CoSA & Preservica Practical Digital Preservation 2017

Best Practices in Digital Preservation: International Perspectives

June 13, 2017

CoSA Council of State Archivists

Preservica Digital Preservation

Practical Digital Preservation 2017

Welcome!

Sarah Grimm
Wisconsin Historical Society

PDP Briefings

- Protecting and Preserving Long-Term Digital Information — January 24
  - For IT Professionals & practitioners
- The Governance of Long-Term Digital Information — May 23
  - For Senior Managers & Budget Administrators

PDP "Hot Topic" Webinars — Tuesdays 2-3pm Eastern

- Preserving and Protecting Audio-visual Files — April 11
- Preserving Digitized State Government Records — May 9
- Best Practices in Digital Preservation: International Perspective — June 13

PDP Online Workshops - Digital Preservation 101:
'State Archives and Agencies Putting Digital Preservation into Action'

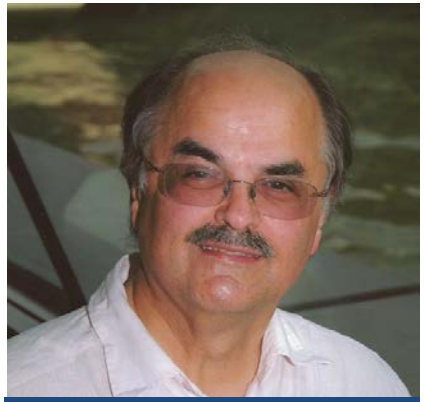
- Part 1: Practical Training in the Key Concepts — February 14
- Part 2: Practical Training in the Key Concepts — February 28
- Part 3: Case Studies — March 14

Sign up today on the CoSA website – PERTTS Portal > Education -Training

CoSA Council of State Archivists

Preservica Digital Preservation

Today's Presenters



David Giaretta
PTAB

Tamara Thornhill
Transport for London

Digital preservation best practices: questions to ask and checks to be made

Will your digital holdings survive?

David Giaretta

david@giaretta.org

www.iso16363.org

13 June 2017

Outline

- Digital Preservation – what is it?
- Threats
- Challenges and Types of Objects to preserve
- OAIS
- Common problems
- International Standards
- Certification: ISO 16363 and the ISO audit process
- Costs and Benefits

Digital Preservation – what is it?

- Keep the bits unchanged?
- Keep converting from one format to another?
- Be able to render (i.e. print/ play/ display etc) the same in the future?
- Provide the bits to the user in 10 seconds?
- Available on a nice website?
- Preserving:
  - Word documents
  - PDF documents

More complicated things:

