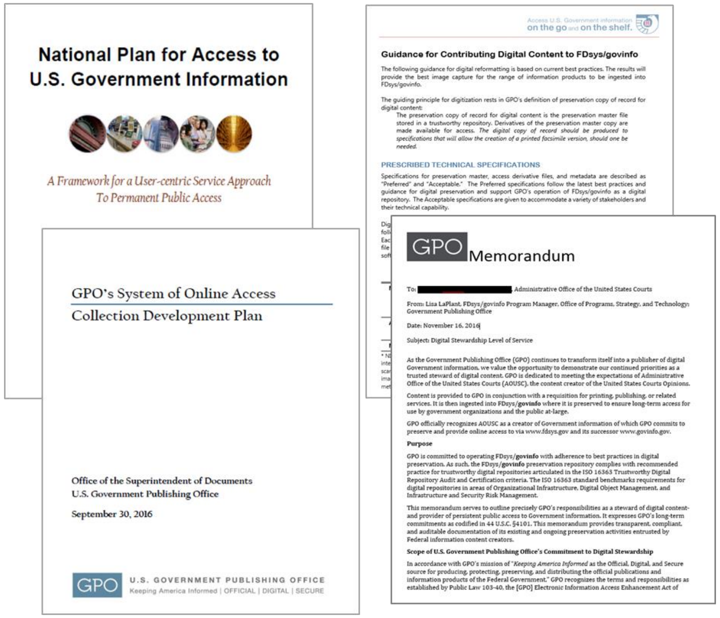
Planning and Managing Storage for Digital Collections

Jessica Tieman
Library Services & Content Management
Government Publishing Office

FDLP Webinar Series
April 2017

What is Digital Preservation?



National Plan for Access to U.S. Government Information

A Framework for a User-centric Service Approach To Permanent Public Access

GPO's System of Online Access
Collection Development Plan

Office of the Superintendent of Documents
U.S. Government Publishing Office

September 30, 2016

GPO U.S. GOVERNMENT PUBLISHING OFFICE
Keeping America Informed | OFFICIAL | DIGITAL | SECURE

Access U.S. Government information
on the go and on the shelf.

Guidance for Contributing Digital Content to FDsys/govinfo

The following guidance for digital reformatting is based on current best practices. The results will provide the best image capture for the range of information products to be ingested into FDsys/govinfo.

The guiding principle for digitization rests in GPO's definition of preservation copy of record for digital content:

> The preservation copy of record for digital content is the preservation master file stored in a trustworthy repository. Derivatives of the preservation master copy are made available for access. *The digital copy of record should be produced to specifications that will allow the creation of a printed facsimile version, should one be needed.*

PRESCRIBED TECHNICAL SPECIFICATIONS

Specifications for preservation master, access derivative files, and metadata are described as "Preferred" and "Acceptable." The Preferred specifications follow the latest best practices and guidance for digital preservation and support GPO's operation of FDsys/govinfo as a digital repository. The Acceptable specifications are given to accommodate a variety of stakeholders and their technical capability.

GPO Memorandum

To: [redacted], Administrative Office of the United States Courts

From: Lisa LaPlant, FDsys/govinfo Program Manager, Office of Programs, Strategy, and Technology; Government Publishing Office

Date: November 16, 2016

Subject: Digital Stewardship Level of Service

As the Government Publishing Office (GPO) continues to transform itself into a publisher of digital Government information, we value the opportunity to demonstrate our continued priorities as a trusted steward of digital content. GPO is dedicated to meeting the expectations of Administrative Office of the United States Courts (AOUSC), the content creator of the United States Courts Opinions.

Content is provided to GPO in conjunction with a requisition for printing, publishing, or related services. It is then ingested into FDsys/**govinfo** where it is preserved to ensure long-term access for use by government organizations and the public at-large.

GPO officially recognizes AOUSC as a creator of Government information of which GPO commits to preserve and provide online access to via www.fdsys.gov and its successor www.govinfo.gov.

