

# Current Issues in Digital Stewardship

2013 Depository Library Council Meeting  
and Federal Depository Library Conference  
July 31, 2013



Butch Lazorchak  
Digital Archivist  
Library of Congress  
National Digital Information Infrastructure and Preservation Program  
wlaz@loc.gov  
facebook.com/digitalpreservation  
@ndiipp



## Outline

- Background on the National Digital Information Infrastructure and Preservation Program at the Library of Congress, including:
  - Viewshare
  - Federal Agencies Digitization Guidelines Initiative
  - Personal Digital Archiving
  - Community Building and the National Digital Stewardship Alliance
- Digital Stewardship issues of special concern to the Depository Library Community:
  - File Formats for Long-term stewardship
  - Preserving Digital Legislative Information
  - Web Archiving
  - Digital Geospatial Data (very briefly!)

## Repository of the Past?



Predictor of the future?

## National Digital Information Infrastructure and Preservation Program



**MISSION:** Ensure access over time to a rich body of digital content through establishment of a national network of partners committed to selecting, collecting and preserving at-risk digital information.

<http://www.digitalpreservation.gov>

## NDIIPP Focus Areas



- Content
- Partnerships: Government, Industry, Academia
- Technical Infrastructure
- Sustainability
- Outreach and Education

## Some Initiatives

- Viewshare
- Federal Agencies Digitization Guidelines Initiative
- Personal Digital Archiving
- Community Building and the National Digital Stewardship Alliance
- Outreach and Communication

## Technical Infrastructure: Viewshare

The screenshot displays the Viewshare website interface. At the top, the Viewshare logo is visible with the tagline "interfaces to our heritage". Below the logo, the title "Desegregation of Virginia Education (DOVE)" is shown. A sub-header indicates the data was created by DOVE on Nov. 17, 2011, and based on the "Desegregation of Virginia Education (DOVE) - Nov 2011" data set. Navigation options for "Inspect" and "Embed" are present. A sidebar on the left lists repositories such as "Albert and Shirley Small Special Collections Library" and "American Friends Service Committee". Below the sidebar, a "Geographic school" filter is shown with options for "Alexandria, Arlington, Fairfax County, VA" and "Arlington County and Norfolk". The main content area features a map of Virginia with 126 data points plotted as circles, and navigation tabs for "LIST", "MAP", and "TIMELINE".

Viewshare is a free platform for generating and customizing views, (interactive maps, timelines, facets, tag clouds) that allow users to experience your digital collections. <http://viewshare.org/>

## Federal Agencies Digitization Guidelines Initiative (FADGI)

- Started in 2007, FADGI is a collaborative effort by federal agencies to define common guidelines, methods, and practices for digitizing historical content. As part of this, two working groups are studying issues specific to two major areas, Still Image and Audio Visual.
- Digital Imaging Standards
- Format Considerations in Audio Visual Preservation Reformatting
- <http://www.digitizationguidelines.gov/>

## Personal Digital Archiving

Guidance to the “general public” on saving their digital stuff: documents, photos, music, video, email, websites etc.

- Public Events
- Further How-to's and tutorials
- Preservation Week events such as “Personal Archiving Day”
- Co-Hosting Personal Digital Archiving 2014 in Indianapolis



<http://www.digitalpreservation.gov/personalarchiving/>

## Community Building



- Preserving.exe: Toward a National Strategy for Preserving Software 2013
- Personal Digital Archiving 2013
- Designing Storage Architectures for Preservation Collections
- Science at Risk: Toward a National Strategy for Preserving Online Science 2012
- Upcoming Meetings on History and Social Science
- <http://www.digitalpreservation.gov/meetings/>

## National Digital Stewardship Alliance



- <http://www.digitalpreservation.gov/nds>
- <https://twitter.com/NDSA2>
- [http://www.loc.gov/extranet/wiki/osi/ndiip/nds/index.php?title=Main\\_Page](http://www.loc.gov/extranet/wiki/osi/ndiip/nds/index.php?title=Main_Page)

## National Digital Stewardship Alliance



### NDSA

- An initiative of the National Digital Information Infrastructure and Preservation Program (NDIIPP)
- A collaborative effort among government agencies, educational institutions, non-profit organizations, and business entities to preserve a distributed national digital collection for the benefit of citizens now and in the future.
- Community-driven, action-oriented

## NDSA Membership



- Over 150 member organizations have joined the NDSA since it was founded in July 2010.
- Examples: Federal Library and Information Center Committee (FLICC), U.S. Fish and Wildlife Service, Northeast Document Conservation Center, National Park Service, New York Public Library, National Library of Medicine, Thomson Reuters, Academy of Motion Picture Arts and Sciences, California Digital Library, Harvard University, NARA, OCLC, Public Broadcasting Service, many more.
- Membership is now open to any interested U.S.-based organizations. Very lightweight membership agreement!

