digital library collections naturally seems unnecessary and even anachronistic. For those who value long-term, free, public access to information, leaving control of information in the hands of those who will control use, limit access, and charge fees is anathema. If libraries choose to have no digital collections, it will almost certainly result in licensing constraints, DRM constraints, loss of information, loss of free access, loss of usability of information, and more.

Society needs institutions that select that information that deserves preserving from the plethora of information that surrounds us; it needs institutions that then acquire, organize, and preserve that information and that provide trusted, free, privacy-respecting, secure access to and service for that information. Society

needs institutions that have the complete mix of all of these roles as their primary mission (not a secondary mission or a by-product of publishing, or dissemination, or making money). In the case of government information in a participatory democracy it is particularly important, even essential, that society has such institutions. We call them libraries. 🌱

Endnotes

1. Böhner, Dörte. “Digital rights description as part of digital rights management: a challenge for libraries.” *Library Hi Tech* Vol. 26, no. 4 (2008): 598-605 (Accessed on March 20, 2009) <http://www.emeraldinsight.com/10.1108/07378830810920923> Internet.

Rumors from page 22