PRESERVATION OF ELECTRONIC LEGAL MATERIALS

UELMA PRESERVATION GROUP

Leah Prescott, chair
Steven Anderson
Erik R. Beck
Diane Boyer-Vine
Daniel Cordova
Susan David DeMaine
Amy Emerson
Emily Feltren
David Greisen
David Hansen
Jason Judt
Jane Larrington
Margaret Maes
Michelle Pearse
Mendora Servin
Anthony Smith
Leslie Street
David Walls

Rev. 4/2018

Contents

Introduction and Brief Background

Formation of ad hoc preservation group

In 2014, the Washington D.C. Counsel's office was looking for a way to preserve the source materials for the Code of the District of Columbia. They reached out to the Georgetown University Law Library to inquire about becoming a member of the Chesapeake Digital Preservation Group - made up of Georgetown Law Library, the state libraries of Maryland and Virginia, and Harvard Law Library - a program created to preserve born digital legal materials. The DC code had been converted to XML documents, and because of the dynamic nature of that system, it was determined that the Chesapeake repository was probably not the best solution.

With the participation of the AALL Government Relations Office, this conversation transformed into a small group of people from other states who were grappling with the same issue of how best to preserve official electronic legal materials - one of the primary requirements of the Uniform Electronic Legal Material Act (UELMA).

The UELMA Preservation Group agreed that the main purpose of the group should be to investigate preservation strategies, and provide tools and assistance to those states that have adopted or plan to adopt UELMA. It was determined that deliverables for the group might be:

- Guidance on what digital preservation entails, and establishing guidelines and best practices
- Trying to educate people on costs related to levels of preservation
- Examples of technical papers and documents that have been developed, documentation and sample language to help with advocacy
- A toolkit and case studies

Survey

The group determined that there was not enough information available about the status of state electronic legal materials, and decided that a survey could provide useful data. After locating contacts in each state, the survey was launched in 2015, with the goal of determining the state of preservation activities for electronic legal materials in general, and to determine what (if any) open tools might be useful for the community as a whole.

The compiled results of the survey can be viewed in Appendix I of this paper, and in general revealed that while legal materials are being created digitally (born digital) as well as being digitized from paper-based materials, many are not yet considered official. In addition, the survey suggested that there is not a strong desire for either a consortial solution, or an open source tool. Consequently, the group decided that the best deliverable at this time would be a guidance document in the form of a white paper - this white paper.

UELMA and Electronic Legal Material

For those who are new to UELMA, the following section will give basic information about the act.

The American Association of Law Libraries (AALL) maintains a webpage of UELMA resources (https://www.aallnet.org/advocacy/government-relations/state-issues/uelma-resources/), and it describes UELMA in this way:

"The Uniform Electronic Legal Material Act (UELMA) is a uniform law that addresses many of the concerns posed by the publication of state primary legal material online. UELMA provides a technology-neutral, outcomes-based approach to ensuring that online state legal material deemed official will be preserved and will be permanently available to the public in unaltered form."

Text of the Act Relating to Preservation

SECTION 7. PRESERVATION AND SECURITY OF LEGAL MATERIAL IN OFFICIAL ELECTRONIC RECORD.

(a) 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.

(b) If legal material is preserved under subsection (a) in an electronic record, the official publisher shall:

(1) ensure the integrity of the record;

(2) provide for backup and disaster recovery of the record; and

(3) ensure the continuing usability of the material.

In terms of what "electronic" and "legal material" means in the context of UELMA, the Act provides the following guidance:

(1) "Electronic" means relating to technology having electrical, digital, magnetic, wireless, optical, electromagnetic, or similar capabilities.

(2) "Legal material" means, whether or not in effect:

(A) state constitution

(B) session laws

(C) state code

(D) a state agency rule that has or had the effect of law

(E) categories of state administrative agency decisions

(F) reported decisions of state courts

(G) state court rules

(H) any other category of legal material to be included by individual states

---

It is a strength of the act that it does not prescribe a technological strategy for electronic documents, thereby allowing for a full range of solutions to deal with a full range of digital format types. That flexibility is also a challenge when trying to develop a standard set of solutions for the widest possible set of users. The strategies used by the states that have thus far enacted UELMA however, fall into only a few categories, and this paper provides case studies from some of those states that will illustrate their strategies.

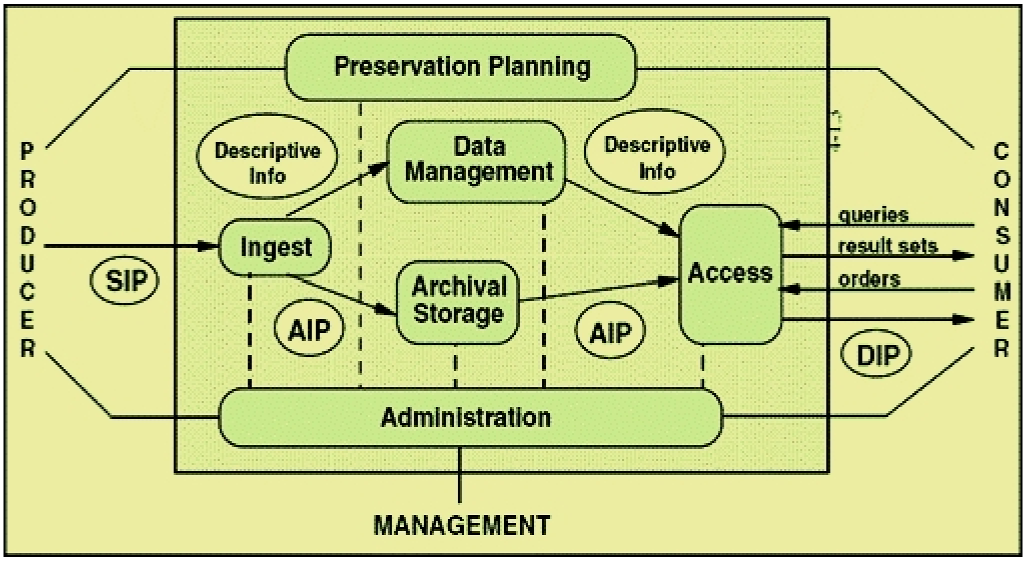
Digital Preservation

There are those who interpret digital preservation to be the act of protecting paper-based materials through digitization, but there is increasing recognition that a digital object - regardless of whether it was born digital or created through a digitization process - is an object separate from any analog form, and as such has a separate set of capabilities, a separate set of preservation challenges, and an equal need to be preserved.

The National Digital Stewardship Alliance (NDSA) defines digital preservation as "The series of managed activities, policies, strategies and actions to ensure the accurate rendering of digital content for as long as necessary, regardless of the challenges of media failure and technological change."

OAIS (Open Archival Information System) Functional Model - *ISO 14721*

Many preservation systems refer to the OAIS Functional Model to frame the range of capabilities that they offer, so it is useful to give a brief explanation about what it is. OAIS is a theoretical model rather than an actual system structure and as such, it describes the basic functions that a compliant system must perform. Graphics such as this one from Wikipedia are often used to represent the model:



(https://en.wikipedia.org/wiki/Open_Archival_Information_System)

Preservation system descriptions will often refer to SIPs, AIPs, and DIPs and these acronyms come from the OAIS model. They refer to:

- SIP - Submission Information Package; the information coming into the system
- AIP - Archival Information Package; the archival objects and data created and packaged from the SIP
- DIP - Dissemination Information Packages); the object and data that is created from the AIP and made available through an access system.

Trusted Digital Repositories

In a 2002 report, OCLC (a global library cooperative) defines a Trusted Digital Repository (TDR) as "one whose mission is to provide reliable, long-term access to managed digital resources to its designated community, now and in the future." A related standard that applies to digital preservation is the Trustworthy Repositories Audit & Certification (TRAC) checklist (ISO 16363). The purpose of this checklist is to define the criteria for certification of a repository system as a TDR.

The metrics of the checklist are split into three topical areas:

- Organizational Infrastructure - the repository's administrative, staffing, financial, and legal functions
- Digital Object Management - the handling of digital objects from ingest to access
- Technology, Technical Infrastructure, and Security - the technology used to handle ingested objects

The OCLC report further summarizes the responsibilities of a TDR. It states that a TDR must:

- Accept responsibility for the long-term maintenance of digital resources on behalf of its depositors and for the benefit of current and future users
- Have an organizational system that supports not only long-term viability of the repository, but also the digital information for which it has responsibility
- Demonstrate fiscal responsibility and sustainability
- Design its system(s) in accordance with commonly accepted conventions and standards to ensure the ongoing management, access, and security of materials deposited within it
- Establish methodologies for system evaluation that meet community expectations of trustworthiness
- Be depended upon to carry out its long-term responsibilities to depositors and users openly and explicitly
- Have policies, practices, and performance that can be audited and measured

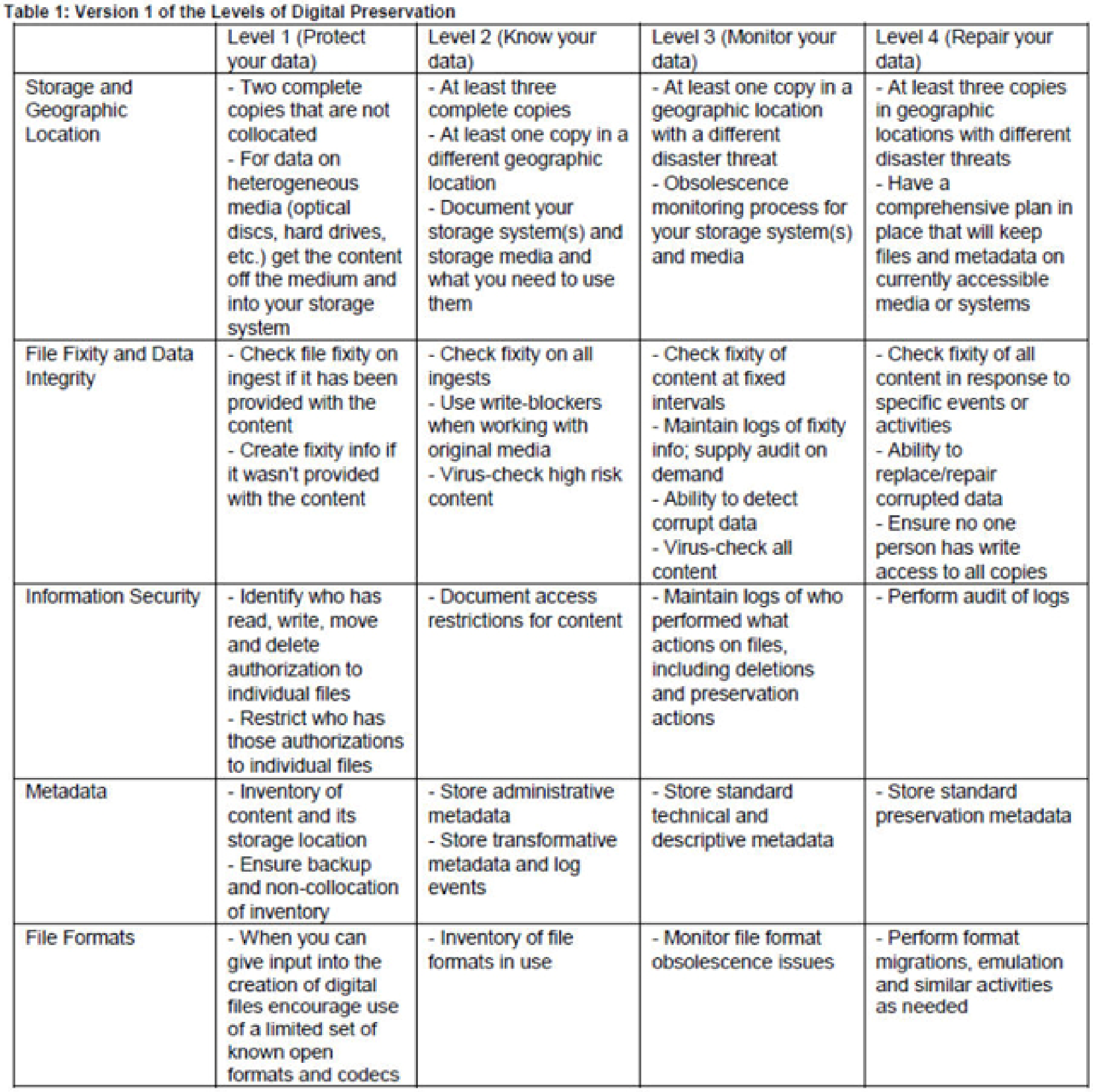
While the process of becoming a TDR is quite rigorous, the certification checklist is a useful tool to help an institution move in the right direction - even if never actually becoming certified.

Levels of Preservation

Digital preservation is distinctly different from preservation of analog objects. While preservation of paper-based materials generally requires establishing a stable environment and then minimizing interaction with the materials, digital preservation requires cyclical and relatively frequent interaction with objects being preserved - to perform functions like fixity checking, refreshing, and format migration when needed. This reality along with standards such as the TDR certification checklist can often create the impression that digital preservation is an unreachable goal for many institutions.

The Infrastructure Working Group of the NDSA created a set of tiered benchmarks for digital preservation activities (see below), and while these levels provide guidance for initiating a preservation strategy, the implicit goal is to continue to move up the tiers as far as is practicable. It is the judgment of the preservation group that Level 3 is the minimal level for compliance with the act.

Table 1: Version 1 of the Levels of Digital Preservation



	Level 1 (Protect your data)	Level 2 (Know your data)	Level 3 (Monitor your data)	Level 4 (Repair your data)
Storage and Geographic Location	- Two complete copies that are not collocated - For data on heterogeneous media (optical discs, hard drives, etc.) get the content off the medium and into your storage system	- At least three complete copies - At least one copy in a different geographic location - Document your storage system(s) and storage media and what you need to use them	- At least one copy in a geographic location with a different disaster threat - Obsolescence monitoring process for your storage system(s) and media	- At least three copies in geographic locations with different disaster threats - Have a comprehensive plan in place that will keep files and metadata on currently accessible media or systems
File Fixity and Data Integrity	- Check file fixity on ingest if it has been provided with the content - Create fixity info if it wasn't provided with the content	- Check fixity on all ingests - Use write-blockers when working with original media - Virus-check high risk content	- Check fixity of content at fixed intervals - Maintain logs of fixity info; supply audit on demand - Ability to detect corrupt data - Virus-check all content	- Check fixity of all content in response to specific events or activities - Ability to replace/repair corrupted data - Ensure no one person has write access to all copies
Information Security	- Identify who has read, write, move and delete authorization to individual files - Restrict who has those authorizations to individual files	- Document access restrictions for content	- Maintain logs of who performed what actions on files, including deletions and preservation actions	- Perform audit of logs
Metadata	- Inventory of content and its storage location - Ensure backup and non-collocation of inventory	- Store administrative metadata - Store transformative metadata and log events	- Store standard technical and descriptive metadata	- Store standard preservation metadata
File Formats	- When you can give input into the creation of digital files encourage use of a limited set of known open formats and codecs	- Inventory of file formats in use	- Monitor file format obsolescence issues	- Perform format migrations, emulation and similar activities as needed

Matrix Categories

- Storage and Geographic Location

This category addresses the issue of how many copies of a file should be kept, with the expectation that if one should be threatened, another could be copied to replace the loss. Included in this consideration is the physical location of the digital file, because different geographical locations will have different threats. The further that copies are from each other, the less risk that a single event will destroy all copies. The LOCKSS Program (Lots Of Copies Keep Stuff Safe) based at Stanford University Libraries, is an example of a preservation strategy that is based on distributed geographic locations.