## Benefits of Participation

- Have a Voice
- Share knowledge
- Drive action
- Learn from Experts
- Impact change



## NDSA Governance

Eight Member Coordinating Committee drawn from all sectors.

Members commit to participating in one of the five working groups:

- Content
- Infrastructure
- Innovation
- Outreach
- Standards and Practices



## NDSA Working Group Activities



- Preservation Levels
- PDF/A white paper
- Cloud Presentations
- Innovation Awards
- Geospatial Content Appraisal white paper
- Digital Preservation in a Box
- Kickstarter Curator Page
- Standards and Best Practices Knowledge Base



## Outreach



- <http://www.digitalpreservation.gov>
- Monthly newsletter (19,100+ subscribers)
- Twitter: [twitter.com/ndiipp](https://twitter.com/ndiipp) or @ndiipp (13,700+ followers)
- Facebook: [facebook.com/digitalpreservation](https://facebook.com/digitalpreservation) (5,300+ likes)
- Blog: <http://blogs.loc.gov/digitalpreservation/>
- Videos
- Podcasts
- National Book Festival
- Printed brochures
- Smithsonian preservation events
- Preservation Week/Day Events

# And Wait! There's More!



<http://ptup.editme.com/files/2007-SJ-Monterey-Tour/TheWholeEnchilada.jpg>

## A Guide to Distributed Digital Preservation

Edited by Katherine Skinner  
and Matt Schultz



Educopia Institute  
MetaArchive Cooperative Publications

- Distributed Architectures
- Technical Considerations
- Organizational Considerations
- Content Selection, Preparation, and Management
- Content Ingest, Monitoring, and Recovery
- Network Administration

<http://www.metaarchive.org/GDDP>

## Digital Stewardship issues of special concern to depository libraries



- File Formats for Long-term stewardship/PDF-A
- Web Archiving
- Preserving Digital Legislative Information
- Digital Geospatial Data
- Distributed Digital Preservation/Cloud Computing

## File Format Sustainability



- Library of Congress Sustainability of Digital Formats: Planning for Library of Congress Collections
- PDF/A



**Sustainability of Digital Formats**  
**Planning for Library of Congress Collections**

Introduction | **Sustainability Factors** | Content Categories | Format Descriptions | Contact

---

**Sustainability Factors**

**Table of Contents**

- [Disclosure](#)
- [Adoption](#)
- [Transparency](#)
- [Self-documentation](#)
- [External dependencies](#)
- [Impact of patents](#)
- [Technical protection mechanisms](#)

**Overview of factors**

In considering the suitability of particular digital formats for the purposes of preserving digital information as an authentic resource for future generations, it is useful to articulate important factors that affect choices. The seven sustainability factors listed below apply across digital formats for all categories of information. These factors influence the likely feasibility and cost of preserving the information content in the face of future change in the technological environment in which users and archiving institutions operate. They are significant whatever strategy is adopted as the basis for future preservation actions: migration to new formats, emulation of current software on future computers, or a hybrid approach.

Additional factors will come into play relating to the ability to represent significant characteristics of the content. These factors reflect the quality and functionality that will be expected by future users. These factors will vary by genre or form of expression for content. For example, significant characteristics of sound are different from those of still pictures, whether digital or not, and not all digital formats for images are appropriate for all genres of still pictures. These factors are discussed in the sections of this Web site devoted to particular [Content Categories](#).

**Disclosure**

Disclosure refers to the degree to which complete specifications and tools for validating technical integrity exist and are accessible to those creating and sustaining digital content. Preservation of content in a given digital format over the long term is not feasible without an understanding of how the information is represented (encoded) as bits and bytes in digital files.

