SAGESSE

ISSN 2560-7014

Integrate Digital Preservation into Your Information Governance Program

Advance your understanding and advocacy for long-term electronic records preservation

by Lori J. Ashley

Introduction

Public, private, and voluntary sector organizations are required to keep information and records for as long as necessary to meet a range of legal, financial, operational, regulatory compliance, and cultural memory purposes, depending on their unique mission, history and objectives. An exponentially increasing number of valued information assets are born and will spend their entire "lives" in the digital realm. For those assets that must be retained long-term, there are risks and serious threats from technology obsolescence that should be systematically considered and proactively addressed.

Since the 1970's archivists around the world have recognized that the obsolescence of digital media and storage devices was a risk to accessing authentic, readable, and usable electronic records. Dependence on computer software applications, many of which use proprietary file formats to create, save, store and manage, and retrieve information assets, poses an additional threat. The emergence of technology neutral open standard formats (e.g., JPG, PDF/A, SVG) now commonly used by records producers while the assets are still in their custody has partially addressed the risk but are often inconsistently used and not a one-time solution for indefinite long-term or permanent records. The emergence of Information Governance as a coordinating accountability framework for enterprise lifecycle management is a positive development that can be leveraged to address the vulnerability of long-term digital information.

Given the reliance and economic impact of computer technologies used by all sectors to deliver goods and services, digital preservation strategies and capabilities should become part of every organization's information governance program. Records and archives management professionals are uniquely positioned as part of their established advisory, appraisal, scheduling, and custodial roles to help their organizations assess the maturity of hosted and on-premise information systems to handle the demands of long-term and permanent electronic information assets.

How Long is Long-Term?

Long-term is defined in the digital preservation community's de facto standard as a period of time that is "long enough to be concerned with the impacts of changing technologies, including support for new media and data formats, or with a changing user community. Long Term may extend indefinitely."[1] Technology refresh cycles are relatively short, however, in the three-to five-year range. Like hardware and storage media, file formats can become outdated, obsolete or unsupported, often without users realizing it. Think of once popular formats like Lotus 1-2-3 and WordPerfect, which have come and gone. When obsolescence happens, accessibility becomes a challenge.

The benchmark for long-term retention commonly applied by preservation practitioners is 10 years[2] due to known risks associated with bit corruption, broken links, abandoned or decommissioned applications, and media degradation. Traditional data archiving approaches and common electronic records fixity methods (e.g., saving to a uniform file format like PDF) are proving insufficient to address exponential growth and diversity in the types of digital information that need to be protected from rapid technological change.

The fragility of digital content as a significant potential risk to businesses has been identified by leading analyst firms that include Forrester, Info-Tech Research Group, and Gartner. In 2015, Gartner noted: "As formats change, software is retired and hardware becomes obsolete, the data that organizations might want to keep can be lost forever."[3] These warnings about risks associated with hardware and software obsolescence, link rot, storage degradation, and vendor abandonment are being directed to CIOs, CTOs, and enterprise architects to raise awareness about adding to the organization's technical debt.[4]

Defining Digital Preservation

Digital preservation is a formal set of processes and activities that maintain permanent and indefinite long-term information stored in digital formats in order to ensure continued access. The preservation of digital information is widely considered to require more proactive and continuous attention than preservation of other media.

Digital preservation actions must be taken over the lifetime of electronically stored information due to changes in software and hardware environments, deterioration of magnetic media such as CDs, DVDs, computer hard drives, and to keep pace with evolving business, legal and regulatory requirements for access and re-use. Digital objects in a preservation system are actively migrated over time to newer formats using policies and automated workflows. Preservation actions are captured in the metadata associated with each digital object to demonstrate authenticity and chain of custody.

---

[1] Consultative Committee for Space Data Systems (2012) 'Reference Model for an Open Archival Information System (OAIS)', CCSDA Secretariat, Washington, D.C., p. 1-1.

[2] Peterson, M., Zasman, G. Mojica, P. and Porter, J. (2007) '100 Year Archive Requirements Survey', Storage Networking Industry Association (SNIA), Colorado Springs, CO.

[3] https://www.gartner.com/en/documents/3038917/cool-vendors-in-content-management-2015

[4] https://www.techopedia.com/definition/27913/technical-debt

Digital objects are organized into collections and shared in accordance with the organization's unique permissions and rights.

Many commercial and open source digital preservation tools are based on the Open Archival Information System (OAIS) reference model. OAIS is a conceptual framework for functions and actions that a digital repository must execute to ingest, store, preserve, and provide access to digital objects for a community of users.

First released in 2003 as ISO 14721 and updated in 2012, the standard has helped to raise awareness and understanding of concepts relevant for archiving digital objects, clarify terminology for comparing data models and archival architecture, expand consensus on the infrastructure and processes needed for digital information preservation and access, as well as guide the development of supporting standards.