- Fixity and Data Integrity

The PREMIS Data Dictionary defines fixity this way - “"information used to verify whether an object has been altered in an undocumented or unauthorized way." This is primarily done with the use of checksums, which act as a “fingerprint” of a document, and when they are calculated periodically and then compared with earlier checksums, it is apparent when something in the document has changed, either through human intervention, or because of spontaneous bit changes (“bit rot”). Much more information about fixity can be found at http://www.digitalpreservation.gov/documents/NDSA-Fixity-

Guidance-Report-final100214.pdf. This category also relates to processes used to extract data from obsolete media, as well as ensuring a virus-free environment.

- Information Security

  This category is for policies and practices relating to who has access to files and what their file-level permissions are, as well as documentation of file access - such as with a transaction log that records action, user, time, etc. The Digital Repository Audit Method Based on Risk Assessment (DRAMBORA), developed by the Digital Curation Centre, is a methodology to support this type of assessment, as well as many other types.

- Metadata

  The type of metadata that this category is concerned with is not only descriptive metadata, which most are familiar with, but also technical metadata such as capture equipment, file measurements, such as file size, resolution, pixel dimensions, for visual files, or time for an audiovisual file. A useful paper on different types of metadata for digital preservation purposes can be found at https://www.loc.gov/standards/premis/FE_Dappert_Enders_MetadataStds_isqv22no2.pdf

- File Formats

  Some file formats are appropriate for long-term preservation because they have qualities such as being uncompressed or lossless in their compression, they have open and available specifications, they are widely adopted, etc. To view what file formats the Library of Congress accepts as preservation-worthy, and why, see https://www.loc.gov/preservation/digital/formats/intro/intro.shtml

Document Strategies

The two most common formats for preservation-worthy electronic legal materials are PDF and XML

PDF Strategies

The Portable Document Format (PDF) is a very common specification used for presenting documents online, and is increasingly used as an archival master format (particularly PDF/A). PDF was developed by Adobe Systems in the early 1990's and the specification was made available at no cost in 1993, even though it was still a proprietary format. In 2008, the specification was officially released as a fully open and non-proprietary standard, leading the way for its use as an archival master format (although there are allied technologies such as PDF Forms that are still proprietary). Several states that have approved UELMA are using PDFs as their primary format for official electronic publications.

PDF/A is a multi-part ISO standard for long-term archiving format for electronic documents, based on the PDF specification. Because a PDF/A document contains everything it needs to present the document, including embedded fonts, it is intended to be stable over time, and can be used on any platform. In addition to the original release (Part 1), two additional parts have been made available, one in 2011 and another in 2012. In *PDF/A in a Nutshell 2.0*, the PDF Association states:

> "Put in the simplest possible terms, PDF/A is a PDF which forbids certain functions which could hinder long-term archiving. PDF/A also demands that the file meet certain requirements which guarantee reliable reproduction.
>
> For example, files must not be encrypted with a password, as all content must always be fully available. Embedded video and audio data are also prohibited: PDF/A consciously avoids anything that requires external software for display or playback. JavaScript and certain actions are also forbidden, as executing them could potentially alter the PDF.
>
> PDF/A also places higher demands on the information it contains. All required fonts (or at least all glyphs for the specific characters used) must be embedded within the PDF. To ensure a uniform colour appearance on a variety of platforms and devices, colour information must be given in a platform-independent format using ICC colour profiles. The software must also use the XMP format for metadata (which is used to store the data identifying the file as a PDF/A, for example).
>
> PDF/A also sets technical limits: for example, the page size is limited to an edge length of either 5.08 metres (PDF/A-1) or up to 381 kilometres (PDF/A-2 and PDF/A-3)."

There are many tools available to create PDF and PDF/A documents, perhaps the most obvious being those by Adobe, including Adobe Acrobat.

XML Strategies

Extensible Markup Language (XML) is a language for marking text similar in some ways to Hypertext Markup Language (HTML), the language of web pages. The difference is that HTML primarily defines how text will appear on a web page, whereas XML is designed to help define what data actually is. Another difference is that HTML fields are predefined and everyone uses the same tags, whereas XML is defined uniquely by the community that is utilizing it for a distinct purpose. This makes XML useful for sharing data.

An example of an instance of XML within the legal realm is the Global Justice XML Data Model (GJXDM). It is defined by the Department of Justice as "an XML standard designed specifically for criminal justice information exchanges, providing law enforcement, public safety agencies, prosecutors, public defenders, and the judicial branch with a tool to effectively share data and information in a timely manner."

Another example is LegalDocumentML, which “provides a common legal document standard for the specification of parliamentary, legislative and judicial documents, for their interchange between institutions anywhere in the world and for the creation of a common data and metadata model that allows experience, expertise, and tools to be shared and extended by all participating peers, courts, Parliaments, Assemblies, Congresses and administrative branches of governments.” The standard aims to provide a format for long-term storage of and access to parliamentary, legislative and judicial documents that allows search, interpretation and visualization of documents.” LegalDocumentML is part of the Akoma Ntoso specification - *Example of XML markup of a legislative document*.

Metadata for Digital Preservation

Best practice for digital preservation includes the creation of different types of metadata that will help to ensure long-term stewardship of digital materials. Metadata for preservation is information that supports and documents the process of digital preservation, and inherent in the OAIS model (see pg. 4) is the concept that all of the information packages (SIP, AIP, and DIP) consist of the digital content that is "packaged" along with descriptive information.

At a more granular level, the content (or "Content Information" in OAIS lingo) may include the digital object along with "Representation Information" (as also defined in the OAIS model). Representation Information is data that is necessary to interpret or use the digital object. If the digital object were a dataset, for instance, the Representation Information could possibly be information about how the data was generated, and what the structure of the dataset is, for example. Both the digital object and the Representation Information must be equally preserved as Content Information.

The "Preservation Description Information" describes what is required to preserve the Content Information, and might include elements of administrative, structural, technical, or rights metadata. The de facto international standard for preservation metadata is the PREservation Metadata: Implementation Strategies (PREMIS) standard.

PREMIS

From the PREMIS maintenance page, *"The PREMIS Data Dictionary for Preservation Metadata is the international standard for metadata to support the preservation of digital objects and ensure their long-term usability. Developed by an international team of experts, PREMIS is implemented in digital preservation projects around the world, and support for PREMIS is incorporated into a number of commercial and open-source digital preservation tools and systems. The PREMIS Editorial Committee coordinates revisions and implementation of the standard, which consists of the Data Dictionary, an XML schema, and supporting documentation."*

The full PREMIS specification, or Data Dictionary, is quite long and involved, but there is a much more accessible document called Understanding PREMIS, and here is part of the introduction:

> "…metadata is categorized according to what it is intended to accomplish: descriptive metadata helps in discovery and identification of resources, administrative metadata helps in managing and tracking them, and structural metadata indicates how complex digital objects are put together so that they can be properly rendered. Similarly, preservation metadata supports activities intended to ensure the long -term usability of a digital resource…
>
> Here are some examples of preservation activities and how metadata can support them:
>
> - A resource must be stored securely so that nobody can modify it inadvertently (or maliciously). Checksum information stored as metadata can be used to tell if a stored file has changed between two points in time.
> - Files must be stored on media that can be read by current computers. If the media are damaged or obsolete (like the 8" floppy disks used in the 1970s) it can be difficult or impossible to recover the data. Metadata can support media management by recording the type and age of storage media and the dates that files were last refreshed.
> - Over long periods of time even popular file formats can become obsolete, meaning no current applications can render them. Preservation managers must employ preservation strategies to ensure the resources remain usable. This might mean migrating old formats to newer equivalents, or emulating the old rendering environment on newer hardware and software. Both migration and emulation strategies require metadata about the original file formats and the hardware and software environments supporting them.

- Preservation strategies may entail changing original resources (migration) or changing how they are rendered (emulation). This can put the authenticity of the resource in doubt. Metadata can help support authenticity by documenting the digital provenance of the resource -- its chain of custody and authorized change history."

It further defines categories that are excluded from the Data Dictionary:

> "The Data Dictionary is not intended to define all possible preservation metadata elements, only those that most repositories will need to know most of the time. Several categories of metadata are excluded as out of scope, including:

- Format-specific metadata, i.e., metadata that pertains to only one file format or class of formats such as audio, video or vector graphics.
- Implementation-specific metadata and business rules, i.e., metadata that describes the policies or practices of an individual repository, such as how it provides access to materials.
- Descriptive metadata. Although resource description is obviously relevant to preservation, many independent standards can be used for this purpose, including MARC, MODS, and Dublin Core.
- Detailed information about media or hardware. Again, although clearly relevant to preservation, this metadata is left to other communities to define.
- Detailed information about agents (people, organizations or software) other than what is needed for identification.
- Extensive information about rights and permissions; the focus is on those that affect preservation functions."

METS

One of the strategies used in PREMIS is the concept of extension schemas. These are other related XML schemas that define elements that are useful in PREMIS but do not need to be redefined as part of PREMIS. One of these extension schemas is the Metadata Encoding and Transmission Standard (METS). This is the metadata standard often used for packaging the SIP, AIP or DIP for importing or exporting.

The most commonly used sections of a METS record are:

- Header – contains information about the METS document itself such as creator.
- Descriptive Metadata (dmdSec) - uses extensions also to utilize descriptive metadata schemes such as MARC, MODS and Dublin Core. Can be embedded in the METS record or point to external records.
- Administrative Metadata (amdSec) – information about how files were created, rights data, Masterfile/derivative information, and migration data can be recording in this section. This also can be recorded within the METS record or have pointers to external records.
- Files (fileSec) – this is a listing of all of the files that comprise the digital object
- Structural Map (structMap) – a mandatory METS section that outlines the hierarchical structure of the digital object. It also links the various elements within the structMap to their corresponding elements in the fileSec or dmdSec. This is critical for digital objects that are made up of many files but represent a single whole, such as a publication that might have hundreds of files that need to be arranged in a hierarchy (sections, chapters, etc.) with various descriptive metadata records that need to connect to specific places within that hierarchy.

Digital Storage

Cloud Storage

Storage of digital materials has changed dramatically since the late 1990's when the primary form of long-term storage was on local media such as CDs and DVDs and magnetic tape. Today it is common practice to use "spinning disk" for storage, either on local drives, institutional networked drives, or increasingly on cloud services. According to the Digital Preservation Handbook by the Digital Preservation Coalition (DPC), "cloud computing" is *"a term that encompasses a wide range of use cases and implementation models. In essence, a computing 'cloud' is a large shared pool of computing resources including data storage. When someone needs additional computing power, they are simply able to check this out of the pool without much (often any) manual effort on the part of the IT team, which reduces costs and significantly shortens the time needed to start using new computing resources. Most of these 'clouds' are run on the public Internet by well-known companies like Amazon and Google."*

Here is some basic information about these cloud solutions:

Amazon

- **Standard Simple Storage (S3)** - http://www.aws.amazon.com/s3

  Amazon provides two methods of cost for their cloud storage: on demand or reserve pricing. If the size of annual storage needed is known, prepayment is an option, which can save up to 75% on the cost.

  On Demand Cost:

Up to 50 TB Storage	51-100TB Storage	500TB+ Storage
0.023 GB/month	0.022 GB/month	0.021 GB/month

- **Amazon Glacier** - www.aws.amazon.com/glacier
  - Standard Infrequent Access (I/A)
  - From the website - "Customers can store data for as little as $0.004 per gigabyte per month, a significant savings compared to on-premises solutions. To keep costs low yet suitable for varying retrieval needs, Amazon Glacier provides three options for access to archives, from a few minutes to several hours."
  - Developer Resources - Amazon provides an API that allow developers to write interfaces to cloud storage systems or use third party solutions that provide user interfaces.

Google Cloud

- https://cloud.google.com/storage/archival/
- April 2018 cost - Capacity pricing is 1 cent per GB / month for data at rest for Nearline and 0.7 cents per GB / month for data at rest for Coldline.

Local digital storage

The size of local hard drives continues to grow, with at 60 terabyte (TB) solid state drive (SSD) announced in 2016. Tape storage also continues to be improved with IBM and Sony working on technology that could potentially store 330 TBs in a single cartridge that take less space than a hard drive.

- Local storage has greater management overhead
  - Must be backed up
  - Most hard drives have an average lifespan of about 5 years

- Easy to overwrite so protection must be put in place, such as Write once read many (WORM)
    - From Wikipedia: *"Write once read many (WORM) describes a data storage device in which information, once written, cannot be modified. This write protection affords the assurance that the data cannot be tampered with once it is written to the device.*

      On ordinary (non-WORM) data storage devices, the number of times data can be modified is limited only by the lifespan of the device, as modification involves physical changes that may cause wear to the device. The "read many" aspect is unremarkable, as modern storage devices permit unlimited reading of data once written."
    - View the Minnesota Case Study to see how WORM storage is being used for UELMA preservation.

Case Studies

California

Description

In 2012, through Senate Bill 1075, California enacted the Uniform Electronic Legal Material Act. In adding Article 4 (commencing with Section 10290) to Chapter 1 of Part 2 to Division 2 of Title 2 of the Government Code, the Legislature identified the Legislative Counsel Bureau as the official publisher for electronic legal material. "Electronic" and "legal material" is specifically defined in Section 10291 of the Government Code. Legal material is defined as the California Constitution, the statutes of the State of California, and the California Codes (hereafter referred to as "SB 1075 Legal Material").

Under the act, an official publisher that publishes legal material in an electronic record and also publishes in a record other than an electronic record may designate the electronic record as official if the publisher authenticates the electronic record, preserves the record, and ensures that the record is reasonably available for use by the public on a permanent basis (Secs. 10293, 10294, 10296, and 10297, Gov. C.). The Legislative Counsel Bureau publishes legal material both in an electronic record and in a record other than an electronic record and has designated the electronic record as official. These records generally originate in the Legislative Counsel Bureau as legislative measures that are eventually enacted into law. The Legislative Counsel Bureau also incorporates changes in law made through the initiative process into its database. The electronic legal material is then published at www.leginfo.legislature.ca.gov in both PDF and HTML.