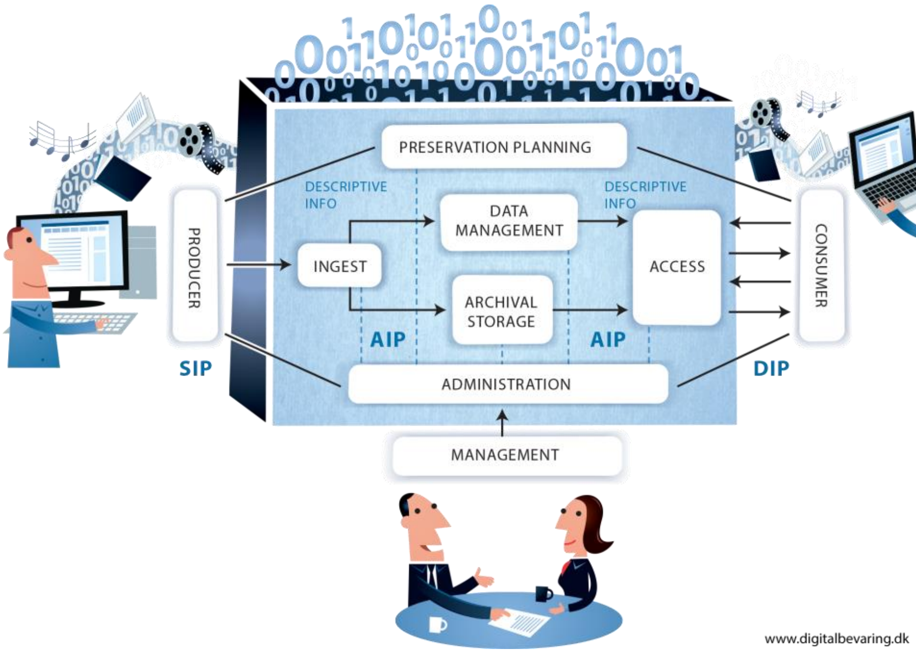
Purpose

GPO is committed to operating FDsys/**govinfo** with adherence to best practices in digital preservation. As such, the FDsys/**govinfo** preservation repository complies with recommended practice for trustworthy digital repositories articulated in the ISO 16363 Trustworthy Digital Repository Audit and Certification criteria. The ISO 16363 standard benchmarks requirements for digital repositories in areas of Organizational Infrastructure, Digital Object Management, and Infrastructure and Security Risk Management.

This memorandum serves to outline precisely GPO's responsibilities as a steward of digital content-and provider of persistent public access to Government information. It expresses GPO's long-term commitments as codified in 44 U.S.C. §4101. This memorandum provides transparent, compliant, and auditable documentation of its existing and ongoing preservation activities entrusted by Federal information content creators.

Scope of U.S. Government Publishing Office's Commitment to Digital Stewardship

In accordance with GPO's mission of "*Keeping America Informed* as the Official, Digital, and Secure source for producing, protecting, preserving, and distributing the official publications and information products of the Federal Government." GPO recognizes the terms and responsibilities as established by Public Law 103-40, the [GPO] Electronic Information Access Enhancement Act of

