Distributed Globally, Collected Locally: LOCKSS for Digital Government Information

by **Daniel Cornwall** (Head of Information Services, Alaska State Library, P.O. Box 110571, Juneau, AK 99811-0571; Phone: 907-465-1315; Fax: 907-465-2665) <dan.cornwall@alaska.gov> *http://library.state.ak.us*

and **James R. Jacobs** (International Documents Librarian, Stanford University Library, 123B Green Library, Stanford University, Stanford. CA 94305; Phone: 650-725-1030; Fax: 650-723-9348) <jrjacobs@stanford.edu> *http://jonssonlib.stanford.edu*

Introduction

Ever since the **Government Printing Office (GPO)** brought **GPO Access** online in 1993 in order to make government information accessible on the Web, some librarians and others have dreamt of a system that would extend the mostly successful, 150 year old geographically distributed **Federal Depository Library Program (FDLP)** model to the digital world.[1] Today, thanks to the efforts of **Carl Malamud**, the **Stanford-based Lots of Copies Keep Stuff Safe (LOCKSS)** team, and 15 libraries around the country, a successful model for the digital **FDLP** has been launched.[2]

This paper will describe the **LOCKSS** model of digital preservation and why that model is beneficial to apply to the realm of digital government information. Next, we will illuminate **Carl Malamud's** herculean efforts toward better access to government information. We will then discuss how we've built the **USDocs Private LOCKSS Network (USDocsPLN)** using those documents harvested by **Malamud**. The paper concludes with a call to action.

Benefits of a Distributed Collection

The subject of digital preservation is of vital concern to libraries and other cultural institutions; Organizations like the **Library of Congress**, **U.S. National Archives (NARA)**, **Internet Archive** and many others have been working on solutions to preservation and long-term access to digital information.[3] Within the government documents library community, there is one school of thought that local digital collections of government documents are wasteful duplication of resources.[4] In this view, **GPO's** assumption of storage and preservation duties has freed libraries from the burden of being documents storehouses to let them focus solely on public services. As the **LOCKSS** model demonstrates, this school of thought is mistaken and in fact will endanger long-term access to and preservation of government information.

The **LOCKSS** model[5] is a proven distributed preservation model based on a peer-to-peer (P2P) architecture[6] in which each node in the **LOCKSS** network locally hosts an exact replica of the content being preserved. The open-source **LOCKSS** software then compares content on each host and repairs any differences, thus assuring preservation and authenticity. Approximately 200 libraries in the global public **LOCKSS** network have successfully preserved e-journals and publisher content for over ten years.[7] Fifteen **LOCKSS** libraries have now embarked on a project to apply this successful model to government documents.

There are myriad reasons why a distributed digital preservation system for government information is necessary. Among them are: protection from natural disaster, server outage, etc.; assurance of authenticity; prevention of surreptitious withdrawal or tampering of information; and building local services for local collections.

A system of geographically disbursed digital collections provides resiliency in the aftermath of a disaster. After Hurricane Rita, the **McNeese University Library** in Lake Charles, LA, lost a large amount of their physical collection, including many Louisiana state documents.[8] Imagine that instead of physical documents, **McNeese** had held the ONLY copy of digital documents and that other LA libraries had relied on **McNeese** rather than building their own digital collections. When the hurricane hit and washed away **McNeese's** servers, all libraries in Louisiana would have lost access. Even if **McNeese** followed best practices and kept an offsite backup of their materials, libraries might still be without access for weeks or months while waiting for **McNeese** to come back online.

While this imaginary wipeout of LA state documents did not happen, we face that very real situation with digital federal documents. **GPO** has been tasked since 2001 to provide a mirror server for **GPO Access**.[9] As of this writing, **GPO** has still not done so. If anything happens to **GPO's** servers, we'll lose access to hundreds of thousands of born-digital federal documents.

Local digital collections also insulate against Internet outages and server downtimes. According to the **FDLP-L** archives, **GPO** servers were taken offline seven times in 2007. During those periods, no one could access **GPO's** documents. With a **USDocsPLN** in place, users would not notice down times because they would be automatically rerouted to their nearest collection.