In November 2016, California voters, through the initiative process, approved Proposition 54. As part of that initiative, the Legislature is required to cause audiovisual recordings to be made of all public legislative proceedings and to make those recordings available to the public through the Internet within 24 hours after the proceedings have recessed or adjourned for the day. The Legislature is also required to maintain an archive of the audiovisual recordings, which are to be accessible to the public through the Internet and downloadable for a period of no less than 20 years (para. (2), subd. (c), Sec. 7, Art. IV, Cal. Const.). In addition, Proposition 54 requires the Legislative Counsel Bureau to make the audiovisual recordings available to the public, with each recording to remain accessible to the public through the Internet and downloadable for a minimum period of 20 years following the date on which the recording was made, and to also then be archived in a secure format (para. (6), subd. (a), Sec. 10248, Gov. C.). The Legislative Counsel Bureau will make the audiovisual recordings available at www.leginfo.legislature.ca.gov and will preserve these recordings.

Implementation considerations

SB 1075 Legal Material

In developing preservation practices the Legislative Counsel Bureau must meet the requirement that the record is reasonably available for use by the public on a permanent basis as specified in Senate Bill 1075, California's enactment of the Uniform Electronic Legal Material Act. In that regard, the Legislative Counsel Bureau had three main considerations in implementing a solution for preservation of SB 1075 Legal Material:

1. Would the solution meet the standards for long-term preservation;
2. Would the solution be cost-effective; and
3. Would non-technical staff be able to use the solution.

The Legislative Counsel Bureau also wanted to meet Level 3 of the National Digital Stewardship Alliance levels of digital preservation. To reach that goal, the Legislative Counsel Bureau would have to:

- Store at least one copy in a geographic location with a different disaster threat;
- Engage in a monitoring process for our storage systems and media to determine obsolescence;
- Check fixity of content at determined intervals;
- Maintain logs of fixity information;
- Have the ability to detect corrupt data;

- Virus-check all content;
- Maintain logs of who performed which actions on files;
- Store standard technical and descriptive metadata; and
- Monitor file-format obsolescence issues.

To meet the aforementioned requirements, the Legislative Counsel Bureau considered three options:

1. Cloud storage using both preservation-specific cloud solutions and general cloud solutions;
2. Standard internal storage systems with standard backups already in use; and
3. Offsite optical WORM (write once read many) technology.

After considering the advantages and disadvantages of each of the three options, the Legislative Counsel Bureau decided to use standard internal storage. In addition, the Legislative Counsel Bureau has undertaken a pilot project to preserve the legal material in a preservation-specific cloud storage solution.

Audiovisual

The Legislative Counsel Bureau is working with the California Senate and Assembly to meet the requirements of Proposition 54. Under this proposition, audiovisual recordings must be made readily available to the public, in the downloadable format, for a period of no less than 20 years (audiovisual archive), and stored in a secure format. In addition, the Legislative Counsel Bureau is evaluating how to preserve the audiovisual archive. The goals of preservation encompass the following:

1. The use of methods and technologies to ensure digital content is usable by the public.
2. The use of methods and technologies that maintain the digital content as digital-content standards change.
3. The provision of a perpetually accurate rendering of the audiovisual recordings. In that regard, the audiovisual recordings would be retained in the original file format created by the audiovisual infrastructure. The Senate currently uses a file format known as "LXF" or "LEGO Digital Designer Model Files," while the Assembly uses a file format known as "TS" or "Transport Stream." These files are the original source files from which any future file conversion would be derived to meet new digital file format standards.
4. The storage of one copy of the audiovisual recordings file in a geographic location with a different disaster threat.
5. The provision of secure access to the audiovisual recordings file to ensure that the recordings are not modified.

Solution

Business Process Adjustments

SB 1075 Legal Material

The Legislative Counsel Bureau developed a strategic plan for the authentication and preservation of SB 1075 Legal Material, which included:

1. Identifying legal materials to be preserved, consistent with the Uniform Law Commission's version of the Uniform Electronic Legal Material Act. (California's enactment in SB 1075 covered fewer materials.);
2. Identifying units within the Legislative Counsel Bureau that are responsible for maintaining the legal materials;
3. Formalizing procedures for authentication of the legal material;
4. Establishing guidelines for cost-effective review of preservation needs of different legal materials on a cyclical basis to maintain data fidelity and integrity; and
5. Formalizing update procedures for preservation purposes.

Given the requirements and definitions set forth in Senate Bill 1075, the Legislative Counsel Bureau focused on the legal material developed during the lawmaking process related to the work of the Legislative Counsel Bureau that is made publicly available: the codes, statutes, and constitution. (Additional material generated during the legislative process has been identified as legal material that should be preserved. But that material is not part of the current authentication and preservation strategy.) Steps 1-3 above have been completed for the SB 1075 Legal Material. The systems that are used to draft and publish the SB 1075 Legal Material were adjusted so that the Legislative Counsel Bureau did not need to change its business process. Instead, software handles the authentication process and provides for storage of SB 1075 Legal Material. Also, software was developed so that staff could write to the preservation system at scheduled intervals. To complete steps 4 and 5, the Legislative Counsel Bureau must develop audit procedures for the SB 1075 Legal Material and formalize procedures to update that material and the technology that is used to allow access to the preserved material.

Audiovisual

The Legislative Counsel Bureau is currently evaluating business process changes in order to preserve audiovisual recordings under Proposition 54. One consideration is the establishment of a process by which current digital file standards are assessed, perhaps on a biannual basis following the legislative cycle, since the California State Legislature has many other processes that revolve around this cycle. If digital file standards change, the Legislative Counsel Bureau would begin the process of converting the original source LXF or TS files to the new standard, replacing the out-of-date standard.

Another aspect of digital preservation is the accurate rendering of authenticated content. Since these audiovisual recordings are intended to show the public how the legislative process produced bills that may become law, the Legislative Counsel Bureau must ensure the recordings are presented to the public without modification.

For this purpose with respect to SB 1075 Legal Material, the Legislative Counsel Bureau uses Adobe digital signatures to ensure that the documents have not been modified. There is no similar solution for audiovisual recordings currently available. One solution could be write once, read many technology. WORM data storage technology allows information to be written to a disc a single time and prevents the drive from erasing or editing the data. The implementation of this technology for securely stored, publicly accessible audiovisual recordings would in effect make them authentic.

IT Design/Components

SB 1075 Legal Material

The Legislative Counsel Bureau has undertaken a pilot project using Preservica's cloud-hosted, standards-based (OAIS ISO 14721) active preservation software for preservation of SB 1075 Legal Material. This web-based digital preservation application provides the Legislative Counsel Bureau with the ability to store files and perform preservation tasks. Preservica provides secure authenticated access, automatically classifies documents, and sets access permissions during ingest. Preservica has built-in workflows that are used by the Legislative Counsel Bureau for ingest of data and metadata management. Using Preservica's Submission Information Packet (hereafter referred to as "SIP") packaging desktop client, the Preservica administrator uploads content into organized file hierarchies based on statute year and California Codes updates, both of which take place twice a year.

The data is generated by the Legislative Counsel Bureau's Legal Division during each year of the two-year legislative session in the form of bills that are enrolled as part of the normal legislative business process and sent to the Governor for action. If a bill is either signed by the Governor or the Governor lets it become law unsigned, the systems create the statutes and authenticate the documents using Adobe certificate-based digital authentication. The Legislative Counsel Bureau Preservica administrator extracts the authenticated document and uses Preservica's SIP client to load into the cloud preservation site. At that time, Preservica's SIP client also adds basic descriptive metadata. That metadata was developed as a collaborative engagement between the Legislative Counsel Bureau and the California State Archives using Dublin-Core standards to meet the needs of the Legislature.

Audiovisual

The Legislative Counsel Bureau is implementing an EMC Isilon WORM storage technology for preservation of the audiovisual recordings. Isilon is a scale-out, network-attached storage platform offered by EMC Corporation for high-volume storage. This system will store both the original format and a modified format MP4 file for public access and downloading.

As a long-term data preservation strategy, the Legislative Counsel Bureau will store the audiovisual data at more than one site. Within the Legislative Counsel Bureau's primary location in Sacramento, an EMC Isilon storage system will be implemented consisting of performance-optimized storage nodes and capacity nodes. The strategy to protect the data features redundancy and will place another EMC Isilon storage system at an established off-site location. The video data would be replicated to the off-site location using high-bandwidth, secure and redundant point-to-point wide area network (WAN) connections. If an outage occurred at the primary location, the video data could be accessed and restored from the off-site location.

Licensing will be purchased to enable the write once, read many technology available within the Isilon storage array known as "SmartLock." The Isilon SmartLock technology protects the data against accidental or malicious deletions or alterations. This type of technology helps protect digital files from being modified while those files reside within the SmartLock-enabled file directory. With this technology, any video made available on the Internet would be protected from alteration, ensuring the video file's integrity. Thus, the technology would satisfy data-authenticity criteria within digital preservation, to meet the requirements of Proposition 54 that the audiovisual recordings be in a secure format.

Costs

SB 1075 Legal Material

The Legislative Counsel Bureau decided to use the current business and technical processes for preservation of SB 1075 Legal Material. Therefore, development costs were absorbed into the standard development budget.

The Legislative Counsel Bureau does not allow public access to the Preservica archive. Thus, the Cloud Edition Starter subscription – up to 250GB at a cost of $4,000 per year – meets the Bureau's current needs. The Preservica system will allow the Legislative Counsel Bureau to scale up, as storage needs increase.

Audiovisual

The Legislative Counsel Bureau estimates the average cost of the storage solution for audiovisual recordings, which is required to store the original file format (LXF), will be $250,000 per year. This includes the backup WORM solution recommended for preserving the audiovisual data.

Our Current Assessment

SB 1075 Legal Material

The Legislative Counsel Bureau followed a structured systems development life cycle in designing the archiving process, in order to meet the preservation requirements of the Uniform Electronic Legal Material Act, as enacted in California. Information technology staff from the Legislative Counsel Bureau met with subject-matter experts to understand the legal material that required archiving for preservation purposes. This included understanding what metadata the owners of the legal material considered meaningful. Questions were asked such as: What was the best timing to capture the legal material? How would consumers of the legal material likely identify the material in a search?

IT staff also wrote requirements, including use cases, for procedures to extract files, create metadata, and stage the files in preparation for extraction to a cloud-based archiving platform.

As previously stated, a determination was made to use an industry standard (Dublin Core) in determining the metadata to be captured for the legal material. The use of this standard allowed easy integration with Preservica's ingest workflow. Metadata is easily discoverable for archived files. This standard allowed the Legislative Counsel Bureau to share information with State Archives.

The Legislative Counsel Bureau struggled with how to handle versioning of the legal material. Not all legal material had the same requirements for versioning. The IT staff decided to capture a snapshot of the legal material on a certain date and time, after working with the owners of the legal material to determine that date and time. The Legislative Counsel Bureau ingests and identifies a snapshot of the legal material based on the date the material was extracted from the document repository.

The Legislative Counsel Bureau is also currently manually initiating the archiving process. We need to find ways to leverage the workflow available in Preservica to automate the extraction and ingest process, thereby taking advantage of the dedicated document repository that utilizes data services to make the legal material available. This process would allow for a standard interface to all the current and future documents that need to be preserved. In turn, the process will make extraction of legal material easier.

Another advantage of the current archiving solution is that metadata is seamlessly integrated with the legal material as it is ingested into Preservica. This makes it easy to discover metadata when viewing the legal material in the preservation tool.

There are areas that need further attention, including how to securely share necessary documents and metadata with State Archives. Though the Legislative Counsel Bureau and State Archives use the same archiving tool, there are some redundancies in effort when both entities preserve the same documents.

The Legislative Counsel Bureau has made the archiving effort for preservation under the Uniform Electronic Legal Material Act an ongoing internal project. That will allow the project team to find ways to improve the processes and procedures, particularly if additional types of legal material are added to the project. The project team must communicate with the owners of the legal material and external customers to ensure that their requirements continue to be incorporated into the archiving solution.

Audiovisual

The Legislative Counsel Bureau is too early in the implementation of the requirements of Proposition 54 regarding audiovisual recordings to assess solutions.

Notes

This report is for educational purposes. References to any specific commercial products, process, service, manufacturer, company, or trademark do not constitute an endorsement or recommendation.

Minnesota

Daniel Kruse Systems Analyst/Programmer
Jason Duffing Systems Analyst/Programmer
Jason Judt Data Systems Project Manager

The Minnesota Office of the Revisor of Statutes has constructed KEEPS; a custom software solution to satisfy the requirements for preservation and security detailed in the Uniform Electronic Legal Material Act (UELMA). The Keep Electronic Edicts Preserved & Secure (KEEPS) system is in the testing phase and is scheduled for deployment in 2016 Q4.UELMA

Background

In Minnesota, UELMA was enacted in 2013 as Minnesota Statute chapter 3E. UELMA establishes an outcomes-based, technology-neutral framework for providing online digital legal material with the same level of trustworthiness traditionally provided by paper publications. The Act requires that official electronic legal material be: (1) authenticatable; (2) preserved, either in electronic or print form; and (3) accessible. The KEEPS solution was specifically designed to satisfy requirement (2).

The UELMA requirements for preservation and security are in section 3E.07. Section 7 states that if official legal material is preserved in an electronic record, the official publisher shall:

(1) ensure the integrity of the record;
(2) provide for backup and disaster recovery of the record; and
(3) ensure the continuing usability of the material.

System Description

The KEEPS system's primary goal is to ensure the integrity of official electronic records. The system makes backup and disaster recovery possible in several ways. The use of write once read many (WORM) disk drives and offsite tape backups created from separate document repositories ensures the continuing availability and usability of the material.

The software system was built in-house using staff programmers and existing commercial products (Figure 1). These products are: a virtual machine, write once read many (WORM) disk, a relational database, and a tape backup application. Additionally, a custom software application was deployed to the virtual machine.