<http://www.digitalpreservation.gov/formats/index.shtml>



## What is PDF /A?

- Some useful features of PDF are incompatible with the demands of long-term preservation
- Preserve the “static visual appearance”
- Self-contained
- Must include embedded fonts
- Must include device-independent color
- Must include XMP metadata
- May not include: Encryption\* (digital signatures ok in V2,3), LZW compression, embedded files\* (now OK in V3), external content references, javascript, etc.

## PDF/A-1, -2, -3



- “Real” name of first version is ISO 19005-1:2005 Document management -- Electronic document file format for long-term preservation -- Part 1: Use of PDF 1.4 (PDF/A-1)
- “Real” name of the second version is ISO 19005-2:2011 Document management -- Electronic document file format for long-term preservation -- Part 2: Use of ISO 32000-1 (PDF/A-2)
- “Real” name of the third version: ISO 19005-3.2 Document management – Electronic document file format for long-term preservation – Part 3: Use of ISO 32000-1 with support for embedded files (PDF/A-3)

## PDF/A Resources



- PDF Association PDF/A Competence Center:  
<http://www.pdfa.org/competence-centers/pdfa-competence-center/>
- LC Sustainability Site PDF/A-1 Page:  
<http://www.digitalpreservation.gov/formats/fdd/fdd000125.shtml>
- D-Lib Magazine: PDF/A: A Viable Addition to the Preservation toolkit  
<http://dlib.org/dlib/november10/noonan/11noonan.html>

# Preserving Digital Legislative Information



The screenshot displays the website for the Center for Archival Resources on Legislatures (CAROL) at the Minnesota State Archives. The page features a header with the Minnesota State Archives logo and a navigation menu on the right. The main content area includes sections for Introduction, Topics of Interest, and Foundations, each with descriptive text. A 'POWERED BY META DATA' logo is visible in the bottom right corner of the screenshot.

**MINNESOTA STATE ARCHIVES**

**CENTER FOR ARCHIVAL RESOURCES ON LEGISLATURES (CAROL)**

**Introduction**

The Center for Archival Resources On Legislatures (CAROL) is the final product of the Minnesota NDIPP project related to the preservation of and access to digital legislative content. CAROL pulls together research information and other resources on pertinent topics, while the [Final Report](#) summarizes project activities.

**Topics of Interest**

The resource center is divided into four main categories: Foundations, Access, Preservation, and Authentication. These topics will help you understand your records and responsibilities, options for providing access, methods and tools for preservation, and issues surrounding authentication of legislative and legal materials.

**Foundations**

This portion of the Resource Center covers issues about understanding your records, recognizing your responsibilities, acquiring content, and choosing appropriate formats and standards. These concepts will help you understand and develop strategies for preserving or providing access to digital content. Links to white papers produced by the NDIPP project team as well as outside resources are provided.

State Archives Home  
Research Services  
Government Record Services  
Records Retention Schedules  
Electronic Records Management  
Trustworthy Information Systems  
Recent Acquisitions  
Recordskeeping Metadata Standard  
Center for Archival Resources On Legislatures (CAROL)  
Special Projects  
Contact Us

POWERED BY  
META  
DATA

# Model Technological and Social Architecture for the Preservation of State Government Digital Information (MTSA)

- Authentication Resources
- Authentication White Papers
- Best Practice Principles for Opening Up Government Information
- Business Case for Digital Preservation
- Cloud Computing
- Digital Audio Video White Paper and Resources
- Government Data Mashups White Paper
- Legislative Metadata
- Legislative History Resources
- Options for Improving Access to Legislative Records
- Project Podcast
- Preservation Options
- Record Inventory
- Records Retention Policies for Selected Legislative Records
- Retrospective Digitization White Paper and Resources
- Survey of Partner's Legislative Records on the Web
- Web Archiving and Evaluation
- Web Content Accessibility
- XML Basics
- XML Native Database White Paper
- XML Usage Survey

# UELMA



## • Uniform Electronic Legal Material Act

<http://www.uniformlaws.org/Act.aspx?title=Electronic%20Legal%20Material%20Act>

- If an official publisher publishes legal material only in an electronic record, the publisher shall designate the electronic record as official
- An official publisher of legal material in an electronic record that is designated as official under shall authenticate the record.
- To authenticate an electronic record, the publisher shall provide a method for a user to determine that the record received by the user from the publisher is unaltered from the official record published by the publisher.
- An official publisher of legal material in an electronic record that is or was designated as official under Section 4 shall provide for the preservation and security of the record in an electronic form or a form that is not electronic.