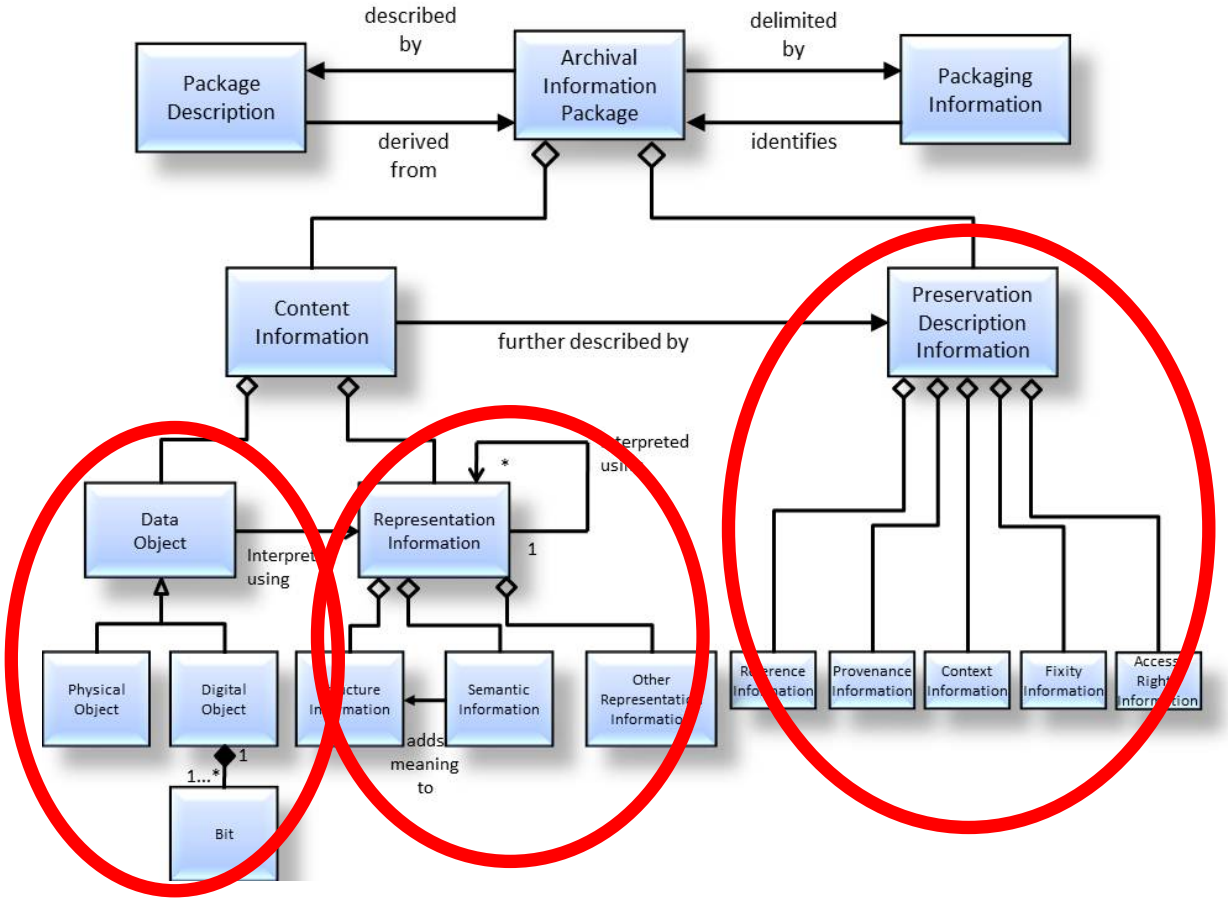
- Modern publications with embedded applications
- MS Word files with embedded links to spreadsheets, databases, Active X apps
- Engineering designs – Computer Aided Design
- Scientific data
  - High Energy Physics
  - Astronomy
  - Biology, Genomics, …
- Finance data
- Websites with embedded applications
- Massively distributed data systems such as (globally) interconnected linked data systems
- ….

Simply being able to printing the numbers or text in the future is not enough!

OAIS (ISO 14721)

- OAIS defines:
  - **Long Term Preservation**: The act of maintaining information, Independently Understandable by a Designated Community, and with evidence supporting its Authenticity, over the Long Term.
- OAIS Conformance:
  - Information model
  - Mandatory responsibilities
- Note the Functional Model provides terminology and concepts but is not part of conformance