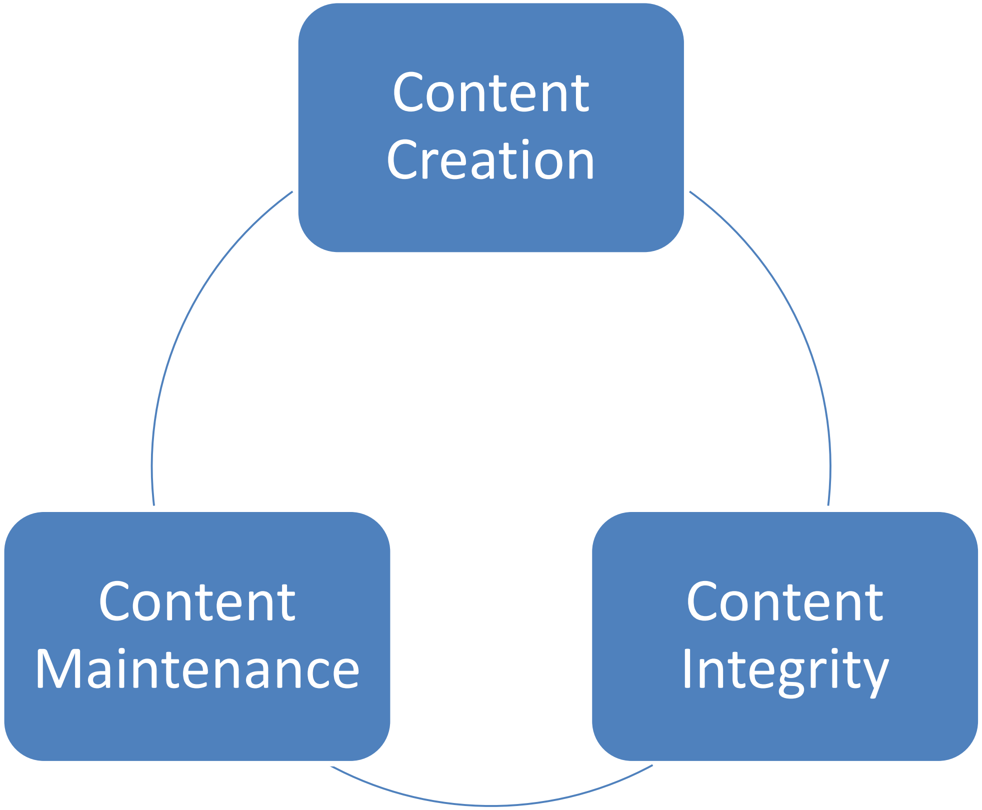


www.digitalbevaring.dk

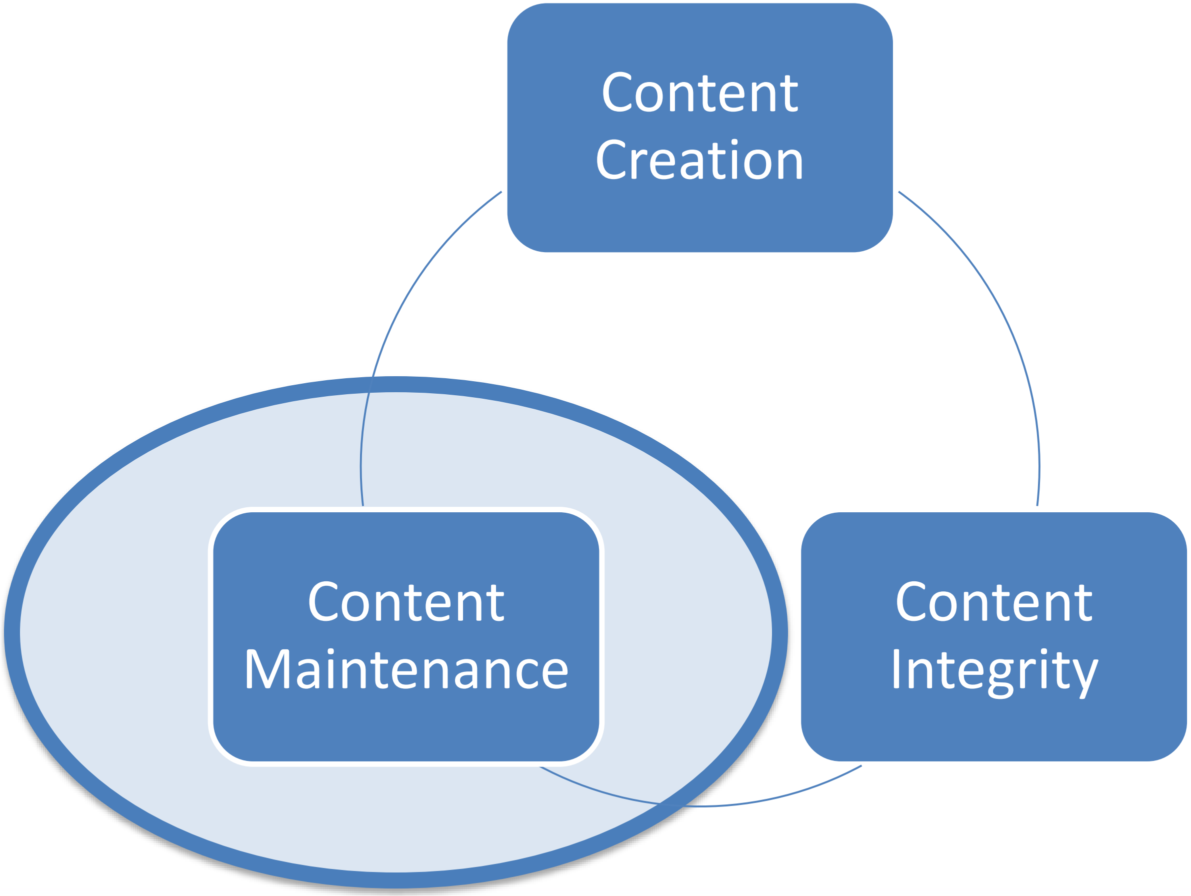
> "Digital preservation combines policies, strategies and actions that ensure access to digital content over time."