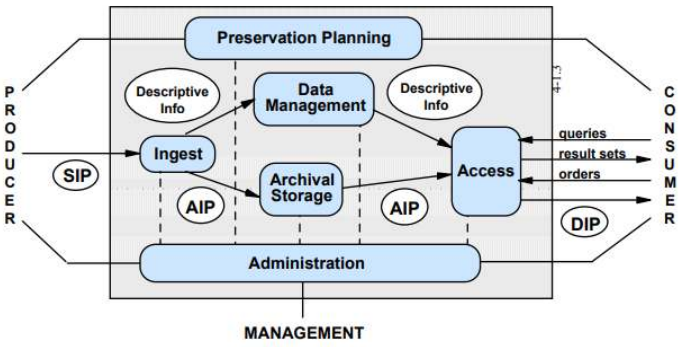

Figure 4-1: OAIS Functional Entities

Common activities for digital repositories conforming to the OAIS reference model include file ingest and characterization, integrity validation and protection, collection management, system and data monitoring, migration of assets from obsolete file formats, replication to multiple geographic locations, robust metadata management to facilitate search and retrieval, and secure access.

A companion standard, Audit and Certification Criteria for Trustworthy Digital Repositories (ISO 16363:2012),[5] includes the OAIS technical functions in addition to identifying organizational and security management capabilities and metrics. Dedicated resources - funding, skilled staff, tools, storage, and, organizational commitment -- are required for a digital repository to persistently monitor risks and adapt to changing conditions.

Two preservation repositories in the world have been certified to ISO 16363 thus far: The Audio/Visual Repository at the Indira Gandhi National Center for the Arts and the United States Government Publishing Office (GPO.) A self-assessment checklist is

[5] Consultative Committee for Space Data Systems, *Auditing and Certification of Trustworthy Digital Repositories: Recommended Practice CCSDS 652.0-M-1.* September 2011. Accessed at http://public.ccsds.org/publications/archive/652x0m1.pdf

available from PTAB – Primary Trustworthy Digital Repository Authorisation Body Ltd[6] – that practitioners can study to learn about the metrics and documentation required to demonstrate that a repository is trustworthy and able to protect and preserve its digital holdings over time.

Active Digital Preservation

The primary role of a digital preservation system is to ensure objects and their metadata remain accessible, useable, and readable over the long-term by providing a proactive way to migrate file formats as they become obsolete or are no longer supported by a vendor or by the organization. Core functions of an integrated digital preservation solution that conforms to the ISO 14721 standard provide for:

Content Ingest – The Ingest function allows users to upload information into the repository. Various methods include simple drag and drop via a web browser, via a secure holding area, using APIs, as well as an optional bulk upload service. Ideally the content hierarchy can be imported from external systems, re-arranged before upload, or modified after the objects are ingested.

Integrity Checks and Metadata Extraction – System workflows during the Ingest process perform quality assurance actions that include virus checking, checksum verification, file format checking against a format registry, as well as metadata extraction and validation. Some organizations choose to migrate (normalize) file formats on ingest using migration pathways. The system also detects duplicate files.

Archival Storage – It is common archival practice to preserve a master file and to create access copies in order to protect the authenticity and integrity of the original object and its metadata. Archival storage means that files are saved to multiple servers in multiple data centers within a given region. All objects held in storage are integrity checked and checksums are calculated on access or at regular time intervals. Archival storage is dependent on other preservation services that include media renewal, security protections, and the availability and enforcement of preservation metadata standards.

Data and Collection Management - A range of capabilities in a digital preservation system allows repository administrators to manage the metadata that describes the content in their collections. Options include providing a custom metadata schema, selecting/editing one or more of standard metadata schemas, adding custom search indexes, viewing audit trails, creating and modifying content hierarchies, synchronizing metadata and collection structures from

[6] The Primary Trustworthy Digital Repository Authorisation Body Ltd (PTAB) is accredited by the National Accreditation Board for Certification Bodies of India (NABCB) to conduct ISO 16363 audits worldwide. A self-assessment checklist is available at http://www.iso16363.org/iso-certification/preparation/

catalogs. It may also be possible to classify personal data information to meet GDPR and other privacy requirements.

Access – Repository administrators can facilitate browsing, searching, and viewing of content as well as display technical and descriptive metadata through a user interface. Search tools support full text search of permitted file formats and metadata as well as limited search within a selected folder. Administrators can launch access workflows to transform files, send notifications, and create content packages for download or delivery. The solution may include a web-based portal that can be customized for internal access or controlled public access to collections including fast playback of audio-visual content. Organizations can also choose to control and provide access to collections through their existing web portals.