OAIS Information Model



OAIS Mandatory Responsibilities

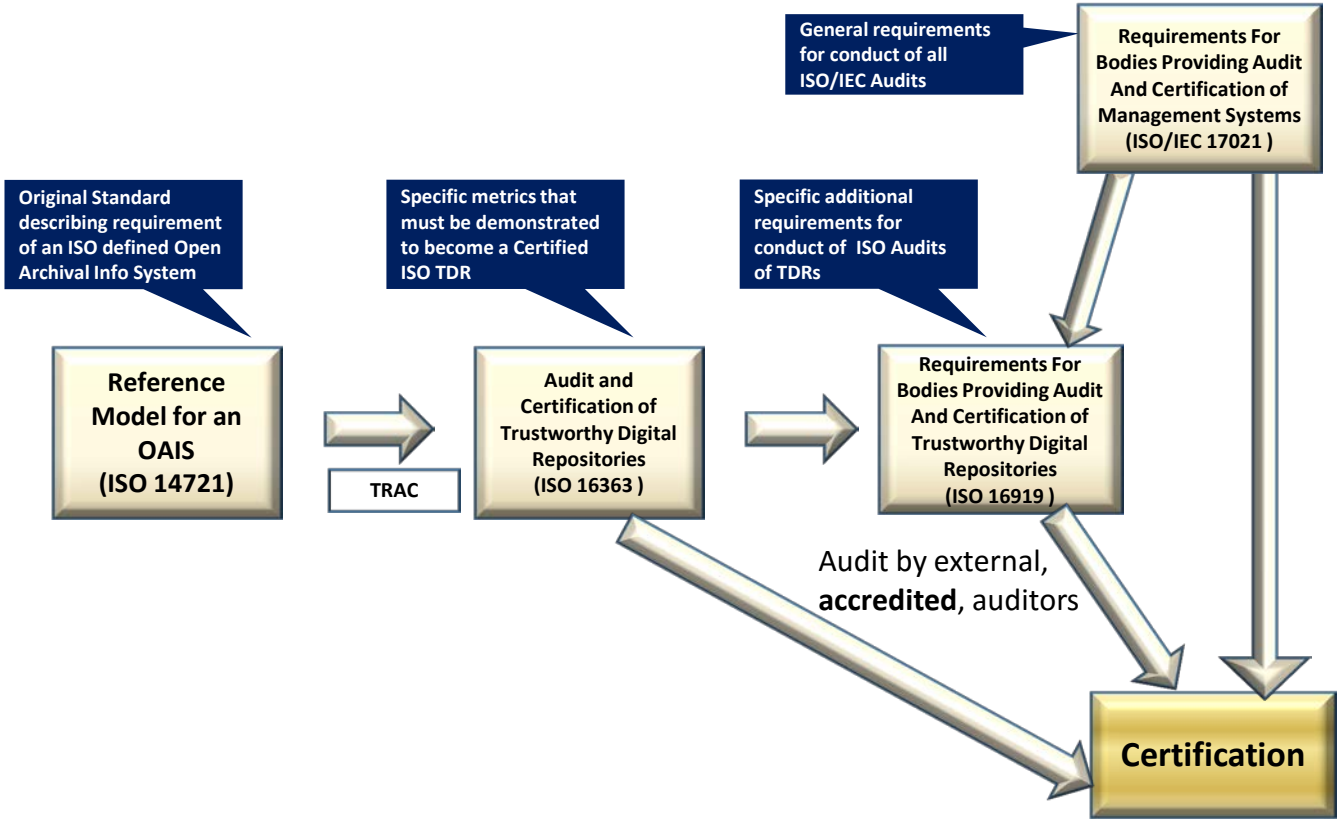
- **Negotiate** for and **accept** appropriate information from information Producers.
- **Obtain sufficient control** of the information provided to the level needed to ensure Long Term Preservation.
- Determine, either by itself or in conjunction with other parties, which communities should become the **Designated Community** and, therefore, should be able to understand the information provided, thereby defining its Knowledge Base.
- **Ensure that the information to be preserved is Independently Understandable to the Designated Community.** In particular, the Designated Community should be able to understand the information without needing special resources such as the assistance of the experts who produced the information.
- Follow documented policies and procedures which ensure that the information is **preserved against all reasonable contingencies, including the demise of the Archive**, ensuring that it is never deleted unless allowed as part of an approved strategy. There should be no ad-hoc deletions.
- Make the preserved information **available to the Designated Community** and enable the information to be disseminated as copies of, or as traceable to, the original submitted Data Objects **with evidence supporting its Authenticity**.