-Association for Library and Technical Services

Digital Preservation Strategies



Preservation Storage Requirements



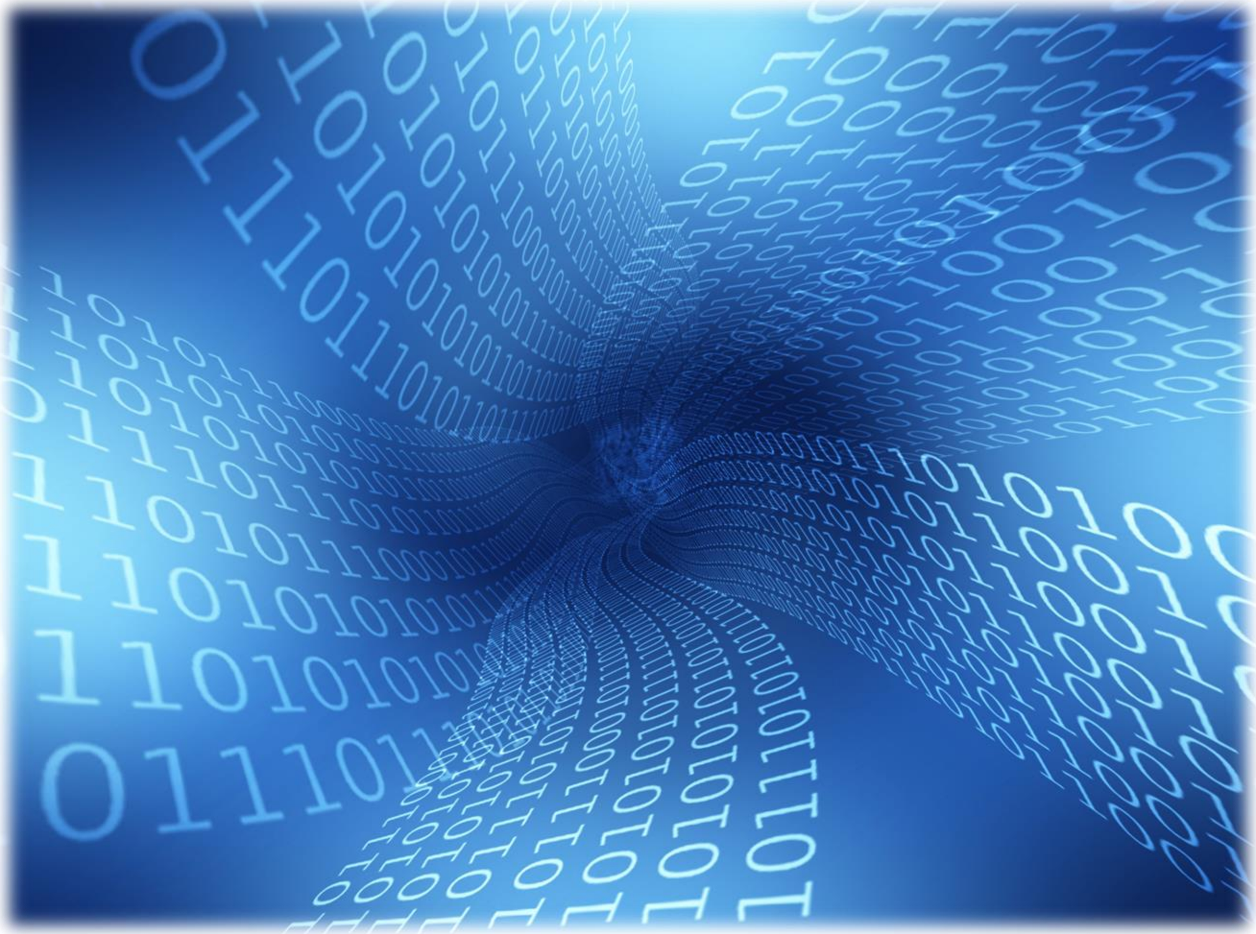
Content Creation

Content Maintenance

Content Integrity

	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, 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 and the 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 storage systems	-At least three copies in a geographic locations with different disaster threats -Have a comprehensive place in a place that will keep files and metadata on currently accessible media or systems
File fixity and Data Integrity	-Check file fixity on ingest and/or create fixity information	-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	-Document access restrictions