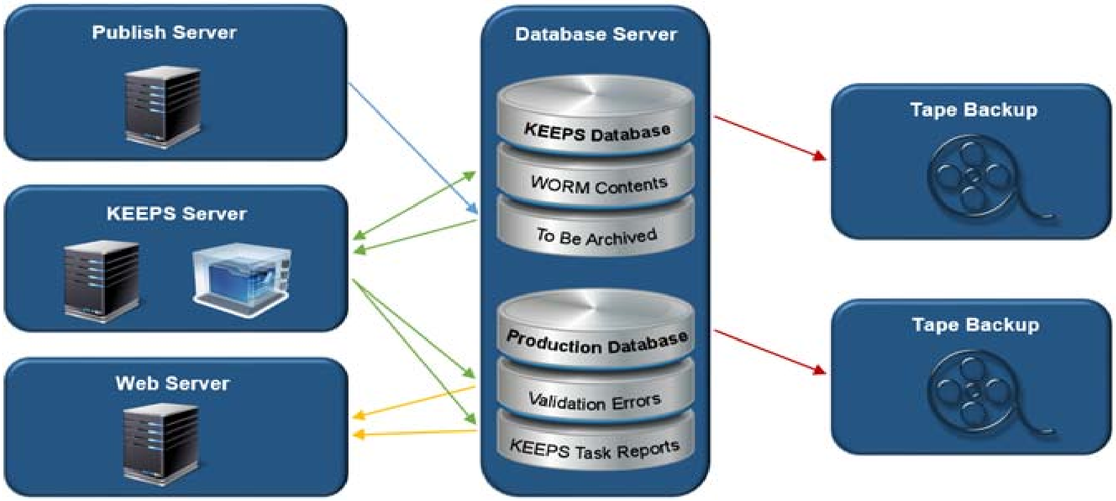


Figure 1 – System Diagram

KEEPS integrates with the legislative publishing system, a backend database, and a private intranet. The KEEPS server has exclusive access to the WORM Disk, and read write access to the database. The web has read only access to the database for the purpose of monitoring the system.

Write once read many (WORM) disk drives are an integral part of this system. WORM disks are essential to ensuring the integrity of the data and are the foundation around which this system was designed. The KEEPS solution will leverage GreenTec WORM Storage Servers as the hardware best suited for this system. These storage devices enforce write-once capabilities through hardware mechanisms rather than software running on a computer.

Development

The basic requirements for the KEEPS system can be summarized as:

- Preserve newly published documents.
- Catch any errors in our publicly available legal material.
- Run independently of our other software without user intervention.

KEEPS Software was written using Java SE8 consisting of three primary modules (Figure 2): an Archiver that writes published UELMA-compliant documents to the WORM Disk, a WORM Contents Populator that records WORM disk contents in a database table which is used in the validation process, and a Validator that works with the database to validate the publicly available legal material and populates the Validation Errors table. Each of the modules records its activity in a database table that can then be seen on an intranet page. The software is built and deployed using the Apache Ant and Ivy projects. The modules can be run at predetermined intervals or on demand and are synchronized by the Schedule Manager. The WORM disk documents and public UELMA documents are backed-up separately to tape. The tapes are stored offsite.

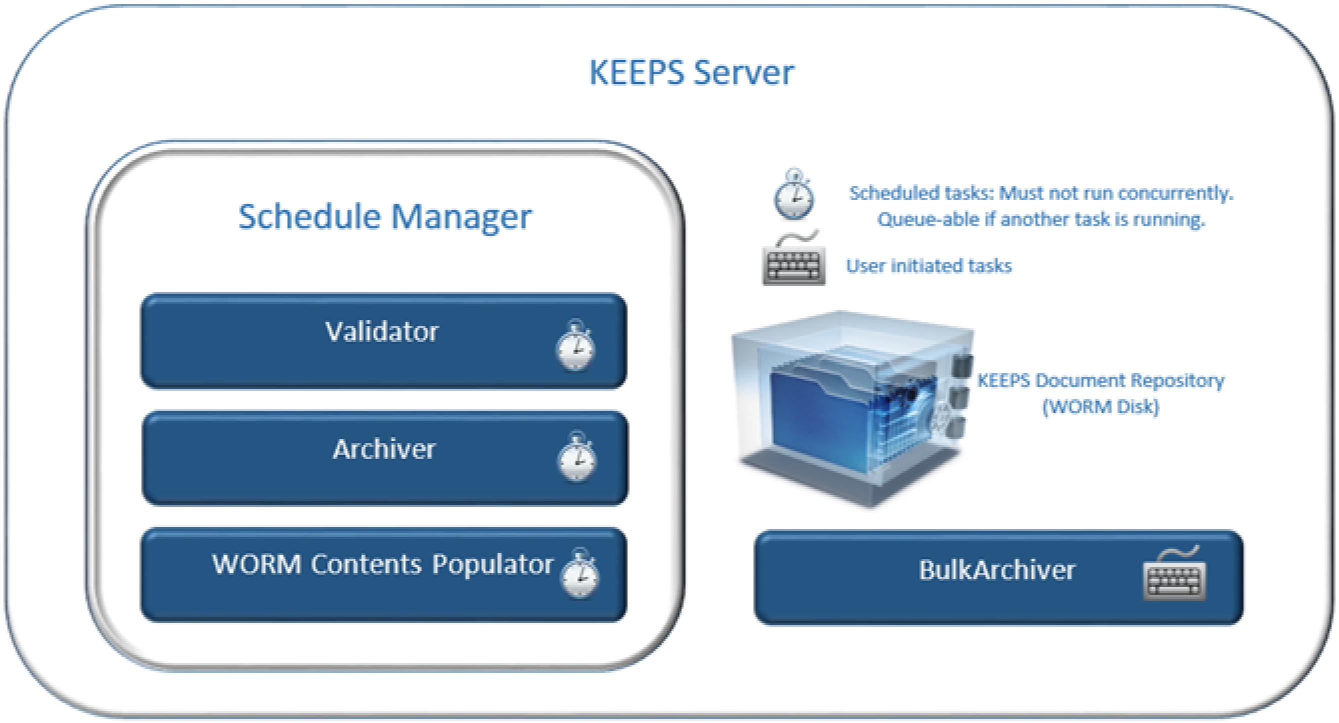

KEEPS Server
Schedule Manager
Validator
Archiver
WORM Contents Populator
Scheduled tasks: Must not run concurrently. Queue-able if another task is running.
User initiated tasks
KEEPS Document Repository (WORM Disk)
BulkArchiver

Figure 2 – KEEPS Architecture

Testing

Tests of the identified scenarios that constituted corruption or failures of the public UELMA documents system were performed. Tests covered: (1) an unauthorized document inserted into the production database, (2) a document removed from the production database, (3) changes of any type to an existing document in the production database, and (4) the inability to archive a document to the WORM disk. In all cases, our validation system correctly identified the issue and reported it. Deployment is effortless and has been repeated many times to ensure the robustness of the software and the ease of installation.

Load testing was conducted on a virtual server with 4 GB of memory running Windows Server 2012 R2. For archive testing, 51,463 Minnesota Statute sections, in the form of PDF, totaling 6.3 GBs were published. Total archival time was 37 minutes. Validation testing occurs daily on 606,105 PDF documents totaling 65 GB. Daily validation completes in under 40 minutes. All load tests are considered successful in our environment.

The system is stable and has been running without programmer involvement since March 23, 2016. It handles errors gracefully and continues processing, providing detailed logs that can be used to troubleshoot issues.

Schedule

- Prototype, 2014-2015

  The first prototype was developed as a proof of concept, by Stephen Segal the Principal at System Specialties, Minneapolis, MN. Written in PHP it provided the basis for good estimates of the time required to validate our entire repository of legal material on a nightly basis.
- Build & Test, January– March 2016

  The system was functionally tested as it was developed, and released to long-term testing.
- Testing, April– September 2016

  The system is stable and running in a simulated production environment.
- Final Deployment, October 2016
  - Production environment will be completed.
  - GreenTec WORM Storage Servers will be purchased.
  - BulkArchiver will write all existing UELMA documents to the WORM Disk.
  - Archiver will write new UELMA documents to WORM Disk.
  - Validator will run daily.

Washington, D.C.

David Greisen - Open Law Library

Vincent Chuang - Open Law Library

The Open Law Platform is a software system created for the purpose of publishing laws, codes, legal interpretations, and any other legal document produced by a government. As part of taking a digital-first approach to legal publishing, the Open Law Platform incorporates UELMA compliance as a core component of the platform.

The Council of the District of Columbia is using the Open Law Platform to publish its laws and code (https://code.dccouncil.us) and provides a case study for replicating key features and processes at other jurisdictions. XML representations of the District's laws and codes can be found at https://github.com/dccouncil/dc-law-xml.

The Platform's version of UELMA compliance is modeled on brick-and-mortar libraries. Lessons about readability over time, information redundancy, version history, and authentication have been learned over centuries in the physical world. And it is useful to apply many of those ideas when considering digital preservation and authentication under UELMA.

Terminology

The Council of the District of Columbia is a *Publishing Entity*. As a *Publishing Entity*, the Council is responsible for publishing and authenticating the *Library* of official legal materials relevant to itself. The Council's *Library* contains various *Documents*, including rapidly changing documents, like the entire District of Columbia Code, and static documents like individual laws. Another *Publishing Entity* could be the Executive Office of the Mayor, and its *Library* could include *Documents* such as the DC Municipal Regulations and the DC Register.

An important difference between a *Library* in the Open Law Platform and a brick-and-mortar library arises in the context of time. The contents of a physical library might change over time, but you can only ever visit the library as it is today. That is to say, if Harvard Law Library throws away its copy of *A Wrinkle in Time*, the library is still the Harvard Law Library, but you can never travel back in time to read *A Wrinkle in Time* there. An Open Law Platform *Library* consists of a snapshot of every version of the library as it has ever existed. For instance, on January 1, 2018, the *Library DC-Law-XML* may have contained one thousand laws. We would refer to that *Library* as *DC-Law-XML* as of January 1, 2018. On January 2, 2018, the Council may pass a new law and add it to *DC-Law-XML*. Unlike a traditional library, you can visit the *Library* as it existed on January 1 or as it existed on January 2.

A *Consumer*, like a citizen of the District, can view *Documents* within a *Library* or download the entire library. And *Hosting Entities*, such as law libraries, can download and host a copy of an entire official *Library*. For instance, if the Harvard Law Library wished to host an authenticatable copy of the Council's *Library* on the Harvard Law Library website, it could do so, just as it could purchase and host an official paper copy of the District of Columbia Code.

Considerations

In addition to UELMA itself, the Open Law Platform was designed with several related and overlapping considerations.

Time

Legal documents have a long history, and that history is itself substantively valuable. As a result, the Open Law Platform is created with the intention that every version of the content it publishes be accessible and authenticatable long into the future. And because legal history is long, this means capturing and maintaining large volumes of documents. The Council's *Library* is only two years old, yet contains more than thirty thousand pages of laws and code, and is growing by over five thousand pages annually. *Libraries* must be manageable, usable and responsive even while containing orders of magnitude more data than traditional libraries.

Authentication

The authentication scheme must also be robust against a wide range of factors from the perspective of *Publishing Entities, Hosting Entities,* and *Consumers*.

Publishing Entities are governments, and governments vary widely in the number of personnel, institutional capacity, and organizational structure. The authentication process must be usable in these varying environments. It must be possible, for any government, to clearly, easily, and securely convey (1) *when* a document was published, (2) *who* published it, and (3) the *authority* of that person to do so. All three questions can be answered with an appropriately designed cryptographic signing framework.

In order to be robust over time, the framework must be resilient to the loss or compromise of private cryptographic keys. The system must also provide for restoration in the event of a government-scale catastrophe: there must be a mechanism for restoring a *Library* after all encryption keys have been lost. And the system must operate on government time scales. Because published documents are intended to be used over the course of decades, accessibility (by way of readability or cryptographic scheme) must keep pace with changing technology. A *Library* must be accessible and authenticatable long after the *Publishing Entity* has abandoned it and moved on to other technologies, just as an official paper copy is at a law library even if the government no longer has that particular version.

A *Hosting Provider* should be able to host authenticatable versions of a *Library* for its patrons. For the *Consumer*, a *Library* must function across every use case. In situations in which the delivery network is compromised (such as hackers taking over the *Publishing Entity*'s web server), a *Consumer* must still have confidence that the *Library* being viewed is authentic. As with physical text, a *Library* should be accessible and authenticatable even without an internet connection. Because *Libraries* have a version component, a *Consumer* should be able to ascertain information regarding both authenticity and versioning information, akin to checking publication information inside a book.

Redundancy

The system must also be distributed. Just as Harvard Law Library and USC Law Library may both carry a copy of *A Wrinkle in Time*, a *Consumer* should be able to access a *Library* from a *Hosting Entity* and be able to confirm that the *Library* is the same as one acquired from the original *Publishing Entity*. Even if a *Consumer* can never access the original *Library* from the original *Publishing Entity*, the hosted *Library* should be authenticatable without reference to the original.

The Open Law Platform Solution

With these various considerations in mind, the Open Law Platform utilizes a combination of technologies, including XML, Git, and strong encryption, to implement a set of authentication techniques.

XML

The Open Law Platform stores almost all documents as plaintext XML. By using plaintext instead of a binary format (e.g., PDF), a *Library* and its *Documents* are virtually guaranteed to be readable for decades to come without relying on legacy software. Plaintext also requires considerably less storage space than binary formats. For the Council, 30,000 pages of XML can be stored in 100 megabytes, while only 10,000 pages of PDFs require fifty times the space when compressed and 500 times the space when uncompressed. This difference means it is feasible to store every version of a plaintext document in less space than a single version in PDF.

XML also has the advantage of being able to store the structure of a document, instead of just presentation information (i.e., how something looks on a screen). This means documents can be converted into any display format in the future and not be tied to any specific software. Together, these benefits of XML make it possible to satisfy the need for usability over time, ability to store large amounts of historical information, and speed of use.

A common concern with XML-based solutions is that XML can appear complicated and requires a different set of tools than most lawyers are used to using. This has resulted in very few UELMA-compliant XML implementations.

The Open Law Platform solves this problem in several ways. First, the platform focuses on making the XML very clean and simple, using, whenever possible, a jurisdiction's terminology to describe a document and its contents (e.g., Title, Chapter, Subchapter, and Section). The platform also stores metadata logically within the document, again using the same terminology as the jurisdiction.

Good tooling (i.e., software for viewing and editing the XML) also goes a long way to making XML more digestible. The platform provides a mix of custom XML schemas and software to ensure XML accuracy, as well as automatic error detection, and other smart editing capabilities. By focusing on user experience, lawyers familiar with the District's laws and code were able to navigate and understand XML representations of law and code with no training.

Converting documents into XML is itself a process. But again, good tooling can make the process feel seamless. The Open Law Platform includes Open Law Draft, a Microsoft Word plugin that helps drafters conform to their jurisdiction's style guides. Once the document conforms to the style guide, Draft can turn the document into correct XML without user input.

An XML-based solution has many benefits inherent in its format, with the biggest barriers being usability and conversion of existing documents into the format. A focus on user experience and good tooling can overcome these high hurdles. Success on this front reveals the downstream benefits of XML that ultimately outweigh the initial costs.