Common problems and their solutions

- Risks to the bits
- Various strategies – checksums, multiple copies, disaster recovery plans etc
- Not enough Representation Information
- Collect Representation Information – remember one does not need everything – just “enough” – need to be clear how much
- Cannot identify all the other things needed – Provenance etc

  i.e. the Archival Info Package (AIP) is inadequate
- Carefully define the AIP – almost certainly have everything needed but remember it does not have to be what the software suppliers say it is

Standards based Repository Audit and Certification (ISO 16363 and ISO 16919)



ISO TDR Certification Framework

- ISO 16363 – the metrics
  - About 120 hierarchically arranged items in 3 main areas
  - Organizational Infrastructure
    - You must be able to explain why you think you have enough staff, enough commitment, enough financial stability
  - Digital Object Management
    - OAIS aspects
  - Infrastructure and Security Risk Management
    - Can you persuade the auditors you have adequately addressed the risks? Are the bits safe?
- A trustworthy digital repository will meet all these requirements adequately
  - Perfection is NOT required
- The auditors will verify that the repository does meet the requirements and if so will issue the ISO 16363 certification

ISO 16363 headlines

- ISO 16363 guides auditors in making a judgement about whether or not a repository can be trusted to play its part in preserving digitally encoded information, following OAIS concepts.
- Fully integrated into the worldwide ISO audit process on which we all depend
  - ISO 16919:2014 specifies requirements for bodies providing ISO 16363 audit and certification – detailed competences that auditors need.
  - 16919 specialises ISO 17021, the fundamental standard for audit and certification
  - Process of continuous improvement
- OAIS and 16363 published by ISO in 2012 so reviews are in progress – see www.iso16363.org , www.oais.info and http://review.oais.info

The importance of ISO accreditation

- Much of our lives and our futures depend on digital information yet the "father of the internet" Vint Cerf worries about the coming "digital dark age".
- We know which systems we can trust for **food, environment, energy** through ISO certification by ISO accredited audit organisations.
- Now there is a way to know who to trust to preserve
- The standard for Trustworthy repositories is ISO 16363, based OAIS
- **The first organisation in the world to receive accreditation for ISO 16363 for Trustworthy Digital Repositories is the Primary Trustworthy Digital Repository Authorisation Body Ltd (PTAB).**
- **As a result, PTAB is authorized to audit digital repositories worldwide and, when appropriate, certify that they are trustworthy.**

ISO 16363 Audit – when is the right time?

- You are ready for an ISO 16363 audit if
  - you think your repository is doing a good job but
  - you know it is not perfect, and need to identify where improvements are needed
    - Or you think it is perfect and need to prove it is!
  - NOTE:
    - The first phase of the audit identifies non-conformances and you will have time to fix these before the audit team visits for the second phase.
    - Then you will have more time to fix the remaining non-conformances before a decision is made on certification.
  - Perfection is not a requirement for certification

Costs and Benefits

- Major (largely one-off) internal cost to prepare for the audit – prepare missing documents and fix the common problems
- Cost of an audit is probably going to be about the same as an ISO 27001 audit i.e. average of 10K/year.
  - For a large repository this would be a few percent of the cost of digital preservation.
- For example initiate Phase 1 to identify the things that need to be fixed for a small registration fee depending on the number of sites, types of data, etc.
- Phase 2 can be arranged when ready and funds are in place
- Benefits
  - An internationally recognised certification
  - Using Independent 3rd party evaluations
  - A process for continuous improvements

BACKUP SLIDES

Digitally encoded information – 1's and 0's

- BITS: 01001110 01001101 01010001 01001101 01010000 01001010 00100000 00100000
- HEX: 4e 4d 51 4d 50 4a 20 20

Example: "ca fe ba be" at start indicates Java class file

- Two IEEE 754 32 bit real numbers: 8.6116461E8 1.35644119E10

Assuming "big-endian"

- Two 32 bit integers 164211241 168379396
- Actually... ....
- ASCII Characters: NMQMPJ

What does this mean?