for content	-Maintain logs of who has performed what actions on files, including deletions and preservation actions	-Perform audits 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	-Encourage use of a limited set of known open formats and codecs	-Inventory of file formats in use	-Monitor file formats obsolescence issues	-Perform format migrations, emulation and similar activities as needed

Preserving the Bits

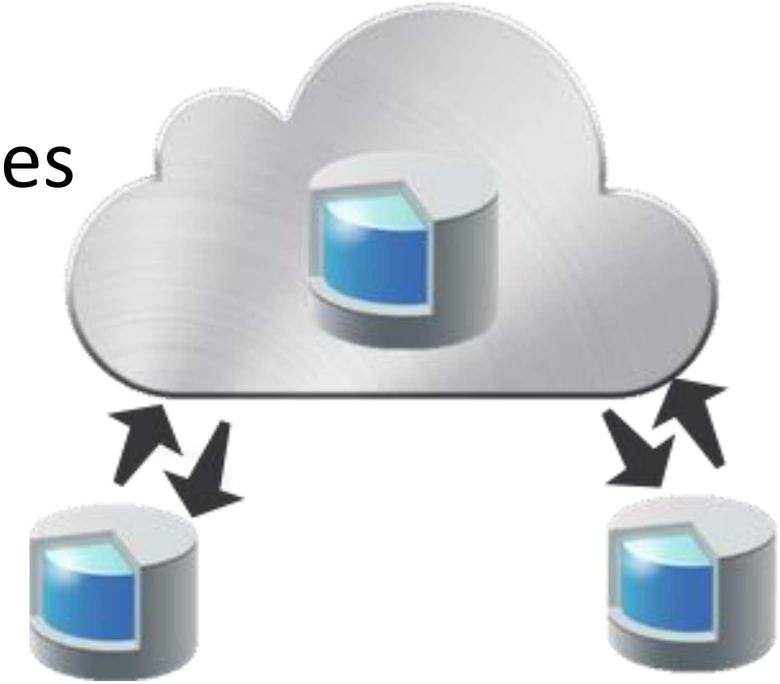


Storage Media

Characteristic	Tape	Optical	Disk	Flash (NAND)	Cloud
Scalability (Capacity)	Medium	Low	Medium	Low	High
Security	High	Low	High	Medium	Medium/ High
Reliability/Performance (Bandwidth)	Low	Low	High	High	High
Cost	Low	Low	High	Low	Low
Lifetime	High	Low	Medium	Low	High
Portability	Medium	High	Low	High	N/A