Git

The Open Law Platform stores XML (and any static PDFs) using the open source Git distributed version control system (https://git-scm.com/). In simplest terms, Git is a piece of software that keeps track of changes to one or more files (each group of one or more files collectively referred to as a "repository"), records the differences between new and old versions of one or more files, and maintains a history of the differences. It does so, in part, by providing the ability to sign each version with a unique cryptographic key (https://git-scm.com/book/id/v2/Git-Tools-Signing-Your-Work). This makes it possible to preserve different versions of documents as they change and creates an immutable chain of authenticatable versions back to the original.

Git makes it easy to copy an entire *Library* from one place to another and then keep the copy up-to-date with the original by just syncing changes. Because every copy of a *Library* has all the historical information and authentication information of the original, it is inherently fraud resistant. In the event a malicious actor attempted to modify the history of the original *Library*, the next time a copy attempted to sync with the now-fraudulently-modified "original", the copy would detect the modification of the history and reject the fraudulent history.

Git is free, open source, and available on virtually every platform. There are also many cloud services that provide Git access. Because of this wide availability, a *Library* that is stored as a Git repository can have all of its historical information hosted on a variety of physical machines located across a large geographic region. And every copy is easily authenticated.

Signing a Library

With XML and Git as the underlying technologies, the Open Law Platform implements specific processes to achieve the needed authentication outcomes.

At the government level, each employee who has authority to publish legal documents receives a smart card (e.g., https://www.yubico.com/products/yubikey-hardware/). A small group of employees (minimum of three, preferably five) or other trusted individuals creates an *Attesting Group*. Each member of the Attesting Group (an *Attestor*) has a smart card that they use to sign *Attestations of Authority*.

Once a threshold of *Attestors* (usually 50%) have attested that a particular person has authority to publish official documents, that person is a *Publisher* (as part of a *Publishing Entity*) and can sign new releases of a *Library*. If a *Publisher* leaves the organization or loses their key, the *Attestors* attest that the old key no longer has authority to publish. If an *Attestor* leaves the organization or loses a key, a majority of the remaining *Attestors* can attest that the old key is no longer valid and can also attest that a new key is a valid attestation key.

Normally, cryptographic signatures are very complicated or very brittle. This system, however, ensures that the system continues to work even if several keys have been lost or compromised. Moreover, encryption keys are stored on physical devices and protected by a password. Even if a jurisdiction's network is compromised, their keys are not.

Authenticating a Library

Attestations of a *Publisher*'s authority are stored in the *Library*. Thus, when a *Publisher* signs a *Library*, all the information needed for authentication is available within the *Library*. This technique combined with the use of Git to create a cryptographically secured history and to create easily replicable repositories results in robust authentication system for *Libraries*.

While a *Consumer* or *Hosting Entity* can confirm that all signatures and all attestations are valid back to the very first release of a *Library*, they will always require at least one out-of-band authentication (i.e., authentication via something other than the original receiving channel) to confirm the very first release. The design of the Open Law Platform aims to decrease the friction required to obtain out-of-band authentication.

For starters, once a *Consumer* or *Hosting Entity* has performed one out-of-band authentication, usually via a telephone call to the *Publishing Entity*, the use of Git to store a *Library* means any future updates can be confirmed authentic without external verification. Just as law libraries currently provide indirect authentication of paper laws—they buy the laws from the official publisher then represent to their users that these are the official laws—law libraries can download a *Library* from the official *Publishing Entity*, perform the single out-of-band authentication, and then represent to their patrons that these are official laws.

Once a *Library* is hosted by more than one *Hosting Entity*, it becomes possible to perform out-of-band authentication by comparing the various hosted *Libraries*. And this comparison can then be automated for ease of use by *Consumers*.

Importantly, this system works without relying on a public root certificate (like those underlying HTTPS) or a web service maintained by the *Publishing Entity*. If the web service goes down, or the *Publishing Entity* stops supporting the web service, the *Library* will still be fully available and authenticatable through the constellation of *Hosting Entities*. In root certificate based systems, compromising the root certificate means compromising all historical documents signed by the certificate. While it may seem unlikely that a root cert will be compromised, this is surprisingly common. Symantec, until recently one of the most trusted root certificate authorities, was forced by Google and Mozilla to divest itself of its root certificate in 2017 because of major systemic security violations. An authentication system premised on a public root certificate system is too fragile to provide authentication over decades. Instead, by intimately tying the authentication mechanism to the preservation mechanism, preserving the documents automatically preserves the authentication.

The discussion up to this point has been regarding *Archival Authentication*, i.e., downloading and authenticating an entire *Library* (along with all historical versions). Most users, such as lawyers and judicial staff will be performing *Transient Authentication* of particular versions of individual *Documents*. For these purposes, a web-based authentication service is ideal, as it makes it trivial for users to authenticate. The Open Law Platform is designed to provide an authentication service through a website, an application programming interface (API), and plugins for all major browsers. The authentication service will compare a hash of the *Document* to be authenticated against the hashes of all versions of all authentic *Documents*. The authentication service can therefore not only tell the user if a *Document* is authentic, but also when the version in question was created and if/when it was superseded by a newer version. Unlike other web authentication services, the Open Law Platform optionally provides the full cryptographic audit chain so an individual can confirm for themselves against a full copy of the *Library* that the *Document* in question is authentic.

Redundancy

Redundancy is built into the system because of the way repositories are stored using Git and because of the authentication process.

With respect to redundancy of information, the wide adoption of Git and the various commercially available Git hosting solutions means that anyone at any time can easily retrieve and host their own copy of a *Library*. This replicability means that *Libraries* can be quickly distributed across large geographic areas and can help recover from data loss. Moreover, each copy of a *Library* is cryptographically signed in a way that permits for corruption detection.

No less important and considerably more complex is the redundancy of authentication. If many *Hosting Entities* are constantly pulling down updates of fully authenticated laws, the constellation of entities can help a *Publishing Entity* recover from catastrophic losses (such as a natural disaster). If all *Attester* keys are lost in, say, a flood, a group of *Hosting Entities* can represent that a new set of *Attester* keys are official keys, helping to rapidly bootstrap a *Publishing Entity* back to an authenticatable state. Moreover, the presence of verifiably authentic copies held by *Hosting Entities* means that any new copies can be authenticated against those copies even if the original *Publishing* entity no longer exists.

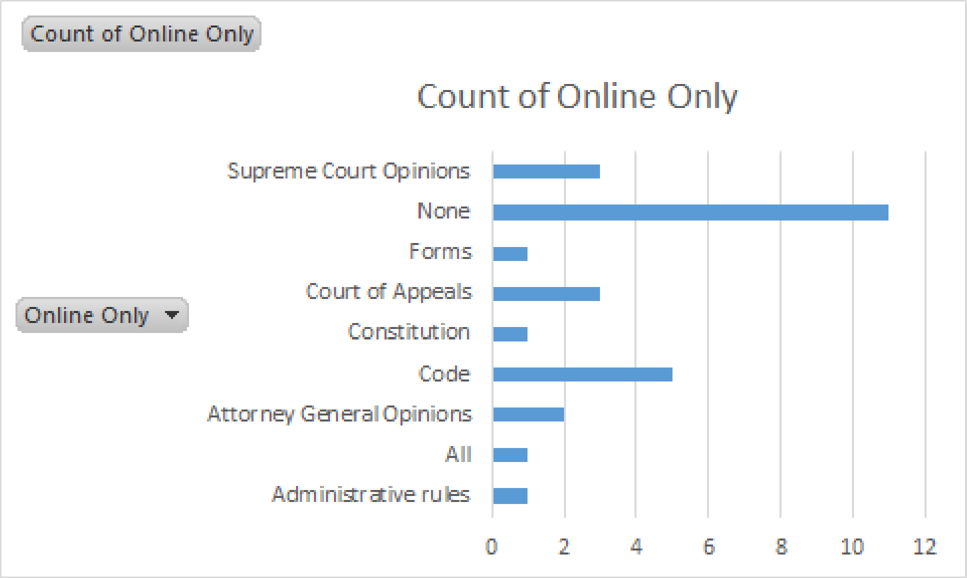
Overall Assessment

The initial cost of developing the Open Law Platform was significant, but it is now a fully generalized legal publishing platform that is available for any jurisdiction to use. Free Git repository hosting is available from several well-established commercial providers including GitHub, Bitbucket, and GitLab. As of February 2018, version 1.0 of the Open Law Platform is complete and running for the District of Columbia. Documents published using the Open Law Platform can be found at https://code.dccouncil.us, and XML representations are available at https://github.com/dccouncil/dc-law-xml. Initial work on *Archival Authentication* is complete and is being rolled out to the Council; *Transient Authentication* is expected June 2018.

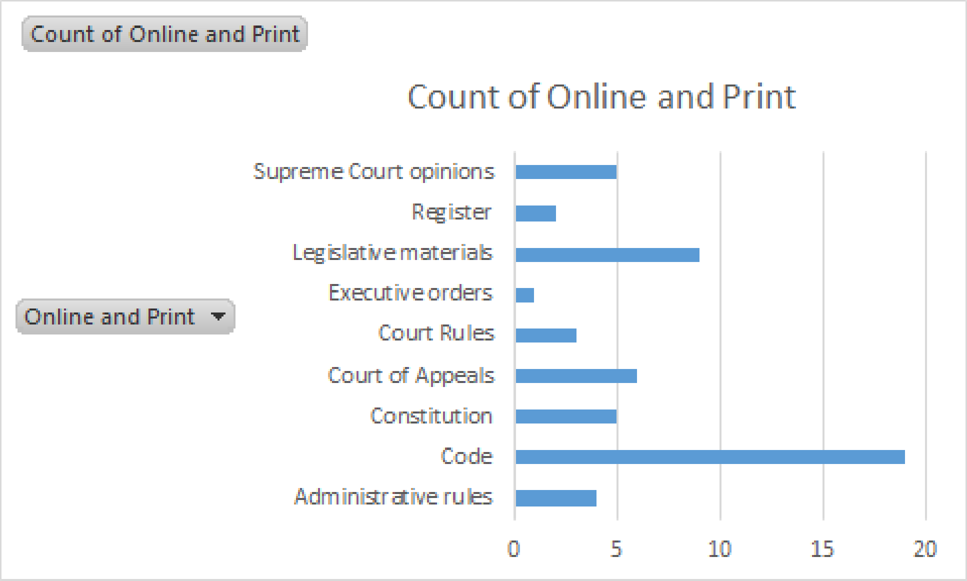
Appendix I: Survey Results

These results are compiled from responses from 20 states

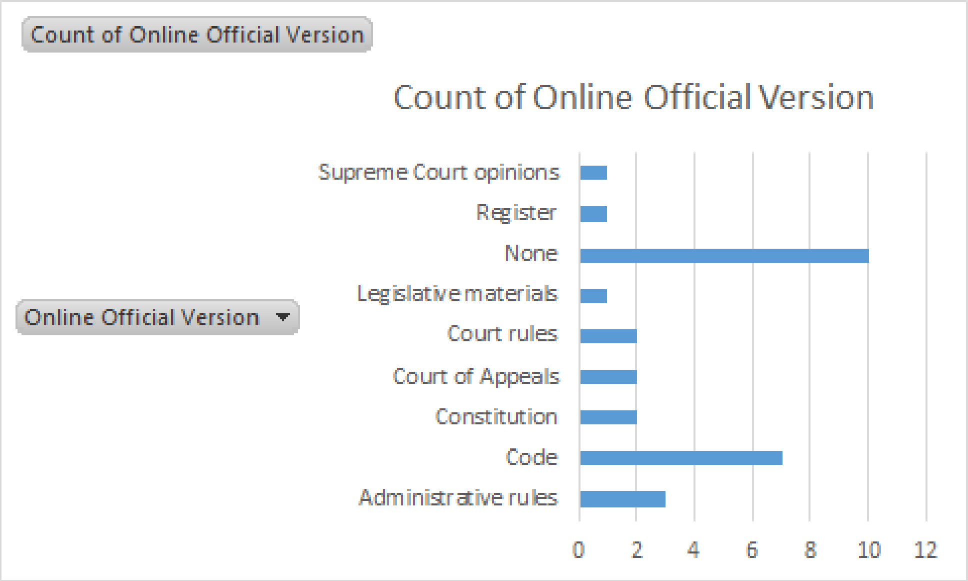
What legal materials does your state publish online only?



What legal materials does your state publish both in print and online?

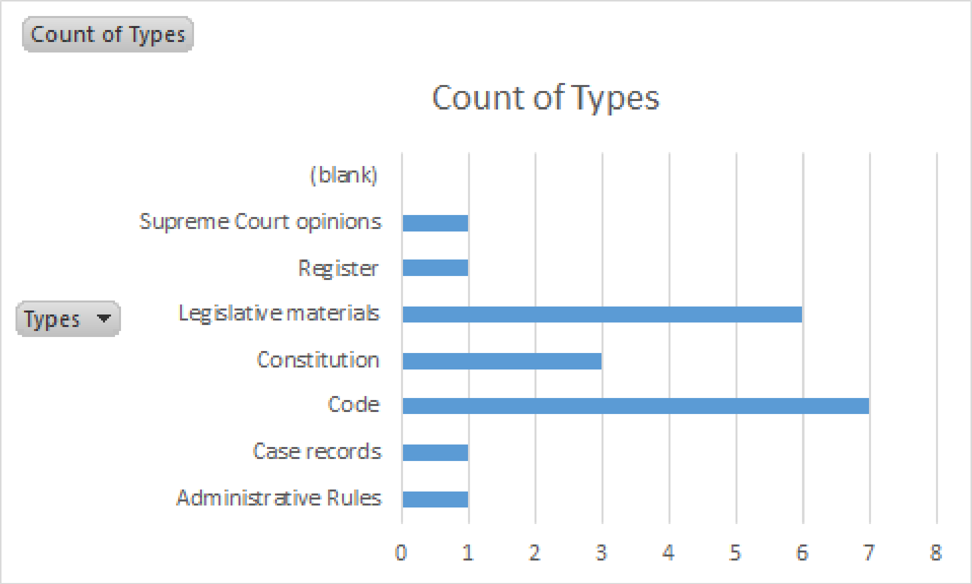


Of the materials identified in questions #1 or #2, which online materials are deemed to be the official version?



Have you digitized any paper-based legal materials?