- ……. Was my flight reference

OAIS (ISO 14721)

- Reference Model for Open Archival Information System (OAIS) provides a very general approach
- OAIS approach to digital preservation:
  - covers all types of digitally encoded information
  - provides a way to **test** whether preservation is successful
  - does not require seeing into the future
  - does require transparency
    - but does not require "open access"
  - does not cover social and organisational aspects
- OAIS does provide a good basis for certification
- Available free from https://public.ccsds.org/Pubs/650x0m2.pdf

Defines useful terminology and important concepts

So many threats to preservation ….

- The bits may be lost – accident or on purpose or floods or earthquakes
- Users may be unable to understand or use the data e.g. the semantics, format or algorithms involved.
- Lack of sustainable hardware, software or support of computer environment may make the information inaccessible.
- Evidence may be lost because the origin and authenticity of the data may be uncertain.
- Access and use restrictions (e.g. Digital Rights Management) may not be respected in the future.
- Loss of ability to identify the location of data.
- The current custodian of the data, whether an organisation or project, may cease to exist at some point in the future.
- The ones we trust to look after the digital holdings may let us down.

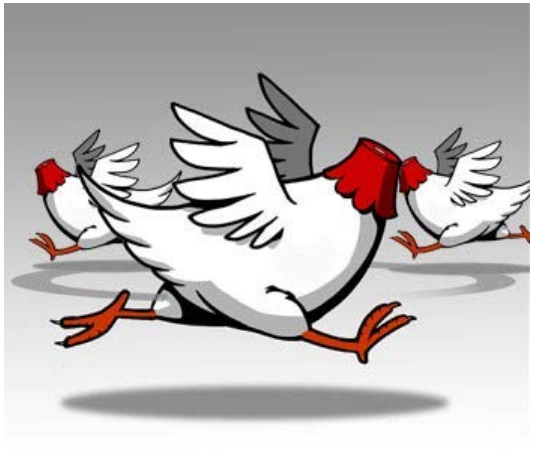
From PARSE.Insight global survey of researchers and data managers

Digital ….. Word and PDF

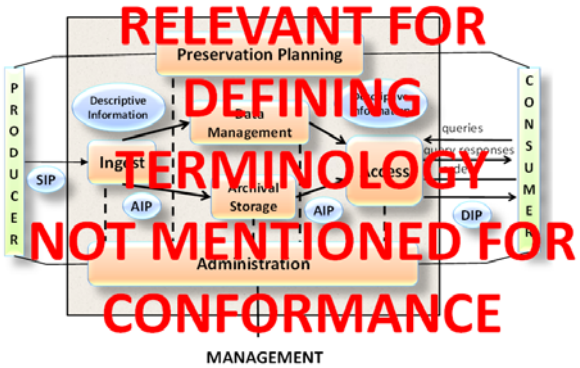
- The obvious analogy with paper documents … (just print and display)
  - WordStar, WordPerfect, MS Word , PDF (various version)
- But……
- Problems include
  - Availability of the software
    - Availability of the Operating systems on which the software runs
      - Availability of the hardware in which the Operating System runs
      - Possible patent issues with x86 emulation
  - Backward compatibility
    - Later version may not be able to read file produced by earlier version of the software
    - Later version produces a different result from earlier version
  - …and what it means e.g. a document written in Chinese or using special abbreviations or technical terms – which the reader does not understand

Misunderstandings

It has been wrongly said that a chicken with its head cut off could be OAIS compliant



... because they thought the Functional Model was the **only** OAIS model



OAIS specifies CONFORMANCE as (section 1.4)

A conforming OAIS Archive implementation **shall support the model of information described in 2.2**.

The OAIS Reference Model does not define or require any particular method of implementation of these concepts.

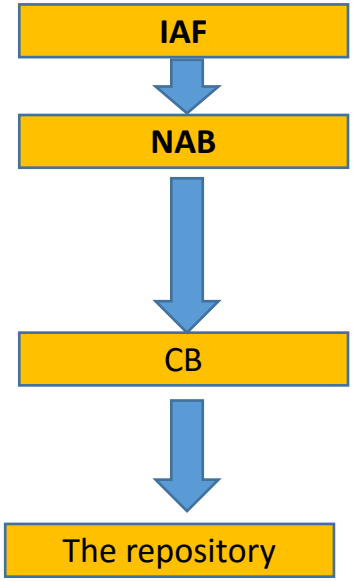
A conforming OAIS Archive **shall fulfill the responsibilities listed in 3.1**. ……..