Storage Media Concepts

- Aerial Density
- Synchronization
- Failure Rate / Reliability
- Mean-time between Failures
- Cyclic Redundancy Check

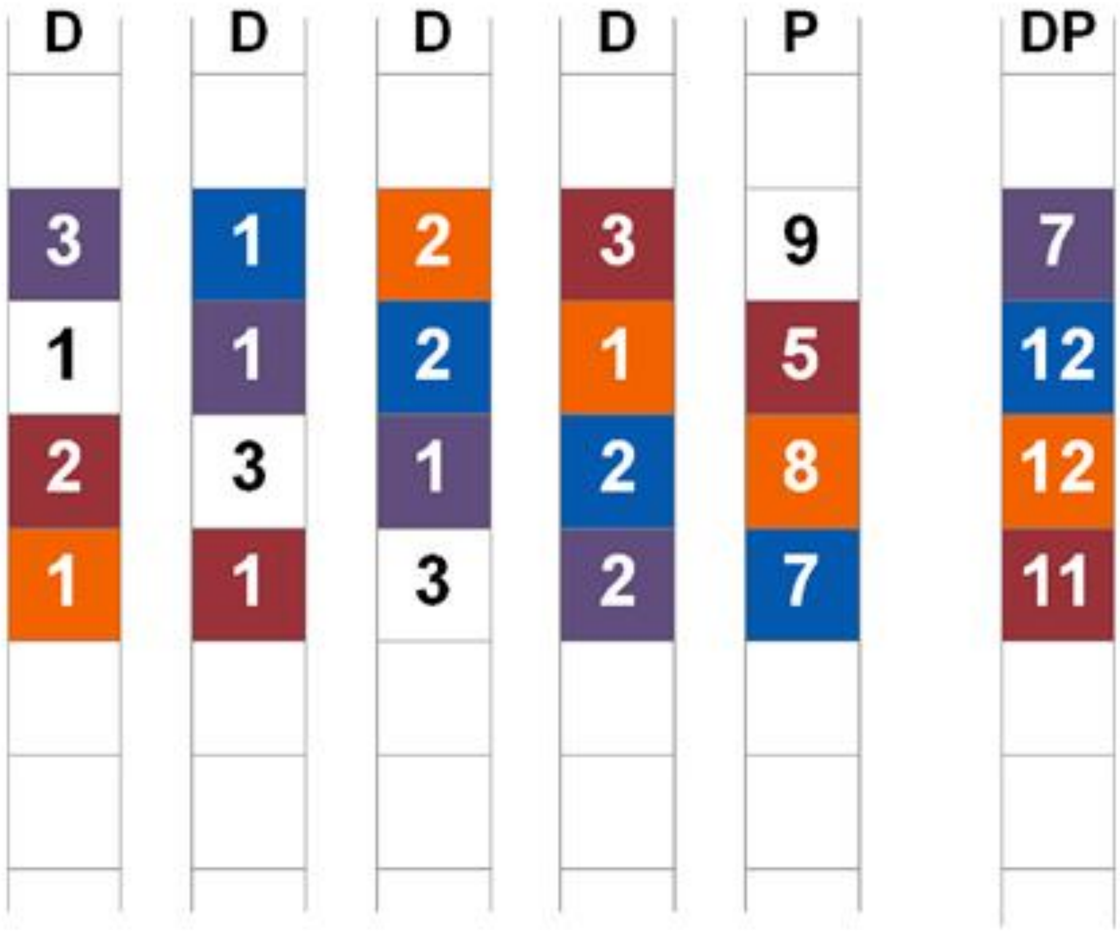


Redundancy & RAID

Redundancy & RAID

Redundancy & RAID

Redundancy & RAID



D	D	D	D	P	DP
3	1	2	3	9	7
1	1	2	1	5	12
2	3	1	2	8	12
1	1	3	2	7	11

How Many Copies?

