Stanford helps to digitally preserve mountains of documents

June 15 2010, By Cynthia Haven

Each year, the U.S. Government Printing Office publishes mountains of paper documents, everything from the Congressional Record to Government Accountability Office reports. But that's only a fraction of its output nowadays. More and more of its content is online-only, and that's a problem.

Cyberspace records are vulnerable to computer crashes; they're also vulnerable to tampering. This week Stanford has joined the effort to protect government documents electronically through LOCKSS (Lots of Copies Keep Stuff Safe), a Stanford-based international consortium of more than 200 university and college libraries that collects and preserves electronic content.

In the latest development, on Monday, June 14, the U.S. Government Printing Office (GPO) announced that it has entered the LOCKSS alliance.

The GPO is the federal government's resource for gathering, cataloging, producing, providing, authenticating and preserving published U.S government information.

For the last 15 years, the GPO has been trying to centralize all its information online - culminating in last year's launch of the Federal Digital System (www.fdsys.gov) to give the public a one-stop site to authentic, published government information.

But what happens if the government servers or databases crash? What if some Manchurian candidate or renegade government agency changes U.S. documents for its own nefarious (or even ostensibly benign) purposes? How does one save democracy from havoc-making hackers?

In the past, we relied on paper: 1,250 libraries nationwide participating in the Federal Depository Library Program. Congress created the program in 1813 to ensure that the American public had access to its government's information. As the decades passed, however, the amount of printed information kept growing. Libraries welcomed the chance to minimize bulky hard copies as the world went digital.

Potential limitations on transparency

When documents go digital, however, "it could severely limit transparency," said James Jacobs, government documents librarian for Stanford University Libraries, who is leading the LOCKSS-USDOCS Project.

"The more you centralize digital content, the easier it is to change things without anybody knowing. LOCKSS is a safety net. The simplicity and beauty of LOCKSS is that there are lots of libraries which preserve that content," he said.

"It's a transparency issue - libraries provide an added level of trust in access to government information."

It's not necessary to be imaginative to envision an administration tampering with data. According to Jacobs, "It has happened before. From the mid-1980s through the late 1990s, the American Library Association published an annual review of instances where the government didn't want citizens to know about something and consciously obscured the record. It happens more than we would like."

Less Access to Less Information By and About the U.S. Government is online at freegovinfo.info/library/lessaccess. On the site, Jacobs praises the work as an "amazing series … a chronology of efforts to restrict and privatize government information."

LOCKSS will prevent such "editing."

"Here's what LOCKSS does: If something happens with the GPO, if a server or database crashes, people can get the information from the library," said Jacobs. Does that mean that anybody can go into a LOCKSS participating library and access the LOCKSS government information? "Yes and no," said Jacobs. "In all practicality no, but in theory yes.

"It's a complete preservation archive - the content only gets made accessible if the live content goes away."

How LOCKSS works

In other words, if a person goes into the library and looks for a copy of a particular congressional hearing, the reader can look for a hard copy or digital copy first. If none is available, the library would release the LOCKSS version from its preservation archive and put it on a public site.

"We've never really had to test it at LOCKSS - but yes, you could definitely get it, in an hour or so," said Jacobs.

Eighteen participating LOCKSS libraries in the United States have signed up for the GPO program, with one Canadian library also interested.

"We've just started harvesting content from FDsys.gov," said Jacobs.

Eventually, as LOCKSS catches up, it will move to a routine where "every time a new document is published on GPO, we get alerted and automatically harvest it."

"It's a strange world we live in," Jacobs admitted. The world of cyberspace, websites and FDsys "is like a cloud. You can see it there, but it's amorphous, and you can't touch it. We're hoping to preserve the cloud - the dot-gov cloud."

It may be less enduring than commonly assumed: "The proselytizers of the Internet have done a very good job assuring everyone that the Internet means it's around forever," he said. Nevertheless, "There's no real consensus about how to preserve digital bits in the long term. The Internet is 25 years old, Google is 10 years old." The truth is, said Jacobs, that most websites disappear within a few months. A lot more become cyberspace corpses.

"LOCKSS explores the possibilities of long-term preservation," said Jacobs, including even a reconsideration of that humble medium, paper.

"Digital content is much more difficult to preserve. If I put a book on my shelf, it can stay there for 200 years," he said. "Paper takes a lot longer to disappear than digital bits."

For the time being, anyway, we live in a "hybrid era," said Jacobs. While digital publications provide quick online access, sometimes paper documents - well-indexed, easy to flip through, pointing readers to citations in other places - are simpler to use than digital records.

"Who wants to download and print a 500-page hearing?" he asked. Yet congressional hearings can run at least that long.

Clearly, long-term data preservation is more complicated and nuanced

than it looks, if that's possible. Jacobs cheerfully admitted, "I'm completely in the weeds, and it's kind of fun."

"This kind of thing is not all that sexy - but it's really important," he said. "Librarians have been doing it for a long time and we want to continue doing it."

Thanks to the new agreement, Stanford University Libraries will be doing it for some time to come.

More information: lockss.stanford.edu/

Provided by Stanford University