## Authentication: Documents, Not People



### • AALL State-by-State Report on Permanent Public Access to Electronic Government Information (2003)

- <http://www.aallnet.org/Archived/Government-Relations/Issue-Briefs-and-Reports/2003/ppareport.html>

### • AALL State-By-State Report on Authentication of Online Legal Resources (2007)

- [http://www.aallnet.org/aallwash/auten\\_rprt/AuthenFinalReport.pdf](http://www.aallnet.org/aallwash/auten_rprt/AuthenFinalReport.pdf)

### • GPO Authentication as part of FDSys (2011)

- <http://www.gpo.gov/pdfs/authentication/authenticationwhitepaper2011.pdf>

# Authentication of Primary Legal Materials and Pricing Options Paper

- Possible Authentication Methods: Secure Web sites; Document hashes; Digital Signatures; Public Key Infrastructures; more
- [http://www.mnhs.org/preserve/records/legislativerecords/docs\\_pdfs/CA\\_Authentication\\_WhitePaper\\_Dec2011.pdf](http://www.mnhs.org/preserve/records/legislativerecords/docs_pdfs/CA_Authentication_WhitePaper_Dec2011.pdf)

Office of Legislative Counsel



OFFICE OF LEGISLATIVE COUNSEL

Authentication of Primary Legal Materials and Pricing Options

December 2011

**ULC** Uniform Law Commission  
The National Conference of Commissioners on Uniform State Laws

Contact Us: 312.450.6600 [Login](#)

Home Acts Committees Legislation Meetings News About ULC

## Acts

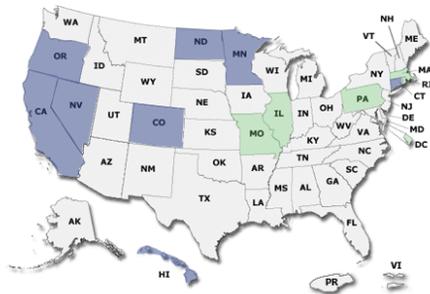
### Electronic Legal Material Act

**Legislative Tracking**

2013 Introductions & Enactments

State	Bill	Sponsor	Status
Connecticut	SB 235		Enacted
District of Columbia	20-221	Wells	Introduced
Hawaii	SB 32	Hee	Enacted
Illinois	SB 1941	Mulroe	Introduced
Massachusetts	HB 38		Introduced
Minnesota	HF 278	Hortman	Enacted
Missouri	SB 478	Lamping	Introduced
Nevada	SB 105		Enacted
North Dakota	HB 1129		Enacted
Oregon	HB 2944	Davis	Enacted
Pennsylvania	SB 601	Scarnati	Introduced
Rhode Island	HB 5850	Hearn	Introduced

**Enactment Status Map**



■ = Enacted   ■ = Introduced this Year

**Act History**

**Origin:** Completed by Uniform Law Commissioners in 2011.  
**Committee:** [Electronic Legal Material Act](#)  
**Resources:** [Draft Materials](#)



EXECUTIVE OFFICE OF THE PRESIDENT  
OFFICE OF MANAGEMENT AND BUDGET  
WASHINGTON, D.C. 20503

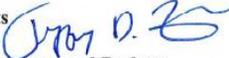


NATIONAL ARCHIVES AND RECORDS ADMINISTRATION  
WASHINGTON, D.C. 20408

August 24, 2012

M-12-18

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND  
AGENCIES AND INDEPENDENT AGENCIES

FROM: Jeffrey D. Zients   
Acting Director  
Office of Management and Budget

David S. Ferriero   
Archivist of the United States  
National Archives and Records Administration