Preservation Planning and Action – The purpose of a digital preservation repository is to ensure the longevity and viability of content and metadata over time. Repository administrators have options to identify file formats, select preservation and presentation actions, and apply migration pathways on individual assets or collections. Standard tools include the ability to create new "Digital Master" preservation copies, create presentation copies for sharing, and implement and monitor all preservation actions via a standard workflow engine.

Administration – Through standard and customized reports, administrators can monitor all functions of the repository including storage and file types. User permissions, roles, and options for two factor authentications are also part of system administration. Tags can be assigned to assets and folders to set user access as well as define which actions can be performed by each role.

Who Uses Digital Preservation Technology?

Curators, librarians, and archivists have long recognized digital preservation capabilities and resources would be needed to protect their digitized and born-digital assets of historical and cultural value. This preservation community represents national, state and provincial libraries, archives and special collection repositories as well as museums, private and corporate archives, and academic institutions around the world.

Many of these organizations are dealing with aging digitized collections of documents, photographs and audio/visual assets, as well as experiencing exponential increases in the volume and diversity of born-digital assets that are being transferred or donated for permanent preservation. The Internet and social media channels enable repositories to make their collections more widely and easily accessible, providing additional incentives to organize and tag digital assets for search and retrieval.

Public Sector Archives

The Library and Archives Canada (LAC), one of the world's largest public sector library and archival institutions, released its strategy for a digital preservation

program in 2017, describing a multi-phase initiative to preserve its existing seven petabyte historical collection and prepare for millions of born-digital government records to be accessioned in the future. Numerous agencies will transfer permanent records from the enterprise GCDOCS (OpenText) system so LAC is working to implement scalable, automated file ingest from provincial and federal agencies to meet the demand.

Academic Archives and Libraries
Founded in 1846, the State University of New York at Buffalo (UB) is the largest campus in the 64-campus public university system. The University Archives preserves and provides public access to varied collections dating back to the 1800's that include student and campus publications, papers of prominent people and area organizations, as well as the private documents of faculty members. The Archives had identified a large amount of content on storage media that was sometimes difficult to identify and classify, and which was at risk of file degradation. A digital preservation system was implemented to speed discovery times and ensure that historical resources remain accessible for the long-term. UB integrated a digital preservation system with its existing ArchivesSpace catalog to increase the amount of digital material available to the university community and the public. Faculty members use the digital archive as part of their curriculum and students reference the collections to complete their capstone projects.

Preservation of electronic information and records is also vitally important for commercial organizations whose global communications, governance, intellectual property protection, brand and asset management, knowledge management, and critical operational processes now rely heavily on born-digital content.

Corporate Archives
Established in 1846, the Associated Press (AP) is an independent news cooperative headquartered in New York City, with photojournalists working in over 100 countries to tell the world's stories. In 2015 the Corporate Archives implemented a cloud-based digital preservation system to systematically acquire, organize, preserve and make accessible the organization's corporate governance records, including Board of Director minutes and monthly CEO memorandums, changes to charters and bylaws, oral histories, documentaries, and financial reports. The historical news collection includes the original wire copy from notable events such the assassination of U.S. President John F. Kennedy as well as award-winning photographs. AP's digital corporate archive has become a trusted source of invaluable resources for departments across the organization that include marketing, legal services, strategic planning, and public relations.

Digital Preservation Moving to the Mainstream

All modern organizations rely heavily on records and information managed exclusively in digital format. In any large enterprise there are likely to be records managed in hundreds of different file types and formats - images, audio and video files, websites, social media, electronic messages, case files, maps, health care documents, as well as

artefacts and assets of historical interest. And many of these assets must be retained over successive generations of hardware and software technologies to remain authentic and usable.

Retention periods for records and information vary widely from brief (e.g., 30 days) to indefinite (e.g., termination plus 30 years) to perpetual (e.g., permanent). Use cases for long-term digital records preservation for the energy, utility, pharmaceutical, financial services and insurance, and consumer goods industries are well documented. Legal liability, digital transformation, mergers/acquisitions, regulatory compliance, application decommissioning, cybersecurity, and other mainstream risk management concerns have made defensibility of data transfer and archiving a compelling concern for IT infrastructure and architecture.

Operational efficiency use cases include secure retention of corporate governance materials, intellectual property, and brand assets. Industry use cases include the retention of pharmaceutical research and clinical trial data to comply with global food and drug regulations; the retention of life insurance and pension records for the lifetime of the client and beyond; and retention for life of equipment in the construction, energy, and utility sectors. A public sector use case is the mandate for recorders of real property documents to make them accessible and reproducible forever.

Most business and enterprise content management applications that are used to manage and store business records do not have adequate capabilities to perform preservation actions. Metadata practices across an enterprise or between business units are often incomplete or inconsistent, hampering the search and re-use of documents over time.