Fixity

Digital repositories should check for fixity prior to any major migrations, backups, or changes in digital storage infrastructure configurations – for some repositories, this may be monthly, quarterly, or annually.

http://www.digitalpreservation.gov/documents/NDSA-Fixity-Guidance-Report-final100214.pdf

Costs

- Moore’s Law

Computing power doubles every 18 months, that the costs of storage will thus go down, and therefore preservation storage costs will be less of a barrier over time (“Moore’s Law—Overview,” Intel Corporation, http://www.intel.com/research/silicon/mooreslaw.htm).

- Kryder’s Law

The importance of computing power is perhaps not as significant as the increased capability of our technologies to store more and more bits onto smaller and smaller hard drives.

Developing a Risk Registry

Threats to persistence most frequently include:

- Improper/negligent handling or storage (e.g. improper environmental conditions).
- Useful life of storage medium is exceeded (e.g. media obsolescence, mean time to failure exceeded).
- Equipment necessary to read medium is unavailable (e.g., punch card readers, 7 track tape drive).
- Malicious damage to medium and/or bit sequences (e.g. purposeful destruction or theft; computer virus).
- Inadvertent damage to medium and/or bit sequences via hardware, software or operator error.

-Vermaaten, Sally, Brian Lavoie and Priscilla Caplan. "Identifying Threats to Successful Digital Preservation: the SPOT Model for Risk Assessment." *D-Lib Magazine* September/October 2012. <http://www.dlib.org/dlib/september12/vermaaten/09vermaaten.html>.

DRAMBORA: http://www.repositoryaudit.eu/

ISO 16363 Self-Assessment

Self Assessment

How do you feel your technology “sits” in relation to the technology that would be considered “state of the art”?

Does the repository see any potential for licenses to dramatically change or are they any situations where a license has changed and the repository had to react to maintain costs and operations?

What’s your oldest piece of hardware? Why is it needed? How is it used? Can you replace it? How quickly?

How do the relationships you have with your IT department support these plans?

Have you investigated opportunities to create more geographically dispersed copies of your data?

Have incidents of data loss been detected from integrity checks in the past?

Does the repository preserve its audit logs or persist evidence of integrity checks within the PREMIS metadata?

Does the repository have an operating procedure for determining how data corruption occurs?

What types of error detection does your repository use?

What data is covered? Where are the error detection codes kept?

Is there a backup?
Do you have a Disaster Preparedness and Recovery Plan? Who has access to it? Who oversees execution of the plan in the TDR?

When was the last time the Disaster Preparedness and Recovery Plan was reviewed/revised? Can outside entities disrupt this plan if they have a copy of it?

Self Assessment continued…

Who has access to backups?

If an error is detected, what happens?

Who reviews logs/reports?

How would the repository recover from a software or security upgrade which perhaps unpredictably breaks functionality of major software or systems?

Does the repository have the ability to conduct risk analysis before applying an update?

Does the IT department provide notification and communication when upgrades are going to be applied?

Have you performed any checks to ensure that your hardware lifetime estimates are accurate?

Do you generate a regular schedule for refresh?

Do you have multiple copies of your AIPs? Where are they located? Who has access to the copies? Are there different rules for how the copies are treated?

Do keep the same number of copies at the same locations for all collections/types of information?

How do you know all the copies are still the same?

What would happen to data being ingested or synced at the exact moment of a power failure?

How quickly could you respond if a drive failed?

Self assessment continued…

Are all of the RAID disks the same model, purchased at the same time?

How does the repository monitor the threat of silent data corruption?

Do you have a risk register? What is the most significant risks for your repository? How likely is the risk? What are you doing to avoid or deal with that risk? Is that process documented?

Do your repository rely on your larger organization to respond to TDR risks? Does their mitigation plan address the need to protect your AIPs?

Are there any aspects of security risk factors associated with data, systems, personnel, and physical plant which are not covered?

Does IT understand what the repository's responsibility is? Describe the AIP creation/dissemination/storage and backup process. Who can perform each of the steps described?

Which member(s) of your staff are in a position to severely compromise the preservation of your digital holdings?