Yes - 68%; **No** - 32%



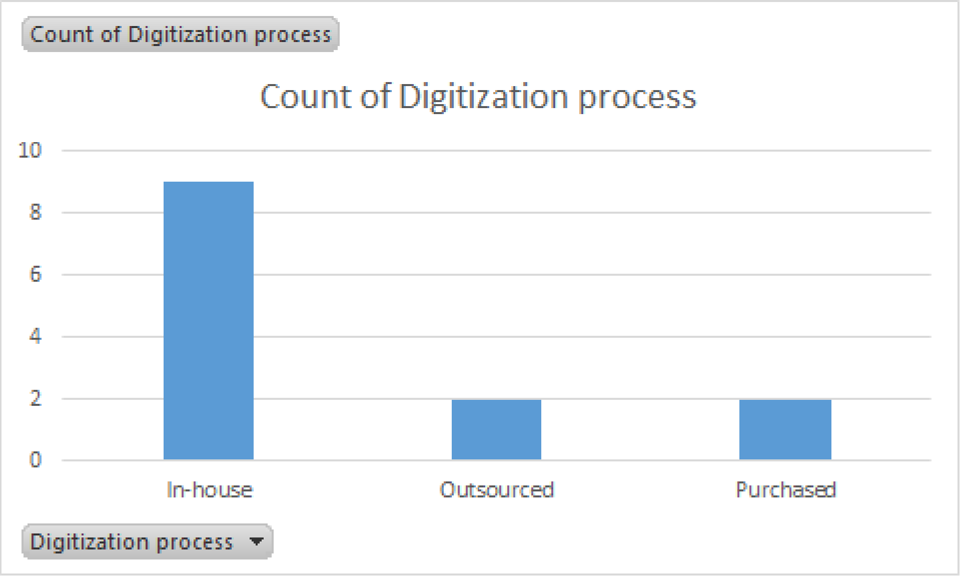
If so, do you intend the digitized materials to be considered official?

Yes - 19%; **No** - 62%; **N/A** - 19%

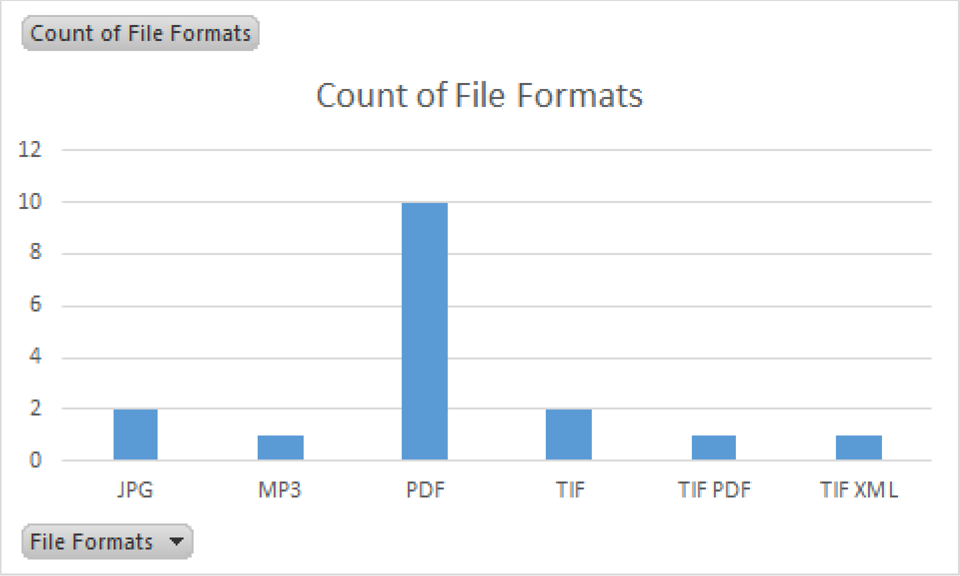
Do you plan to digitize paper-based legal materials within the next 18 months?

Yes - 29%; **No** - 67%; **Maybe** - 4%

What digitization processes are you using?



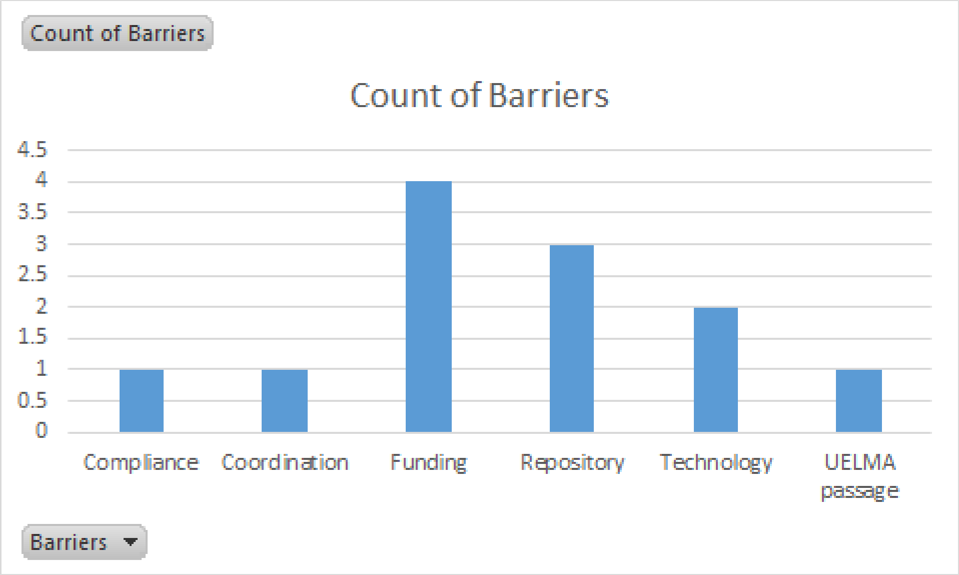
For legal materials that you are digitizing, what is the file format you are using (e.g. pdf, xml, doc, tif, jpg, jp2000)?



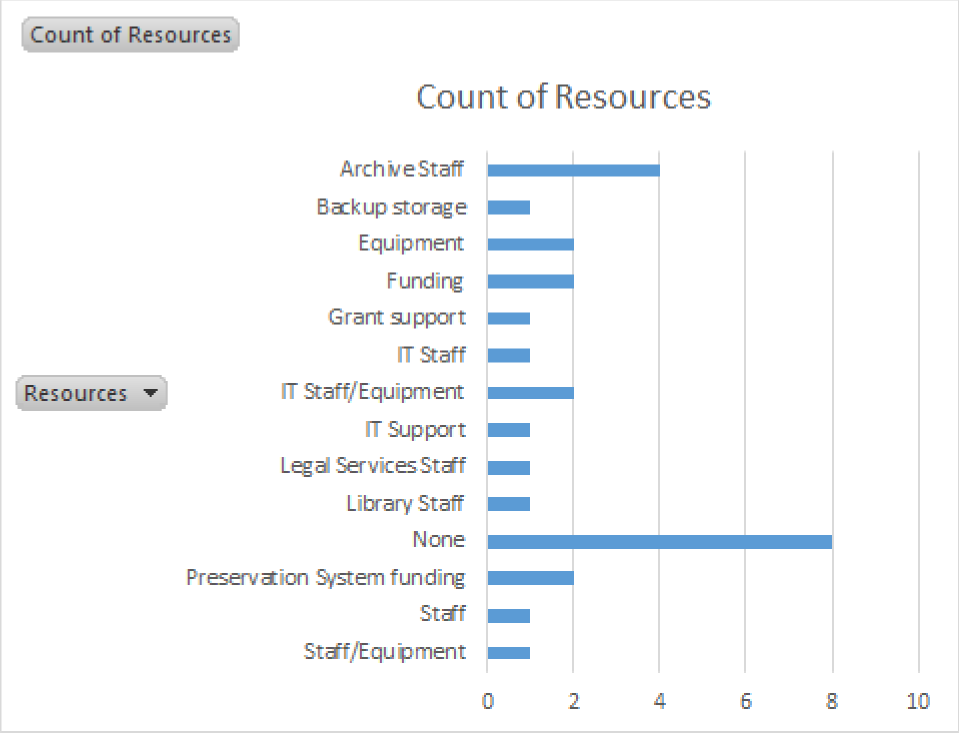
Do you intend to implement a long-term preservation strategy within the next 18 months?

Yes - 50%; **No** - 32%; **Maybe** - 18%

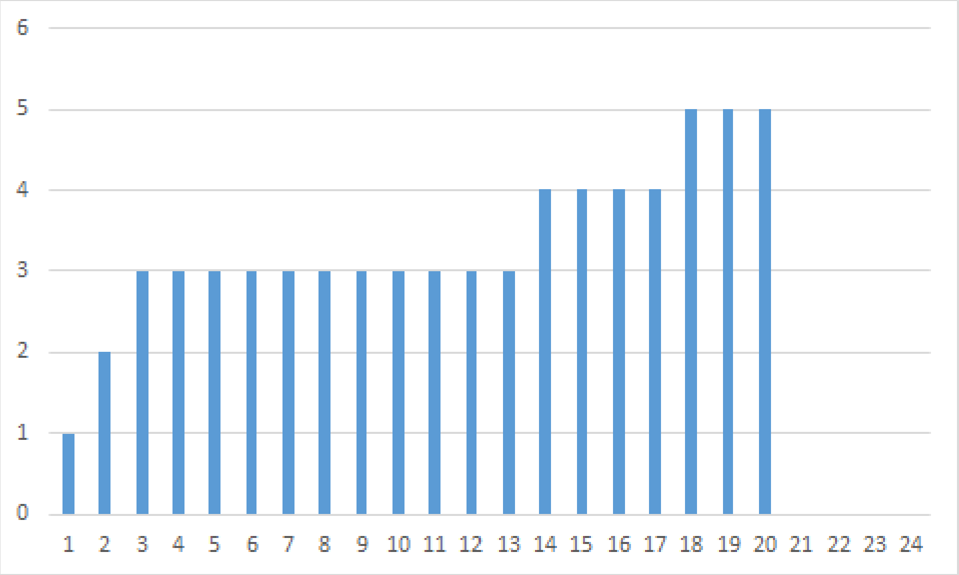
If you do not have a long-term preservation strategy, what are the barriers that you face?



What sort of resources has your state provided for long-term preservation?



How likely would you be to use an open source, out-of-the-box strategy for publication and preservation of electronic legal materials?



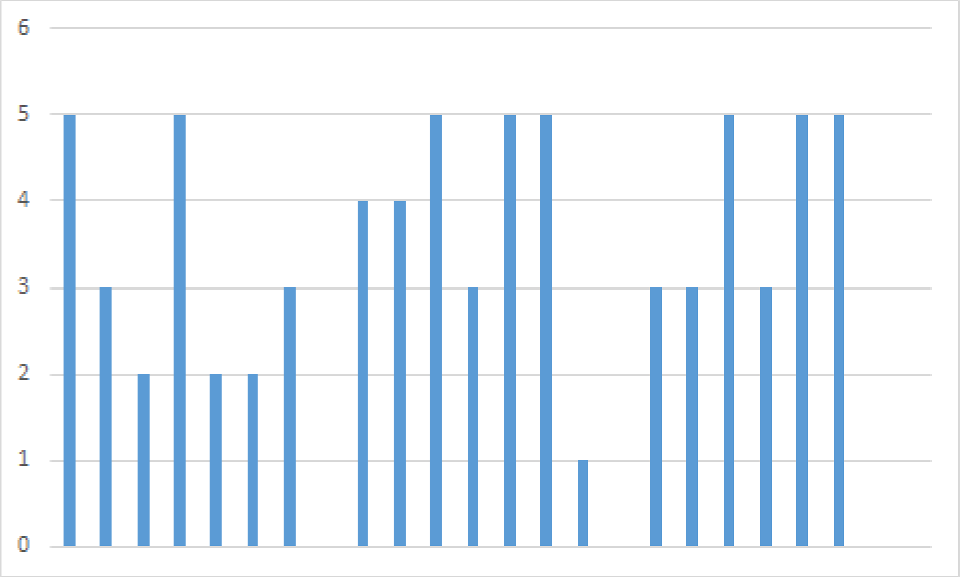
1 = Very likely

5 = Unlikely

Issues:

- Unlikely to engage in preservation of any kind
- Not sure what this would be
- Public trust
- Already developing own strategy
- Support and maintenance
- Staying with print

How likely would you be to participate in an interstate digital storage solution for official electronic documents if one existed?



1 = Very likely

5 = Unlikely

Issues:

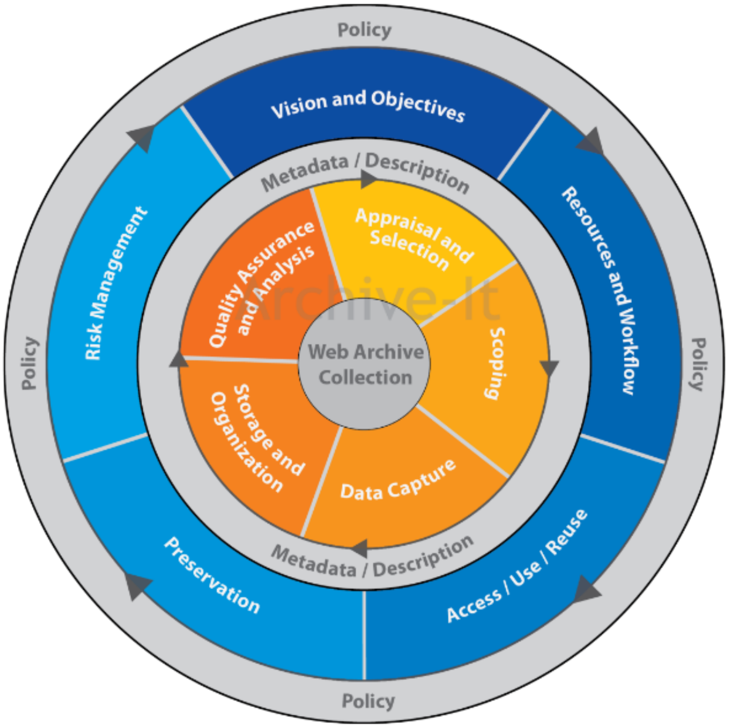
- Unlikely to engage in preservation of any kind
- Resources
- Accessibility; Security
- Constitutional mandates
- State level operation
- Self-sufficient

Appendix II: Open source and commercial preservation systems

Archive-It (www.archive-it.org)

Archive-It is a web archiving service that has been available since 2006 from the Internet Archive. This is the same organization that is responsible for the Wayback Machine (https://archive.org/web), which has been archiving the internet since 1996. Archive-It uses the Heritrix web crawler that was developed by the Internet Archive, and outputs data in the WARC file format, an ISO standard for web archiving.

Archive-It is a subscription service, and is used to archive both the websites of partner institutions as well as topical collections of web sites. The Archive-It Team at the Internet Archive has developed a life-cycle model to help guide in the decision-making needed for a web archiving program. (http://ait.blog.archive.org/files/2014/04/archiveit_life_cycle_model.pdf)



Features:

- Control the extent, depth, and description of collections
- Browse collections by URL, by metadata, and by full-text search.
- Public access via archive-it.org and tools to build custom integrated portals
- File storage in preservation format in multiple, redundant data centers.
- Ability to download WARC files on-demand for local management.