In vast collections of electronic records, tracking file obsolescence, and monitoring media degradation is a complex and resource intensive effort. Digital preservation systems designed to conform to the OAIS model are purpose built to address all the core elements of successful long-term digital preservation, combining durable storage, format migrations, and on-going monitoring.

Successful and sustained digital preservation involves compliance with business rules, records management policies, and data governance principles. Record producers need guidance and support to determine the most efficient, systematic, and controlled approaches and triggers to ensure preservation actions are taken proactively. Record and information management professionals can help to meet these needs by:

- Analyzing and tracking file formats in use
- Identifying systems and repositories where long-term electronic records are stored
- Assessing the active preservation capabilities of systems and electronic record repositories
- Analyzing retention schedules to identify and prioritize ‘at risk’ digital files
- Updating transfer protocols to ensure preservation actions are taken as needed, regardless of whether the records are active, inactive or archival

A Call to Action

While digital preservation has primarily been the focus of archival communities in the past, the world's reliance on computer technologies is moving the discipline into the purview of records management and information governance. Like other allied professions, records management must continue to adapt its practices to address emerging challenges associated with electronically stored information while applying principles and proven methods that ensure compliance and mitigate risk.

Ensuring that unique long-term digital information can be easily found, used and trusted in the future is critical to the mission and success of your organization - regardless of sector or industry. Records managers, archivists, and information governance professionals are well positioned to positively impact their organization's data strategies and preservation practices by focusing on lifecycle requirements and leveraging established practices for protecting the integrity, authenticity, and usability of electronic records.

Key initiatives that may provide opportunities for ARMA and allied professionals to engage stakeholders and advance preservation capabilities for long-term digital assets include:

- Business case development for new or upgraded systems
- Planning for legacy application decommissioning
- Developing preservation and transfer requirements for technology procurements
- Digitization projects (paper to digital and film to digital)
- File classification and defensible disposition initiatives
- File and systems migrations, e.g., Office 365
- Business reorganization, mergers, consolidations and acquisitions
- Employee onboarding and training

Digital preservation requires resources, discipline, and commitment. Through consultation with content creators and users, information governance and records management practitioners can help educate their leadership and peers about the fragility of digital information and promote available preservation tools and community-endorsed practices.

Appendix

Preservica is a privately held company with offices in Boston, Massachusetts (US) and Abingdon, Oxfordshire (UK) that offers SaaS and on-premise solutions which combine essential technical functions of long-term digital preservation into a single integrated platform. Trusted by more than 220 organizations in 14 countries around the world, Preservica's standards-based software has been designed from the ground up to tackle the unique challenges associated with ensuring that fragile digital information remains accessible and trustworthy over decades.

Preservica's active digital preservation solution is architected with extensive workflows and APIs to ingest and manage long-term and permanent digital information transferred from a wide variety of source systems. Files are ingested in their native format and actively migrated to newer formats over time and at scale, ensuring the digital information remains useable and readable. Preservica is storage-agnostic, offering flexibility for where and how data is stored and helping to avoid vendor lock-in.

Since 2012 Preservica teams have collaborated with practitioners from public sector archives and libraries, cultural heritage institutions, corporate archives, and academic institutions to shape the development of tools and techniques capable of ensuring the authenticity of digital information and records that organizations need or want to retain and access long-term. In addition to regularly gathering and analyzing requirements from our vibrant user community to inform the company's product roadmap, Preservica supports research and advocacy with leading professional associations that include: Information Governance Initiative (IGI), Archives and Records Association (ARA), Information and Records Management Society (IRMS), ARMA International, Council of State Archivists (CoSA), the National Association of Government Archives and Records Administrators (NAGARA), Digital Preservation Coalition (DPC), and AIIM.

www.preservica.com

Lori J. Ashley
Industry Market Development Manager
Preservica

Lori Ashley is Preservica's Industry Market Development Manager where she analyzes cross-industry and sector-specific requirements for long-term and permanent records retention and use. Research into archives and records management needs in corporate, government, and voluntary sector organizations are used to develop digital preservation use cases, support marketing initiatives, as well as inform Preservica's product roadmap.

Lori is Preservica's liaison to the Council of State Archivists (CoSA) and the National Association of Government Archivists and Records Administrators (NAGARA). She is a long-time member of AIIM, SAA, and ARMA, and has served in numerous board positions for the ARMA Madison (WI) chapter.

Lori joined Preservica in 2017 after 14 years as an independent management consultant and educator who advised government and commercial organizations on information governance and records management. She is co-developer with Dr. Charles Dollar of the Digital Preservation Capability Maturity Model (DPCMM) and self-assessment tool (www.DigitalOK.org) which has been used by CoSA's membership and more than 120 organizations worldwide. Lori served as Records Coordinator for the Wisconsin Department of Electronic Government. Before her public sector service, she was a business and regulatory strategist for an energy company.