A separate standard, as noted in 1.5, has been produced on which accreditation and certification processes can be built.

A conformant OAIS Archive may provide to users additional services that are beyond those required of an OAIS.

Governance processes

- A government appoints/sets up a National Accreditation Bodies (NAB) and makes links to the International Accreditation Forum (IAF)
- NAB appoints assessors
  - Knowledgeable about ISO 17021 and audit skills
  - Knowledgeable about ISO 16363/ISO 16919
- Any organisation can apply to be a Certification Body (CB) but there are stringent tests – specified in ISO 17021 – for it to be accredited
  - Each CB is assessed/re-accredited on a regular basis
- A repository – if it is certified – received a certificate which gives:
  - The NAB logo and the number assigned to the CB
  - The CB logo and the number it assigned to the repository certificate
- A user of the repository can check that these are valid by looking at the websites of the NAB and the CB



Applications

- Application form asks for a number of details to allow the CB to judge whether it has the appropriate competences
- E.g. http://www.iso16363.org/iso-certification/application-form/
- Other required public information (http://www.iso16363.org/iso-certification/) :
  - audit processes;
  - processes for granting, refusing, maintaining, renewing, suspending, restoring or withdrawing certification or expanding or reducing the scope of certification;
  - the use of PTAB's name and certification mark or logo;
  - processes for handling requests for information, complaints and appeals;
  - policy on impartiality.

Name of organisation	
Address (or addresses if multiple sites)	
Name of legal representative	
Job title	
Email:	
Telephone:	
Scope of audit required Please provide as much detail as you can, such as: a) physical location(s) b) organizational units c) activities d) processes e) number, with examples, of different types of digital objects being preserved e.g. formats, disciplines, semantics etc f) designated community(ies) g) regulatory and legal environment h) languages	
General description Including activities, human and technical resources, functions and relationship in a larger corporation, if any	
Outsourced processes Any used by the organization that will affect conformity to requirements (if any)	
Consultancy used by the systems to be audited in the last 2 years and by whom	
Other certifications held or being sought	
Restrictions on information which cannot be made available for review by the audit team because they contain confidential or sensitive information.	

Useful Links

- OAIS
  - WEB pages: www.oais.info
  - Site to gather and discuss proposals for OAIS updates in 2017: http://review.oais.info
- ISO 16363:
  - www.iso16363.org
- Integrated GLOSSARY of digital preservation http://www.alliancepermanentaccess.org/index.php/consultancy/dpglossary/
  - SKOS ontology to show relationship between terms from different glossaries
  - OAIS, APARSEN, DPC, ANZ, SNIA, INTERPARES, ISO16363

Tamara Thornhill

Corporate Archivist

Transport for London

CoSA Council of State Archivists

Preservica Digital Preservation

TfL Corporate Archives: Context

- Transport for London responsible for London's public transport network and main roads
- TfL dates from 2000, but earliest incarnation was 1933
- 29,000 employees
- 300+ sites
- c.£11bn budget
- Sponsoring Europe's largest construction project
- 31 million customer journeys/day across 13 modes
- 8.2m contacts in our customer database
- 2012 Olympics Bid centred on transport

Who we are, what we do....

- Preserve and make accessible the Corporate Memory
- Evidence decisions, prove rights and responsibilities, evidence cultural and social impact and engagement
- Serve the business and the public
- Comply with FOI and transparency agenda

TfL Corporate Archives: Records

CoSA
Council of State Archivists

Preservica
Digital Preservation

Digital IRM in TfL: the figures...