Where are backups stored?

How many copies of backups are maintained? How would you do a restore of missing Data Object /crashed system/facility? How long would it take?
How often do you test your disaster preparedness and recovery plan(s) and what were the results of the last test?

References

Blue Ribben Task Force on Sustainable Digital Preservation and Access. "Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information." February 2010. Web. 8 March 2017. <http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf>.

Bornholt, James, et al. "A DNA-Baed Archival Storage System." n.d. Web. 8 March 2017. <https://homes.cs.washington.edu/~bornholt/papers/dnastorage-asplos16.pdf>.

Centre, Digital Curation. *DRAMBORA*. 1 February 2008. Web. 8 March 2017. <http://www.repositoryaudit.eu/>.

Digital Preservation Coalition. *Digital Preservation Handbook*. 2017. 8 March 2017. <http://www.dpconline.org/handbook/organisational-activities/storage>.

Fontana, R. and G. Decad. "Storage Media Overview: Projecting Future Trends from 2008-2015 Historic Perspectives." 19 September 2016. Web. 8 March 2017. <https://web.archive.org/web/20161228030617/http://www.digitalpreservation.gov/meetings/DSA2016/Day1/Fontana_LOC%20Talk%20Fontana_Decad_September%202016_09072016_a.pdf>.

Library of Congress. *Digital Preservation Meetings*. 2017. Web. 8 March 2017. <http://www.digitalpreservation.gov/meetings/>.

—. "Preserving Our Digital Heritage: Plan for the National Digital Information Infrastructure and Preservation Program." 2002. 8 March 2017. <http://www.digitalpreservation.gov/documents/ndiipp_plan.pdf>.

National Digital Stewardship Alliance. "Checking Your Digital Content: What is Fixity, and When Should I be Checking It?" 2014. 8 March 2017. <http://www.digitalpreservation.gov/documents/NDSA-Fixity-Guidance-Report-final100214.pdf>.

—. *Fixity Working Group*. n.d. Web. 8 March 2017. <http://ndsa.org/fixity/>.

Panos, Constantopoulos, Doerr Martin and Meropi Petraki. *Reliability Modelling for Long Term Digital Preservation*. n.d. 8 March 2017. <https://pdfs.semanticscholar.org/9b4f/2eb370a81dc79df2fdda1e34ee6d9be1dd6d.pdf>.

Rosenthal and David S.H. "The Future of Storage." 12 May 2016. *DSHR's Blog.* Web. 8 March 2017. <http://blog.dshr.org/2016/05/the-future-of-storage.html>.

Rosenthal, Davd S.H. "Storage Will be a Lot Less Free Than it Used to Be." October 2012. *DSHR's Blog.* 8 March 2017. <http://blog.dshr.org/2012/10/storage-will-be-lot-less-free-than-it.html>.

Rosenthal, David S.H. and Vicky Reich. "Distributed Digital Preservation: Lots of Copies Keep Stuff Safe." n.d. Web. 8 March 2017. <https://lockss.org/locksswiki/files/ReichIndiaFinal.pdf>.

Rosenthal, David S.H., et al. "Requirements for Digital Preservation Systems." *D-Lib Magazine* 2005. Web. 8 March 2017. <http://www.dlib.org/dlib/november05/rosenthal/11rosenthal.html>.

—. *The Economics of Long-Term Digital Storage*. 2012. 8 March 2017. <https://www.lockss.org/locksswp/wp-content/uploads/2012/09/unesco2012.pdf>.

Vermaaten, Sally, Brian Lavoie and Priscilla Caplan. "Identifying Threats to Successful Digital Preservation: the SPOT Model for Risk Assessment." *D-Lib Magazine* September/October 2012. <http://www.dlib.org/dlib/september12/vermaaten/09vermaaten.html>.

Walter, Chip. "Kryder's Law." *Scientific American* 1 August 2005. Web. 8 March 2017. <https://www.scientificamerican.com/article/kryders-law/>.

Contact Information

Jessica Tieman, GPO
jtieman@gpo.gov