Archivematica (https://wiki.archivematica.org)

Archivematica is an open-source preservation application supported by Artefactual Systems (https://www.artefactual.com). The community is supported using a discussion list, user group meetings, training and workshops, and installation and service agreements, and Artefactual Systems has partnerships that provide Archivematica as a hosted service. The evolution of the software can be tracked on the development roadmap.

Archivematica utilizes a set of micro-services to perform fundamental preservation actions. The system comprises standard and open tools that other services also utilize (for a current list, see https://wiki.archivematica.org/Release_1.6.0), as well as specific python middleware that ties the tools together into services and workflows. A list of the micro-services is available at https://wiki.archivematica.org/Micro-services#Archivematica_Micro-services.

From the Archivematica literature:

“The goal of the Archivematica project is to give archivists and librarians with limited technical and financial capacity the tools, methodology and confidence to begin preserving digital information today. The project has conducted a thorough OAIS use case and process analysis to synthesize the specific, concrete steps that must be carried out to comply with the OAIS functional model from Ingest to Access. Through deployment experiences and user feedback, the project has expanded even beyond OAIS to address analysis and arrangement of transferred digital objects into SIPs and allow for archival appraisal at multiple decision points. Wherever possible, these requirements are assigned to software tools within the Archivematica system. If it is not possible to automate these steps in the current system iteration, they are incorporated and documented into a manual procedure to be carried out by the end user. This ensures that the entire set of preservation requirements is being carried out … In short, the system is conceptualized as an integrated whole of technology, people and procedures, not just a set of software tools. For institutions that want technical assistance to install and customize Archivematica, optional technical support services are provided by Artefactual Systems.”

Arkivum (www.arkivum.com)

Arkivum is a UK-based company that offers three distinct storage solutions for enterprise records, data sets, and cultural heritage assets respectively. Arkivum’s digital asset management and preservation system is called Perpetua. Perpetua can be operated as a cloud-based, managed storage storage solution or as a locally maintained, internal system. It uses Archivematica’s tools for metadata creation, performs scheduled data integrity audits, offers data encryption and access controls, and supports a variety of backup storage options including tape-, cloud-, and disc-based systems, all geographically distributed between the United States and the British Isles. To alleviate uncertainty around its hosted preservation model, Arkivum offers contractual commitments to service that exceed 25 years in duration. They also build into all service agreements a transparent exit strategy should institutions choose to migrate to a new system. Arkivum is still a relatively new company, having just transitioned from an internal project at the University of Southampton to a private business model in 2011.

Duraspace Systems (www.duraspace.org)

Duraspace is a non-profit organization devoted to providing long-term support for Dspace, Fedora, and VIVO. Duraspace also supplies digital asset management and preservation services through its DuraCloud and DspaceDirect platforms:

- ### DuraCloud (www.duracloud.org)

  DuraCloud is a cloud storage and content preservation service offered by Duraspace that backs up assets to multiple cloud storage providers while also offering a suite of preservation tools such as data integrity checks, transfer tools, and scheduled synchronization. DuraCloud is mainly a storage solution and must be integrated with an asset management system like Dspace or Archive-It for the capture of preservation metadata. Clients are given a range of cloud storage providers to store their assets with. These include Amazon Simple Storage Service, Amazon Glacier, San Diego Supercomputer Center, Rackspace Cloud Files, and Chronopolis. Pricing for DuraCloud varies depending upon which cloud services the client decides to use. Costs include an annual subscription fee between $1,235 and $5,520 and a per terabyte storage cost between $500/TB to $825/TB. Chronopolis storage includes an additional ingest fee of $310/TB.

- DspaceDirect (www.dspacedirect.org)

  DspaceDirect is a hosted DAMS service wherein a client can contract Duraspace to maintain a cloud-based instance of Dspace for a fee. Since the DspaceDirect system integrates with Duracloud, it is possible to use a DspaceDirect repository as a preservation system. Duracloud is another Duraspace service that offers cloud-based archiving and preservation functions such as checksums, file redundancy, content migration, and access control. Pricing for DspaceDirect can vary greatly depending upon extent of storage needed. A relatively small 250GB allotment of storage costs, as of September 2017, is $8,670.

- Fedora (fedorarepository.org)

  Not to be confused with Red Hat's Linux operating system by the same name, FEDORA, which stands for Flexible Extensible Digital Object Repository Architecture, was developed by the Digital Library Research Group at Cornell University in the 1990s. It is an open source repository system, which offers tools for management and dissemination of digital assets. It is notable for its flexibility and modularity. It can be configured to accept any file type or metadata schema. Additionally, its features can be controlled via an extensive set of APIs, allowing for integration with external applications and devices. Because of this modularity potential, Fedora itself operates as a kind of skeleton platform. It offers several core functions like storage, relational connections, and basic ingest, but most advanced functions that one comes to associate with a typical digital asset management system (DAMS) are integrated into the system as add-ons. Islandora and Samvera (see below) are two such suites of software tools that can be added onto Fedora to facilitate digital asset management and preservation.

Islandora (islandora.ca)

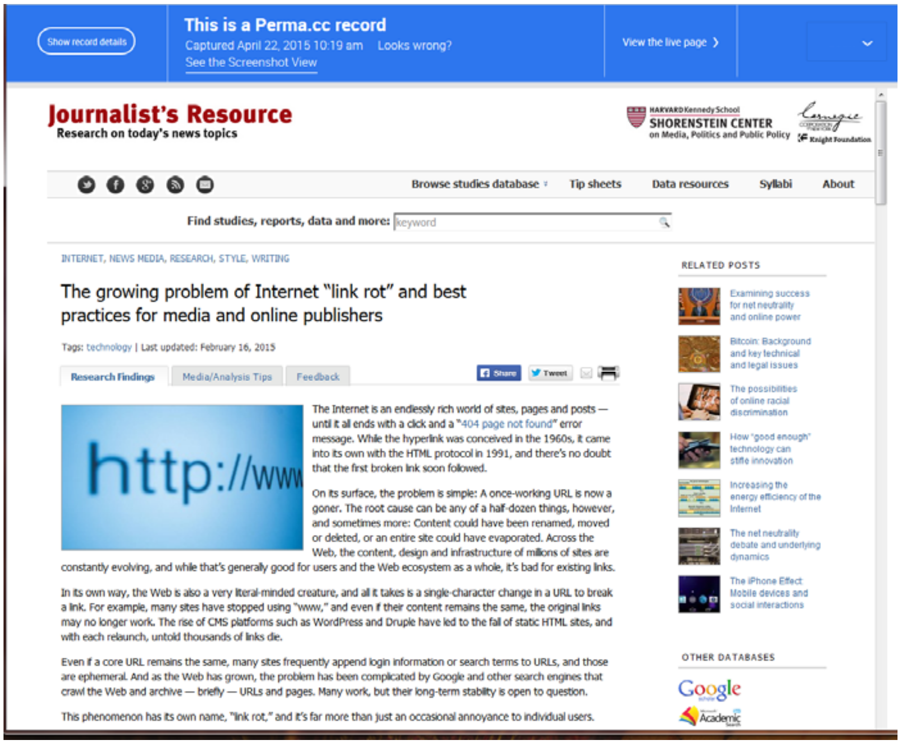
Originally developed by affiliates of the University of Prince Edward Island, Islandora is a Drupal-based framework of digital repository tools that integrate with Fedora. Islandora is an open source platform whose core, components, and documentation are maintained by a growing community of contributors from around the world. Islandora's main components include Drupal web interfaces for administrative functions and end-user experience, Solr search engine for asset discovery, and special content models for different asset types like pdfs, video files, large image files, etc. Specific features are added to Islandora using Drupal's module and theme systems. The Islandora community has created a number of modules that carry out preservation-related operations. Some of these include:

- Islandora Pathauto/Islandora Handle/Islandora DOI for implementing persistent URLs.
- Islandora PREMIS for supporting production and storage of preservation metadata.
- Islandora FITS/Islandora Checksum/Islandora Checksum Checker for carring out data integrity functions like checksum generation, file format identification, and technical metadata extraction.
- Islandora BagIt for depositing backup assets to a BagIt preservation archive.
- Islandora Vault for for depository backup assets to CloudSync or DuraCloud.
- Islandora LOCKSS-O-Matic for depositing assets into a Private LOCKSS Network.

Perma.cc (https://perma.cc)

Perma.cc is a tool built by Harvard University's Library Innovation Lab to specifically combat link rot in citations. It is an online service that will archive the web page for a given URL and add it to the Perma.cc collection and return a unique URL (e.g. "perma.cc/ABCD-1234") that points to the record in the collection. When that URL is then used in a citation, it will give readers a stable view of the page at the time it was archived (even if the original disappears from the web), as well as a link to the page as it currently exists.

Perma.cc is a free service, and anyone can create an account, but unless associated with a vetted organization, a user will be limited to creating 10 permalinks per month. Once an organization has joined Perma.cc, unlimited user accounts can be created and those users can create unlimited permalinks, as long as the links are saved within the member organization on Perma.cc.



Show record details
This is a Perma.cc record
Captured April 22, 2015 10:19 am
Looks wrong?
See the Screenshot View
View the live page
Journalist's Resource
Research on today's news topics
HARVARD Kennedy School
SHORENSTEIN CENTER
on Media, Politics and Public Policy
Knight Foundation
Browse studies database
Tip sheets
Data resources
Syllabi
About
Find studies, reports, data and more:
keyword
INTERNET, NEWS MEDIA, RESEARCH, STYLE, WRITING
Research Findings
Media/Analysis Tips
Feedback
Share
Tweet
http://www
RELATED POSTS
Examining success for net neutrality and online power
Bitcoin: Background and key technical and legal issues
The possibilities of online racial discrimination
How "good enough" technology can stifle innovation
Increasing the energy efficiency of the Internet
The net neutrality debate and underlying dynamics
The iPhone Effect Mobile devices and social interactions
OTHER DATABASES
Google scholar
Academic Search

The growing problem of Internet "link rot" and best practices for media and online publishers

Tags: technology | Last updated: February 16, 2015

The Internet is an endlessly rich world of sites, pages and posts — until it all ends with a click and a "404 page not found" error message. While the hyperlink was conceived in the 1960s, it came into its own with the HTML protocol in 1991, and there's no doubt that the first broken link soon followed.

On its surface, the problem is simple: A once-working URL is now a goner. The root cause can be any of a half-dozen things, however, and sometimes more: Content could have been renamed, moved or deleted, or an entire site could have evaporated. Across the Web, the content, design and infrastructure of millons of sites are constantly evolving, and while that's generally good for users and the Web ecosystem as a whole, it's bad for existing links.

In its own way, the Web is also a very literal-minded creature, and all it takes is a single-character change in a URL to break a link. For example, many sites have stopped using "www," and even if their content remains the same, the original links may no longer work. The rise of CMS platforms such as WordPress and Druple have led to the fall of static HTML sites, and with each relaunch, untold thousands of links die.

Even if a core URL remains the same, many sites frequently append login information or search terms to URLs, and those are ephemeral. And as the Web has grown, the problem has been complicated by Google and other search engines that crawl the Web and archive — briefly — URLs and pages. Many work, but their long-term stability is open to question.

This phenomenon has its own name, "link rot," and it's far more than just an occasional annoyance to individual users.

Preservica (www.preservica.com)

Preservica is a private digital preservation company that operates out of Boston and Oxford. It offers services across the digital asset lifecycle, not just for long term preservation. The Preservica platform is available as a fully hosted, cloud-based service or as an on-premise, locally hosted, customizable installation. Preservica repository adheres to OAIS ISO 14721 preservation standards. It offers tools for metadata creation and harvesting, simple ingest workflow, multiple storage choices, large file transfer, access control, and active file format identification and migration.

Rosetta (www.exlibrisgroup.com/category/RosettaOverview)

Rosetta is a digital asset management and preservation system produced by Ex Libris that offers full lifecycle support for any digital format. Though Rosetta is proprietary software, it exposes parts of its architecture to third-party integration with APIs. Administrators can connect Rosetta to a separate storage device or devices, if desired. Rosetta's workflow models are configurable. It generates checksums, identifies file formats and extracts technical metadata at ingest. The Rosetta preservation planning module enables administrators to schedule data integrity and migration tasks as needed. Rosetta uses the PREMIS data model for collecting preservation metadata. The system's architecture is divided between an operational repository, where functions like publishing and delivery are carried out, and a permanent repository, where preservation functions and long term storage take place.

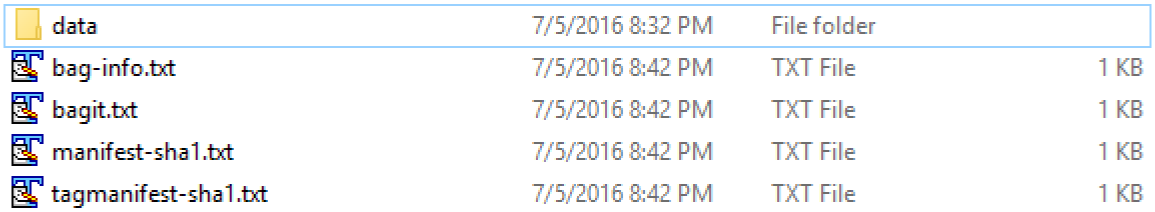
Samvera (samvera.org)

Formerly known as Hydra, Samvera is similar to Islandora in that it integrates with Fedora repository software to provide search engine and interface layers; however, rather than using Drupal to facilitate this, Samvera uses a Ruby on Rails plugin called Blacklight. The Blacklight framework is a web platform that is specifically designed for resource discovery with Solr Indexes. Samvera has been adopted by a number of major digital libraries as a digital asset management system. In 2014, the Hydra user group decided to pursue options for adding digital preservation functionalities to the Hydra/Samvera stack. As of September 2017, a Samvera Digital Preservation Interest Group that is investigating the matter.