- 2,500+ shared network drives
- 120 million+ files on shared network drives
- At least 12 document management systems
- Hundreds of SharePoint sites
- No automated deletion rules in corporate databases/DMS
- Lots of emails that never make it to shared repositories
- Social media channels
- 14,000+ mobile devices deployed, utilising various bespoke apps

Digital Archives in TfL: the figures

- 4 network drives, 300GB born digital (181,000 files)
- Livelink site, 1TB
- USB sticks, Yammer and Diamond Jubilee
- Hard drive web content, 1TB
- Digitised material – 6 hard drives, 2TB
- Mapping Products, 2 TB

Long-Term Solution: AWS for TfL with File Format Migration

- Functionality accessible through a web browser
- Upload content and metadata using an upload tool and pick workflow to accession, ingest, and migrate records
- No third parties (including Amazon) have access to content
- All copies of content held in one geographic region
- Generates multiple copies of content and metadata and replicates this in several data centers to minimize any chance of data loss
- Use fast Amazon S3 for frequently accessed content and lower-cost Amazon Glacier for large infrequently accessed

Valuable Opportunity: Metadata

Field Label	Purpose	Level Captured	ISAD(G)	Dublin Core	PREMIS	Content Descriptive	Technical	Preservation	Needed to know for Access Requirements	Archivists Info	Public Info	Catalogue Description of Phys Record	Technical Management of Digital Record	Administrative Management of Digital Record
Country Code	ID geographical location of record	All	y	n	n	n	n	n	phys	n	y	y	n	n
Repository Code	Tie record to TNA catalogue	All	y	n	n	n	n	n	n	n	y	y	n	n
Level	The position of the record in the hierarchy of the collection	All	y	n	n	y	n	n	n	y	y	y	n	y
Repository	ID physical location of record	All	n	n	n	n	n	y	y	y	y	y	n	y
System ID	Unique identifier used to electronically and therefore intellectually arrange the record	All	n	n	n	n	y	n	n	y	y	y	n	n
Alt Ref	Former references used to identify the record by the Archive	File; Item	n	n	n	n	n	n	n	y	y	y	n	n

CoSA Council of State Archivists

Preservica Digital Preservation

Valuable Opportunity: Awareness

CoSA
Council of State Archivists

Preservica
Digital Preservation

Take-Aways

- How much data do you have? What is your projected growth?
- Organisational Security Concerns, IT Structure
- Access Requirements
- Opportunities to Address Wider Archival Problems
- Potential Wider Use of Software

Next Steps & Resources

CoSA/Preservica Practical Digital Preservation Program

PERTTS Portal

https://www.statearchivists.org/pertts/education-training/cosa-preservica-practical-digital-preservation/

Upcoming Conferences

- CoSA/NAGARA: Boise, ID (July 12-15)
- SAA: Portland, OR July 23 – 29

Watch the next LIVE demo

June 22, 10am US ET, 3pm UK - Register for our live online demonstration of the current Preservica version
Register at www.preservica.com

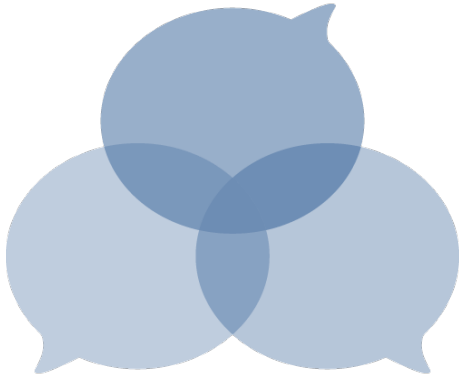
RESOURCES on Preservica.com

From whitepapers, case studies, videos, resources and past webinars - www.preservica.com/resources

CoSA Council of State Archivists

Preservica Digital Preservation

Questions?



CoSA Council of State Archivists

Preservica Digital Preservation

Thank You!

www.preservica.com
info@preservica.com
@preservica
@dPreservation

www.statearchivists.org/

CoSA Council of State Archivists

Preservica Digital Preservation