SUBJECT: Managing Government Records Directive

## Legislative Preservation Resources

- Preserving Digital Legislative Information: Wrapping Up the MTSA Project
  - <http://go.usa.gov/ylr>
- NARA Records Express Blog
  - <http://blogs.archives.gov/records-express/>
- Center for Archival Resources On Legislatures (CAROL)
  - <http://www.mnhs.org/preserve/records/legislativerecords/carol/index.htm>
- Drafting Legislation Using XML at the U.S. House of Representatives
  - <http://xml.house.gov/drafting.htm>
- Authentication at the Government Printing Office
  - <http://www.gpo.gov/authentication/>
- Document Naming Conventions: To accompany the House standards for Electronic Posting of House and Committee Documents and Data
  - [http://cha.house.gov/sites/republicans.cha.house.gov/files/documents/committee\\_docs/naming\\_conventions\\_v\\_1\\_2.pdf](http://cha.house.gov/sites/republicans.cha.house.gov/files/documents/committee_docs/naming_conventions_v_1_2.pdf)

# Web Archiving

- LC web archiving activity
- Terminology
- Infrastructure
- Memento

The screenshot shows the Library of Congress website for Web Archiving Projects and Collections. The page features a navigation bar with the Library of Congress logo and the text 'WEB ARCHIVING PROJECTS AND COLLECTIONS'. A search bar is present with the text 'LOC INTRANET' and 'LOC.gov'. The main content area is divided into several sections:

- Web Archiving**: A sidebar with a search box and a list of links including 'Collection Proposal Process', 'DigBoard Tool', 'Training Materials', 'Frequently Asked Questions', 'Selection Criteria Wiki', 'Copyright & Permissions', 'Terms and Definitions', 'LCWEBARCH List', and 'Production Portal for IA Crawls'.
- Web Archiving Projects and Collections**: The main content area, which includes a search box, a paragraph of introductory text, and a list of 'Active Collections' such as 'Brazil Cordel Literature Online', 'Burma/Myanmar Election 2012', 'Civil War 150th Web Archive', 'Congressional and Government', 'Digital Formats Web Archive', 'Egyptian Election 2012', 'Leral Blansa', 'Malaysian Election 2012', 'Manuscript Division Archive', 'Pakistan Nationalisms: Part 1 & 2', 'Performing Arts Collection', 'Public Policy Topics', 'Single Sites', 'Small Press Expo Comic and Comic Art Collection', 'Timor Leste Election 2012', 'Tunisia & North Africa & the Middle East 2011', 'U.S. Election 2012', and 'War of 1812 Bicentennial Web Archive'.
- Highlights & Statistics**: A section with a bulleted list of statistics: '316 terabytes collected', '6.421 billion documents', 'Active Collections Report', and 'Active Statistics & Charts'.
- Archive of loc.gov**: A section with a paragraph of text and a link to 'LCWA: Public Access to Web Archives'.
- All Collections**: A section with a bulleted list of collection counts: 'Collaborative Projects (6)', 'Law Library (5)', 'Library Services (49)', 'OSI (5)', and 'OVOPs (18)'.

At the bottom of the page, the URL <http://www.loc.gov/webarchiving/> is displayed.

## Web Archive Terminology



- **Seed** - the Nominated URL, or starting point for the crawler
- **Seed List** - List of Seed URLs fed to the crawler for harvesting
- **Scoping** - Instructions the team gives to the crawler to include other URLs associated with the seed (for example, crawl of whitehouse.gov would include facebook.com/whitehouse)
- **Crawler**- Software which downloads file, extracts URLs, adds URLs to list and repeats
- **Frequency** - How often a seed is crawled (one time, once a week, once a monthly, twice a year, or annually)

## Web Archiving Infrastructure



- **Crawler: Heritrix**
  - <https://webarchive.jira.com/wiki/display/Heritrix/Heritrix>
- **Access: Wayback Machine**
  - <http://archive-access.sourceforge.net/projects/wayback/>
- **File format: WARC (ISO Standard)**
  - <http://bibnum.bnf.fr/WARC/>
- **Information Resource: International Internet Preservation Consortium (IIPC)**
  - <http://netpreserve.org>

# Memento



## Memento Adding Time to the Web

[About](#) [Demos](#) [Guide](#) [Tools](#)

Memento wants to make it as straightforward to access the Web of the past as it is to access the current Web.

If you know the URI of a Web resource, the technical framework proposed by Memento allows you to see a version of that resource as it existed at some date in the past, by entering that URI in your browser like you always do and by specifying the desired date in a browser plug-in. Or you can actually browse the Web of the past by selecting a date and clicking away. Whatever you land upon will be versions of Web resources as they were around the selected date. Obviously, this will only work if previous versions are available somewhere on the Web. But if they are, and if they are on servers that support the Memento framework, you will get to them.

