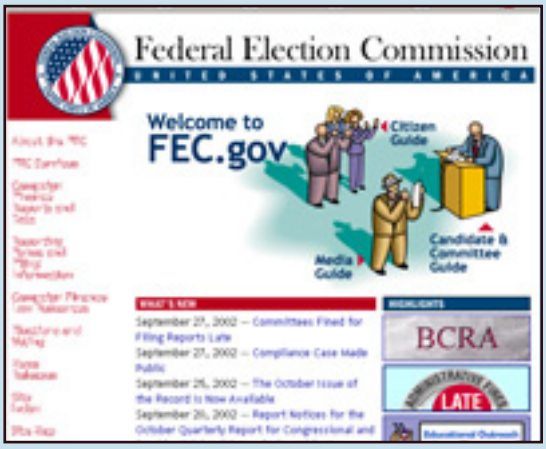
Library of Congress Digital Preservation Newsletter

Preserving Government Web Sites at the End-of-Term



Example Web site from 2002, archived in the Library of Congress Web Archives.

When a new president takes office new policy directions often replace old ones and government agencies must quickly switch gears to serve the public in new ways. Government Web sites often change dramatically in the first crucial weeks of a transition. The Library of Congress and several key partners will preserve public United States Government Web sites at the end of the current presidential administration. "Digital government information is considered at-risk, with an estimated life span of 44 days for a Web site. This collection will provide an historical record of value to the American people," said Director of Program Management Martha Anderson of the Library of Congress' National Digital Information Infrastructure and Preservation Program (NDIIPP). The Internet is one of the primary ways the United States Government interacts with the public. The intent of this project is to save at-risk resources for future historians and scholars, in addition to building on existing election and government Web site collections.

The Internet Archive will undertake a comprehensive crawl of the .gov domain. The Library of Congress has been preserving congressional Web sites on a monthly basis since December 2003 and will focus on development of this collection for the project. The University of North Texas and the California Digital Library, who are also involved with the NDIIPP Web-At-Risk initiative, will focus on in-depth crawls of specific government agencies. The project will also call upon government information specialists -- including librarians, political and social science researchers, and academics -- to assist in the selection and prioritization of Web sites to be included in the collection, as well as identifying the frequency and depth of the act of collecting. The Government Printing Office will lend expertise to the curation process along with libraries in its Federal Depository Library Program.

For more information about the End-of-Term Web site preservation project, see the original press release. Government document librarians and others interested in participating in the program, please email eotproject@loc.gov. ▪

JHOVE and the Development of JHOVE2

In late 2003, engineers at Harvard University Library and JSTOR developed an open-source tool called JHOVE (the JSTOR/Harvard Object Validation Environment) to validate file formats. Since its release, JHOVE has gradually gained acceptance worldwide as an essential tool for format validation.

Format validation is crucial to digital preservation and access. If you don't know what a file is, or if its integrity is damaged, you may not be able to read it or hear it or see its content.

JHOVE was designed to process a digital object and determine what the object claims to be (identification), if the object conforms to requirements (validation) and the properties of the object (characterization). When JHOVE finds a file that it cannot validate, it flags the file. Though the process is automated, only a human can decide whether to accept the file as is or try to get a better version.

JHOVE is easy to install and run. Some users embed the JHOVE Java code into their existing system and integrate it into their digital-preservation workflow.

As adoption spread so did awareness of the original tool's limits. "We came to realize a number of

Stephen Abrams of the California Digital Library

shortcomings," said Stephen Abrams of the California Digital Library and one of the developers. "Some things we now know we could've done better and some things we just didn't have the opportunity to do."

Equipped with a new set of requirements, and with support from the Library, Abrams, and colleagues at Portico and Stanford began work on JHOVE2. Their goals are to:

- Change the existing JHOVE architecture to get better performance, enable more simplified system integration, and encourage third party development and enhancement
- Provide significant new functions
- Implement existing and new functionality

The team is in the requirements gathering and design phase.

The terminology in JHOVE2 has changed a bit from JHOVE. Identification and validation are the same but characterization is now called feature extraction, which Abrams explains as, "Being able to examine formatted objects and extract and report on their salient internal properties."

And a new function, assessment, will determine acceptability under local policy rules. Even if a file is not perfect, the assessment feature can tell you if it's good enough to keep.

JHOVE2 will be able to process more sophisticated digital objects. Abrams said, "In JHOVE there's an assumption that a single digital object is always manifest in a single file, and is always an instance of a single format." But that isn't always true.

Digital objects are becoming more complex: they are made up of several elements. A TIFF file, for instance, holds raster image data but may also contain an embedded color profile or XMP metadata. The JHOVE2 team wants to break from the assumption of "one object, one file, one format," and accept a more arbitrary model: one object potentially comprised of multiple files in multiple formats.

JHOVE2 will implement modular plugins over a lean code framework. "In JHOVE, you were only able to invoke one module over a single file or set of files," Abrams said. "In JHOVE2 – because over time there will be an expanding number of modules– you'll be able to define a sequence of modules that will be iteratively invoked for each file."

The JHOVE2 team aims to make it easier for others to write plug-ins. "We've had four years of experience and feedback," Abrams said. "We're in a much better position to make something cleaner, easier to understand and better documented."

JHOVE2 promises to be a vital tool for long-term archival maintenance of digital assets. The JHOVE2 team will work closely with an advisory board of diverse stakeholders, a mixture of academic libraries, national libraries and archives, repository projects and international preservation projects. JHOVE2 will be released under an open-source license. Early prototypes should be available in 2009, but the project team will continue work through 2010. ▪

Announcement: Symposium on Mass Storage Systems

A workshop on digital preservation and sustainability will be conducted in conjunction with the 25th IEEE Symposium on Massive Storage Systems and Technologies on September 22 from 8:30-5:00 at the Sheraton Inner Harbor, Baltimore, Maryland. The Library of Congress, National Library of Medicine and the National Agricultural Library will lead a panel to highlight challenges for long-term preservation followed by panels on scientific data collection management and technologies, tools, and standards for long-term preservation of data. For more information about the program and to register, see http://storageconference.org/daps/index.html. ▪

To subscribe to this newsletter, go to https://service.govdelivery.com/service/multi_subscribe.html?code=USLOC&origin=http://www.loc.gov type in your e-mail address, scroll down and click on "Digital Preservation."