Appendix III: Stand-alone Preservation Tools

Aside from preservation systems that strive to accomplish an end-to-end OAIS-compliant preservation workflow, there are a number of open source tools that will perform different and discreet parts of the process. Here are some that are in widespread use:

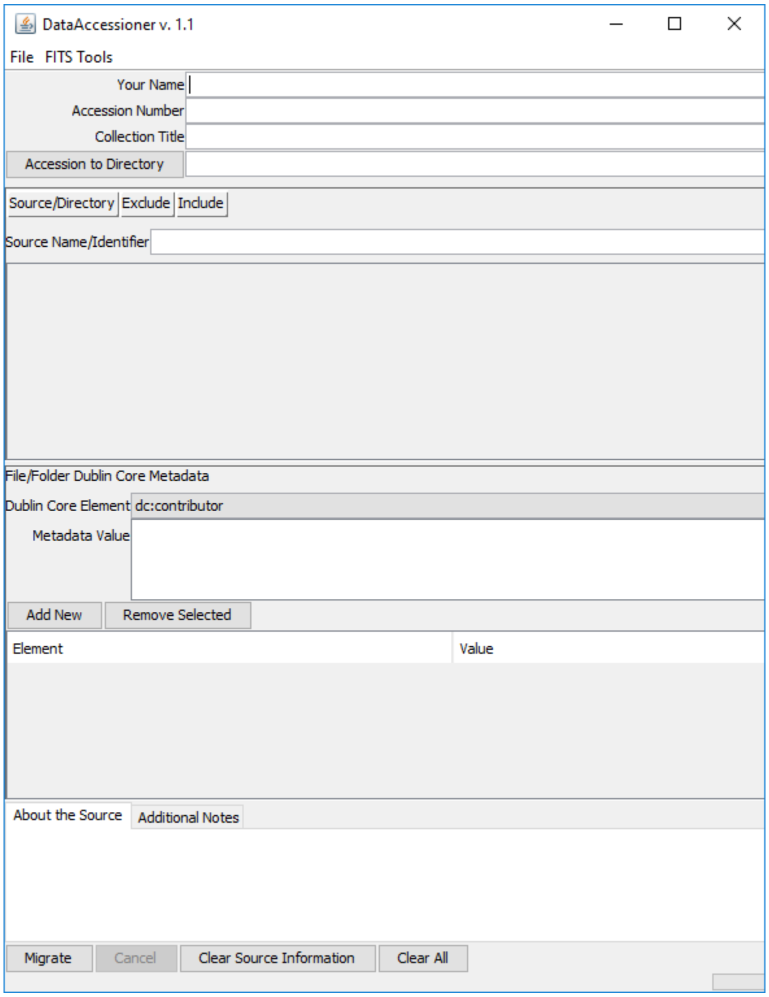
- BagIt
  Developed by the Library of Congress and the California Digital Library to define a strategy for transferring digital content. It specifies the elements and structure of a "bag" that includes the files in a standard container. The tool will create a manifest of checksums of all of the digital files in that container as well as metadata about the package. On receipt of the package, the checksums can be validated to make sure that no corruption occurred during the transfer.



data	7/5/2016 8:32 PM	File folder	
bag-info.txt	7/5/2016 8:42 PM	TXT File	1 KB
bagit.txt	7/5/2016 8:42 PM	TXT File	1 KB
manifest-sha1.txt	7/5/2016 8:42 PM	TXT File	1 KB
tagmanifest-sha1.txt	7/5/2016 8:42 PM	TXT File	1 KB

Example of a BagIt bag

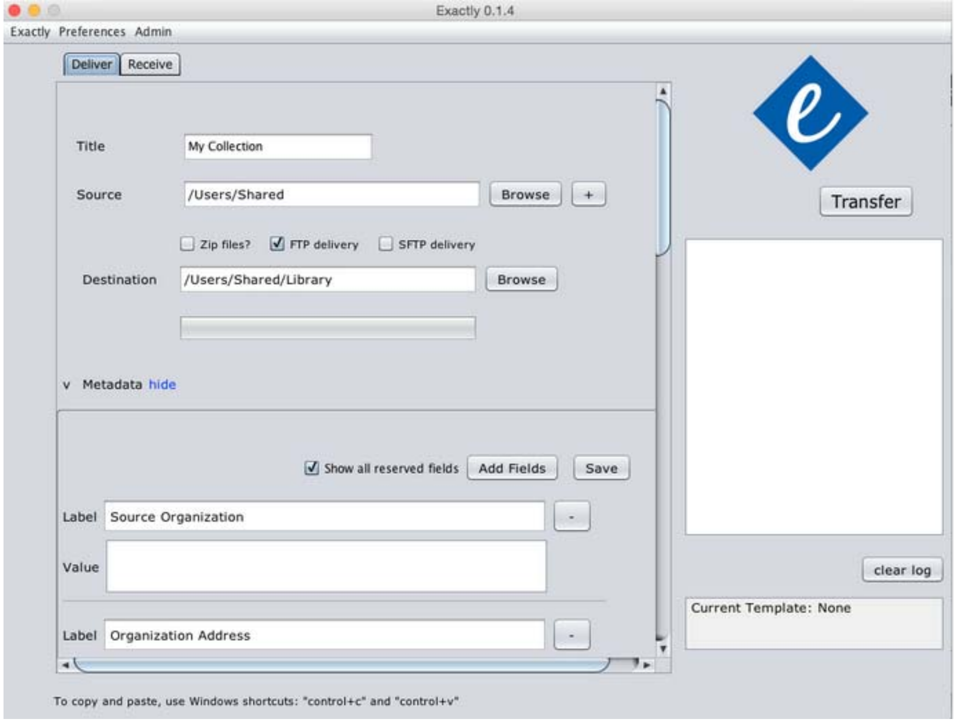
- Data Accessioner
  A simple tool for transferring files from one media to another. It too will create checksums and gives the user the option of creating Dublin Core metadata.



DataAccessioner v. 1.1

Dublin Core Element: dc:contributor

- Exactly
  Another transfer tool – from the website: *"Exactly allows recipients to create customized metadata templates for senders to fill out before submission. Exactly can send email notifications with transfer data and manifests when files have been delivered to the archive."*



Title: My Collection

Source: /Users/Shared

☐ Zip files? ☑ FTP delivery ☐ SFTP delivery

Destination: /Users/Shared/Library

☑ Show all reserved fields

Label: Source Organization

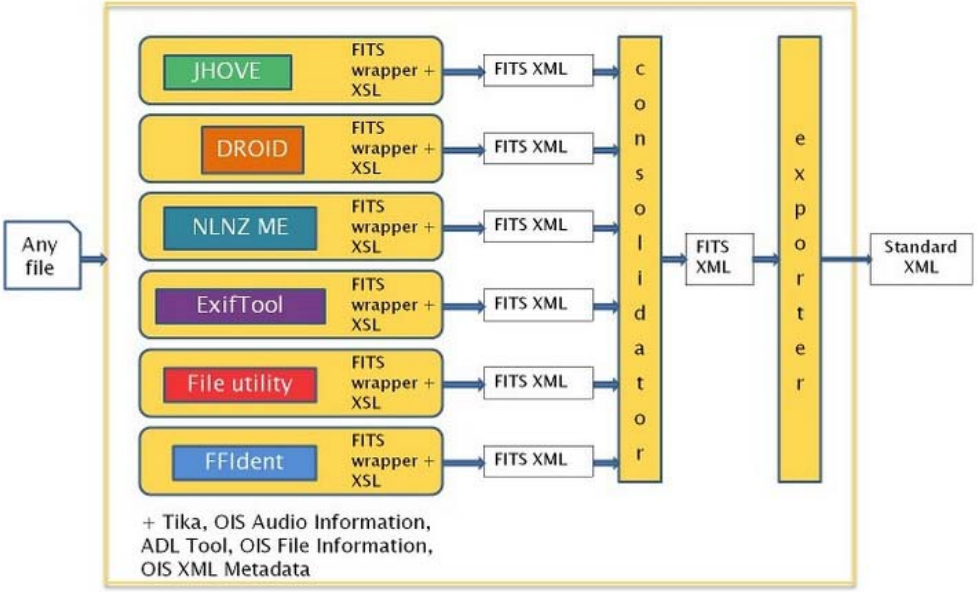
Value:

Label: Organization Address

Current Template: None

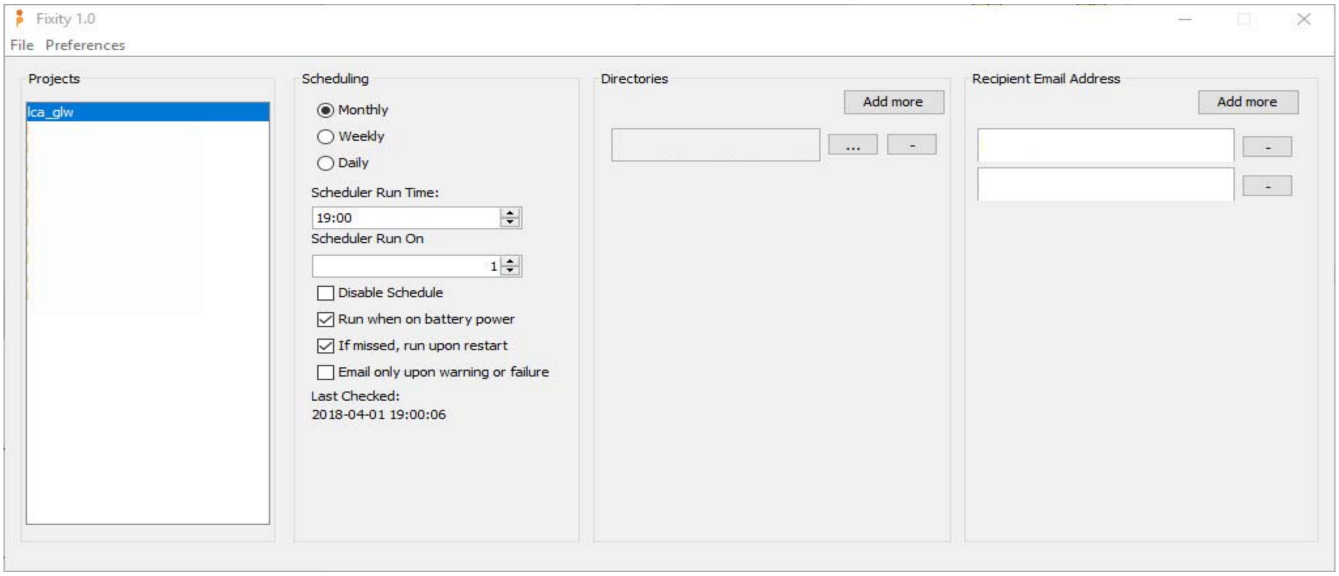
To copy and paste, use Windows shortcuts: "control+c" and "control+v"

- FITS
  This tool consolidates the use of other open source tools for the purpose of file characterization. There are a number of individual tools that will identify a file format and output various pieces of technical information about that file. Because each individual tool has strengths and weaknesses, it is good practice to run files through multiple characterization tools, and FITS will do this in one process and output a consolidated set of data.



From the FITS online User Manual

- Fixity
  While several of the tools mentioned above will create checksums that act as something like a digital fingerprint of a file, without the ability to validate that checksum periodically, it serves no good purpose. Fixity is a tool that allows for the scheduling of regular checksum validations, and will send a report on the results.



lca_glw

● Monthly
○ Weekly
○ Daily

Scheduler Run Time: 19:00

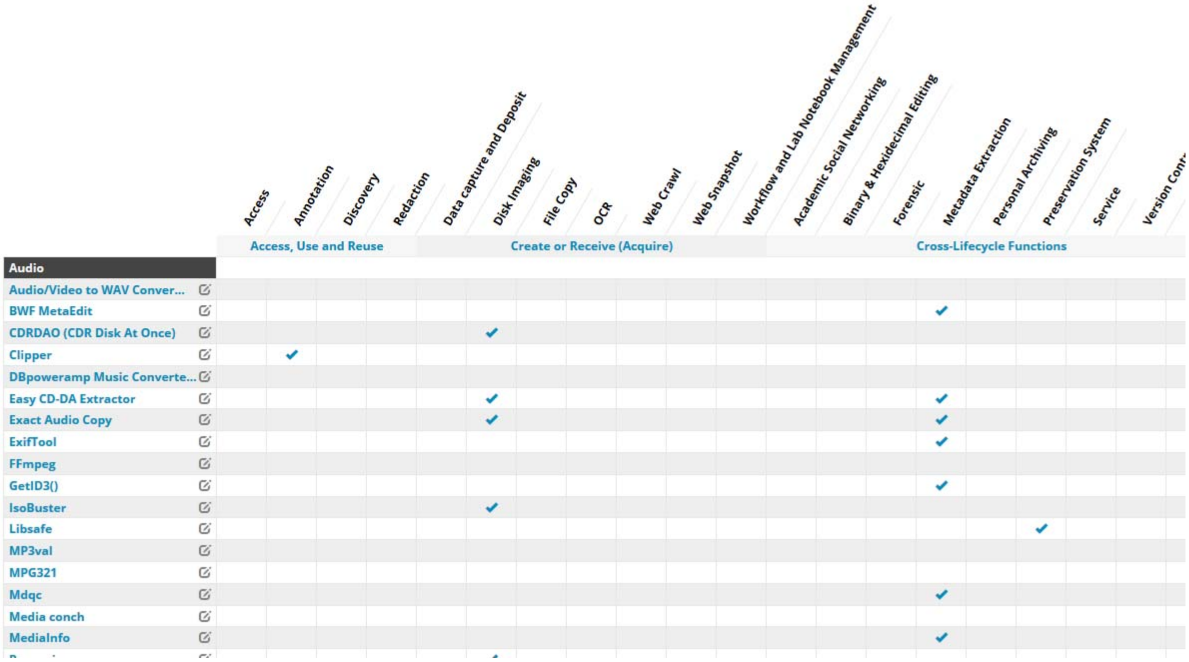
Scheduler Run On: 1

□ Disable Schedule
☑ Run when on battery power
☑ If missed, run upon restart
□ Email only upon warning or failure

Last Checked:
2018-04-01 19:00:06

For an exhaustive list of available tools and systems for digital preservation, go to the POWRR Tool Grid v2

POWRR – **P**reserving Digital **O**bjects **W**ith **R**estricted **R**esources



	Access	Annotation	Discovery	Redaction	Data capture and Deposit	Disk Imaging	File Copy	OCR	Web Crawl	Web Snapshot	Workflow and Lab Notebook Management	Academic Social Networking	Binary & Hexidecimal Editing	Forensic	Metadata Extraction	Personal Archiving	Preservation System	Service	Version Cont-
	Access, Use and Reuse				Create or Receive (Acquire)							Cross-Lifecycle Functions							
Audio																			
Audio/Video to WAV Conver...																			
BWF MetaEdit															✓				
CDRDAO (CDR Disk At Once)						✓													
Clipper		✓																	
DBpoweramp Music Converte...																			
Easy CD-DA Extractor						✓									✓				
Exact Audio Copy						✓									✓				
ExifTool															✓				
FFmpeg																			
GetID3()															✓				
IsoBuster						✓													
Libsafe																	✓		
MP3val																			
MPG321																			
Mdqc															✓				
Media conch																			
MediaInfo															✓				