Authenticity, a critical feature to have in any trusted government information infrastructure, is enhanced with a distributed collection. Digital government information has been altered without notice.[10] While there are no documented instances of this happening to **GPO** content, the potential is there as long as **GPO's** servers continue to be the exclusive source for government information. Multiple copies on geographically disparate servers allow possible alterations to be inspected and corrected, thus protecting against deliberate tampering. The **USDocsPLN** explicitly does this. Research suggests that only a large-scale network attack lasting months could successfully change content stored in a **LOCKSS** network.[11]

Related to the problem of alteration is that of outright withdrawal. In the **FDLP** world of distributed physical collections, there are processes in place to protect against this. In order to withdraw a publication from depository collections, **GPO** must notify the holding libraries of the item to be withdrawn and order them to either return the publication to **GPO** or destroy it. Sometimes withdrawal is appropriate and libraries comply.[12] But in some instances, publications are withdrawn needlessly or explicitly to protect the government's reputation. In these instances, depository librarians have been known to create a loud hue and cry that usually results in the withdrawal order being cancelled.[13] In the current centralized digital model, this protection does not exist. No public process need be followed. A simple delete command is all it takes. A cached copy can sometimes be found in **Google** or the **Internet Archive's Wayback Machine**, but often not.

Besides the preservation aspects, building local digital collections can serve to provide unique services for local communities. For instance, text mining is becoming a useful way of analyzing documents either one at a time or in large collections. It could be as simple as a tag cloud of a speech[14] or as complex as analyzing patent applications.[15] Local digital collections could provide researchers with a full or selected amount of **GPO Access** to analyze without requiring access to **GPO** servers that could potentially impact security or performance. Those collections could also be repurposed and remixed to facilitate new ways of analyzing information and creating new bodies of knowledge.

continued on page 43

Distributed Globally ...
from page 42

Libraries Need a Little Help From Their Friends

Libraries have traditionally taken an active role in collecting content to meet the needs of their local user base. This was a straightforward process in the print world, with vendors galore and, in the case of U.S. government documents, the **FDLP**. In the digital world, things are much murkier, the process a little more convoluted. The responsibility to collect and preserve content remains but the process is more challenging; on the open Web, there are no vendors to pull together disparate publishing streams or depository systems for easy inclusion into local library collections. On the Internet, libraries need to implement a more aggressive approach toward collecting Web-based materials as well as identifying new partners in their efforts — libraries must rely on the kindness of strangers and library fellow travelers.

One such fellow traveler to the government documents community is **Carl Malamud**.[16] **Malamud** is an Internet- and open government activist who runs the Website, *public.resource.org*. Since the U.S. government has been producing digital public domain government information, **Malamud** has been successfully shaking it free from government control and onerous access fee structures and making it more accessible to citizens. **Malamud's** overarching goal is to release government information into the open so that others can build more advanced interfaces and facilitate better access to the workings of our governments.[17]

His first campaign led to the creation of the **Securities and Exchange Commission's EDGAR** database of **SEC** filings and corporate disclosure documents (which has recently had a name change to **IDEA**).[18] He has since, in his efforts to "open source America's operating system,"[19] set his sights on Federal and State Courts and case law, State and municipal codes, **U.S. Copyright Office**, **National Technical Information Service (NTIS)** videos, **Government Accountability Office (GAO)** legislative histories, and, of most interest to Government Documents Librarians participating in the **FDLP**, documents from the **GPO**.

The **GPO** is the official publisher of the U.S. Government and manages the **FDLP**. They publish and distribute to libraries publications from 21 federal agencies as well as such integral publications as the Federal Register, Congressional Record, Congressional Reports, Bills, documents and Hearings, Public Laws, Papers of U.S. Presidents and much more. **GPO Access** is built on an older technology called WAIS with a very primitive user interface and limited search capabilities. For that reason, **Malamud**, with the assistance and cooperation of the **GPO**, harvested **GPO Access** documents from **GPO** servers in late 2007 and made them accessible/downloadable via BitTorrent, Rsync, HTTP and FTP. Those documents comprise 200+ gigabytes of data from 1991-2007 amounting to 5,177,003 PDF pages, 54,600 GAO Reports, 448,496 Congressional Reports and more. It's these **GPO** documents upon which the **USDocsPLN** has so far focused.