[Formal technical specifications](#) detailing the Memento framework are available. But if you want a more gentle entry point to gain an understanding of how Memento is trying to change the Web by adding a time dimension to its most common protocol, HTTP, check out the [Introduction to Memento](#).

If you are interested in establishing a Web with a memory, please join the [Memento Development Group](#). And if you want to get involved in the technical discussion of the [Memento specification](#), you probably should subscribe to the [IETF HTTP list](#) or to the [W3C TAG list](#).

<http://www.mementoweb.org/>

# What are the Special Risks to Geospatial Information?

- Unique geospatial data formats
- Spatial database complexity
- Fragility and uncertainty surrounding digital cartographic representation
- Issues related to time-versioned content
- Metadata unavailability or inconsistency
- No generally supported content packaging design for complex geospatial data



The screenshot shows the FGDC website interface. At the top, there is a navigation bar with links for Site Map, Accessibility, and Contact. A search box is also present. Below the navigation bar is a main menu with options like Home, Library, Calendar, and Contact Us. The left sidebar contains a list of categories such as Participants, Data & Services, Standards, Metadata, Framework, Policy & Planning, Training, Grants, International, Geospatial LoB, NGAC, and IFTN. The main content area displays the title "Users/Historical Data Working Group" and a breadcrumb trail: "you are here: home → participation → working groups & subcommittees → historical data working group". The text describes the group's purpose and lists resources like Charter, Membership, Workplans (FY 2006, FY 2007, FY 2008), and Reports (FY 2006).

**NDSA Geospatial Subgroup**

**NDSA WHITE PAPER: ISSUES IN THE APPRAISAL AND SELECTION OF GEOSPATIAL DATA**

*The mapping and drafting section of the Division of Economics and Statistics. Courtesy of the Library of Congress Prints and Photographs Division.*

ABOUT | FAQ

## Geospatial Data Preservation



**Featured Resource**  
[Geospatial Data Transfer, Archival Data Transfer, Validation and Dataset Functional Verification \[Web\]](#)

### QUICK LINKS FOR

- Data Managers
- System Developers
- Researchers

### Featured Practice

## Format Descriptions for Geospatial Data

Explore descriptions of formats used for geospatial data and how to assess them for potential use.



Find resources via free text search on titles, authors, and key terms...

EDUCATION & TRAINING

TOOLS & SOFTWARE

POLICIES & BENEFITS

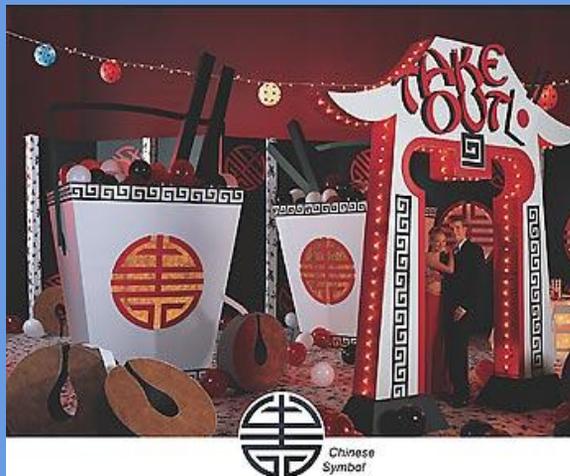
BROWSE BY TOPIC

BROWSE BY TYPE

ADVANCED SEARCH

## Key Takeaways

- NDIIPP and the NDSA are trusted resources for digital stewardship information
- Digital Stewardship is multi-faceted, challenging and complex, but achievable
- We need your help to spread the word



<http://shindigzparty.files.wordpress.com/2008/03/4s058a.jpg>

# Thanks!

Butch Lazorchak  
Digital Archivist  
Library of Congress  
[wlaz@loc.gov](mailto:wlaz@loc.gov)



[http://imagecache2.allposters.com/images/pic/Matted\\_Prints/mp\\_814104\\_b-Man-on-Phone-Thanks-Posters.jpg](http://imagecache2.allposters.com/images/pic/Matted_Prints/mp_814104_b-Man-on-Phone-Thanks-Posters.jpg)