Current Status

The **USDocsPLN** is now up and running. The 200+ gigabytes of digital documents have been downloaded from **Malamud's** site (*http://bulk.resource.org/gpo.gov*) and distributed among the 15 partners in the project, where they will be preserved within the **LOCKSS** network. This was an extremely cost-effective project as 1 terabyte (which equals 1,000 gigabytes) of storage is now below $200, hardware is typically less than $1,000, and there is only minimal administrative cost once the **LOCKSS** box has been configured. The group will continue to evaluate and add to the network other

continued on page 44

against the grain

people profile

Head of Information Services, Alaska State Library
P.O. Box 110571, Juneau, AK 99811-0571
Phone: (907) 465-1315 • Fax: (907) 465-2665
<dan.cornwall@alaska.gov>
http://library.state.ak.us http://freegovinfo.info

Daniel Cornwall

BORN & LIVED: Born in Los Angeles, California. Lived in San Antonio, Texas and Panama City, Florida. For past ten years I've lived in Juneau, Alaska.

FAMILY: Wife, **Louise**, father in California, four brothers around the country and one sister in Canada.

PROFESSIONAL CAREER AND ACTIVITIES: Worked at **Alaska State Library** for ten years, promoted twice. Member of **Alaska Library Association** and **American Library Association**. Frequent presenter at state conferences. Heavy user and promoter of the **ALA** government documents wiki. Project coordinator for state agency databases across the 50 States. Co-blogger/administrator of free government information.

IN MY SPARE TIME I LIKE: I enjoy reading, blogging, hiking, cooking and outdoor astronomy when Juneau's weather permits it.

FAVORITE BOOKS: *The Mage Storms* series by **Mercedes Lackey**, *The Post-American World* by **Fareed Zakaria**, *A Vow of Conversation* by **Thomas Merton**, *Physics for Future Presidents* by **Richard A. Muller**.

PET PEEVES/WHAT MAKES ME MAD: Accusations without evidence (by governments), statements that suggest a lack of basic knowledge, lack of self-consistency, and yes, librarians who buy into the "travel agent" theory of librarianship w/o control of critical resources.

PHILOSOPHY: The less secrecy, the better for everyone. Do as you would have others do to you. Ask nothing of others that you are not prepared to do yourself.

MOST MEANINGFUL CAREER ACHIEVEMENT: Getting Alaska state agency monographs into the **LOCKSS** system back in 2005.

GOAL I HOPE TO ACHIEVE FIVE YEARS FROM NOW: Working with others in my state, I hope that most Alaskans will become aware of the rich feast of state and federal government information that is available for their taking.

HOW/WHERE DO I SEE THE INDUSTRY IN FIVE YEARS: I see librarianship as a thriving, user-centered profession. Librarians will learn how to bankroll the trust they are given by their brick and mortar patrons and become trusted online guides. They will also be curators of specialized local collections in all media for the education and convenience of their patrons. I also see public and special librarians following their academic cousins and giving much more instruction about resources and research than they do today. This instruction will be where the patron is, whether in-person or online. This will happen because libraries will gain fund as they are seen as transformational places.

Distributed Globally ...
from page 43

collections of digital government documents, including, but not limited to, other collections on *public.resource.org*.[20]

Participating libraries in the **LOCKSS-USDocs** private network include:

- **Alaska State Library**
- **Amherst College**
- **Georgia Institute of Technology**
- **Library of Congress**
- **Michigan State University**
- **North Carolina State University**
- **Northeastern University**
- **Rice University**
- **Stanford University**
- **University of Alabama**
- **University of Illinois/Chicago**
- **University of Kentucky**
- **University of Wisconsin-Madison**
- **Virginia Tech**
- **Yale University**

While it's exciting to have this large group of research libraries participating in the **US-DocsPLN**, we realize that the cost of being a **LOCKSS Alliance** member may be a barrier for some libraries — fees range from $1,000 to $10,800 per year, depending on institution size. We are working to increase the number of **LOCKSS Alliance** members in order to distribute software and other development costs across a larger network. More members mean less cost per institution.

How You Can Help

The preservation of federal documents is too important to be left to the federal government alone; we have the makings of a viable system to preserve digital government publications. There are several ways you can help.

- **Join our private LOCKSS Network.** Join the **LOCKSS Alliance**, get a server for under $1,000, and contact us. The more servers in the **USDocsPLN**, the merrier.
- **Notify us of collections of electronic federal documents. LOCKSS** staff can show you how easy it is to allow **LOCKSS** to ingest and preserve your materials.
- **Attack the root problem.** Demand members of **Congress** legislate and *fund* a system that will ensure that **GPO** proactively deposits publications and data through the **FDLP** and other interested partners. While the **USDocsPLN** project is a good start and an excellent ad-hoc effort, it should be the government's responsibility to put information in the hands of taxpayers. We should not have to be prying it out of the government's hands. A distributed digital **FDLP** benefits everyone.

endnotes on page 45

against the grain

people profile

International Documents Librarian, Stanford University Library
123B Green Library, Stanford University, Stanford, CA 94305
Phone: (650) 725-1030 • Fax: (650) 723-9348
<jrjacobs@stanford.edu> *http://jonssonlib.stanford.edu*
http://freegovinfo.info *http://radicalreference.info*
http://questioncopyright.org

James R. Jacobs

BORN & LIVED: Englewood, NJ. See bio at *http://freegovinfo.info/about/jrjacobs* for more.

EARLY LIFE: Northeast states; lots of soccer, tennis, baseball etc.

FAMILY: Spouse, mother/father in PA, youngest of four siblings (brother in NYC, sister in Cleveland, OH, sister in Groton, NY).

PROFESSIONAL CAREER AND ACTIVITIES: I've worked in libraries since I was 15 when I worked in a small public library in Homer, NY. Professionally, I've been a Government Documents Librarian since 2002, first at **UC San Diego** and now at **Stanford University**. I'm active in **ALA's Government Documents Roundtable (GODORT)** and am a moderator for **govdoc-l**, the primary listserv of government information librarians.

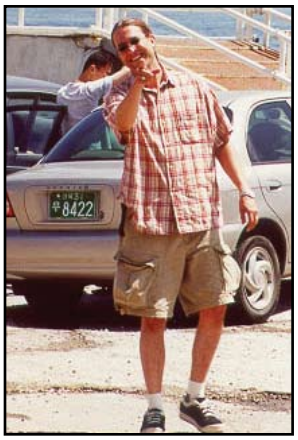
IN MY SPARE TIME I LIKE: Spare time? J information activist/blogger with *http://freegovinfo.info*, *http://radicalreference* and *http://questioncopyright.org*. I also like to dabble in open-source software, hike (urban and rural), eat good food, and read.

FAVORITE BOOKS: *Sometimes a Great Notion*, *Lord of the Rings*, *Baroque Cycle*, *Leaves of Grass*, *Dharma Bums*, *Another Roadside Attraction*, *Tao Te Ching*, *Cat's Cradle*, *People's History of the United States*.

PET PEEVES/WHAT MAKES ME MAD: People who say, "it can't be done" instead of imagining the possibilities; people who act selfishly.

PHILOSOPHY: Information wants to be free; librarians to facilitate that process.

MOST MEANINGFUL CAREER ACHIEVEMENT: Writing "Government Information in the Digital Age: The Once and Future Federal Depository Library Program" which has had over 15,000 downloads; building Radical Reference and Free Government Information to give free reference to activists and independent journalists and advocate for access to and preservation of digital government information.



GOAL I HOPE TO ACHIEVE FIVE YEARS FROM NOW: That a large number of libraries have the technical and administrative wherewithal to be building local digital collections, sharing with each other and building services to increase access and shine light on government activities.

HOW/WHERE DO I SEE THE INDUSTRY IN FIVE YEARS: I'm an optimist. I see libraries continuing their vital work of preserving and giving free access to society's vital information in all formats. I also see them expanding their trusted position by leveraging the Web to make more information available to more people.

Rumors
from page 40

continued on page 47