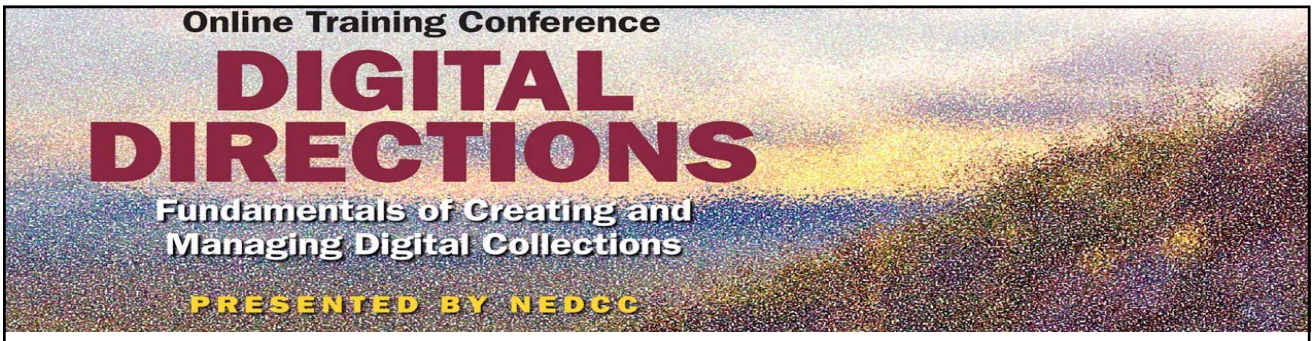


Session 5: Storage 101

Mike Thuman, Digital Transformation Advisor

Digital Enduro, LLC

linkedin.com/in/mike-thuman
@digitalenduro

mike.thuman@gmail.com or mike@digitalenduro.com

www.nedcc.org

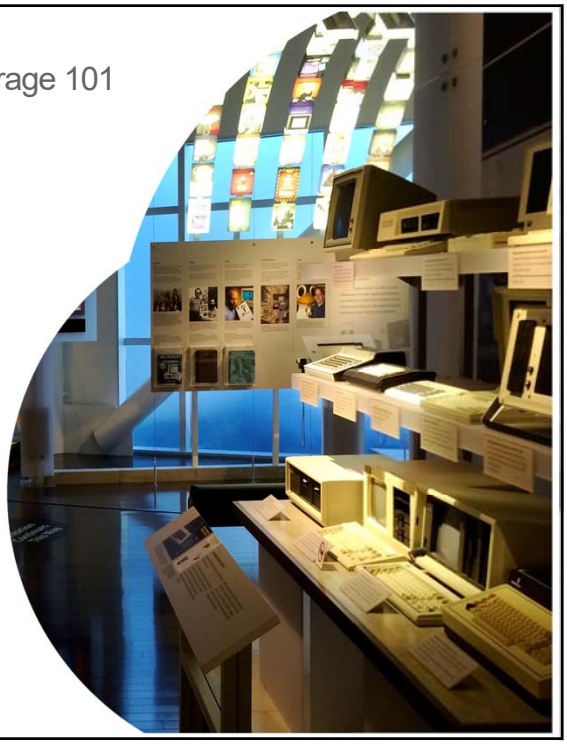
Day 2, 11/16/2021

Storage 101 → Digital Preservation Storage 101

I. Introduction and Storage Notes From Session 1
II. Structured vs Unstructured/Your Inventory
III. Media as Storage
IV. Disk Drives-Bit Rot/Bit Flip Risk
V. NDSA Levels and Storage
VI. Build vs Buy and Deployment Models
VII. On-Prem Storage and Leveraging Cloud
VIII. Digital Preservation Storage
IX. Planning Notes

The Digital Directions conference is geared toward professionals working with digital collections at **archives, libraries, museums, historical organizations, government agencies, corporate archives,** and **other** organizations that steward digital collections.

Digital Directions: comprehensive overview of **digital preservation**



Key Topics Related to Storage from Session 1

Session 1 Digital Preservation: Overview of Concepts, Standards, and Planning (Sam Meister)

*1. A **comprehensive storage strategy** is needed to ensure risks can be prevented and/or mitigated.*

*2. **Good practice** is for a storage strategy to have the following characteristics:*

- multiple copies, geographically separate.
- use different storage technologies, online and offline.
- storage is actively monitored.

*3. **Each institution** will need to evaluate **storage options** in relation to their **specific situation**.*

*4. It is very likely that **some amount of your digital content** will be stored in **"the cloud".***

Unstructured Data and Inventories?

What kind of digital content do you have?

Digital Photos, videos AND their respective hard drives, jump drives, CDs/DVDs.

Documents, images, audio, software files in a wide range of formats.

Artworks and archival records.

Word, PDF, TIFF, jpg, png on shared drives in the company network.

Video on shared drives in a Local Area Network and external hard drives (using compressed and uncompressed formats).

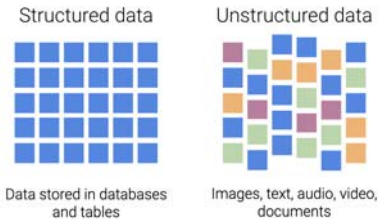
Images and photo-documentation of treatments.

Digital images and videos.

Documents, data, and images. Video in the future.

Documents (conservation reports, technical reports), data, images and videos.

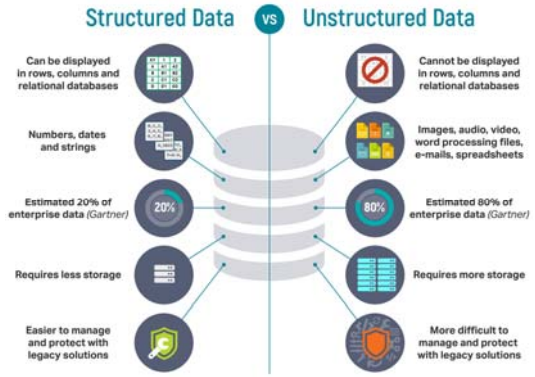
Source: Excerpts from the AIC 2020 Survey



- Rapid growth of video
- Research Data can have its challenges depending on how its packaged
- **Storing hard drives and storage devices is not digital preservation**

https://blog.accern.com/post/the-difference-between-structured-and-unstructured-data

What kind of digital content do you have?



Source: Lawtomated on Medium

The bulk of what you need to store will be unstructured

Unstructured data: more difficult to manage and protect with legacy solutions THUS the need for DIGITAL PRESERVATION and STORAGE

When we get to CLOUD, there are three predominant CLOUD STORAGE technologies:

- BLOCK STORAGE
- FILE STORAGE
- **OBJECT STORAGE**

Do You Have an Inventory of Existing Digital Content?

Sample Basic Inventory

Category: Special Collections - Slides
Title/Description: Circus photographs
Type: Images, digitized
Format: TIFF
Extent: 242 GB, 2250 images
Location: Server (Systems), CDs (Digital Center)
Coverage dates: early 1950s,
Creation date: 2010 - 2012,
Inventoried: by Andrew Huot, November, 2013

DPOE Baseline Modules: Identify, version 2.0, Nov 2011

LIBRARY OF CONGRESS

- Focus on getting a thorough inventory, less focus on the fancy report. Word, Excel and others.
- Make it expandable/scalable.
- Make it a shared doc.
- The TOOLS that you choose will help you interrogate and document file characteristics.
- Anticipate what might be coming your way (e.g., digitization).

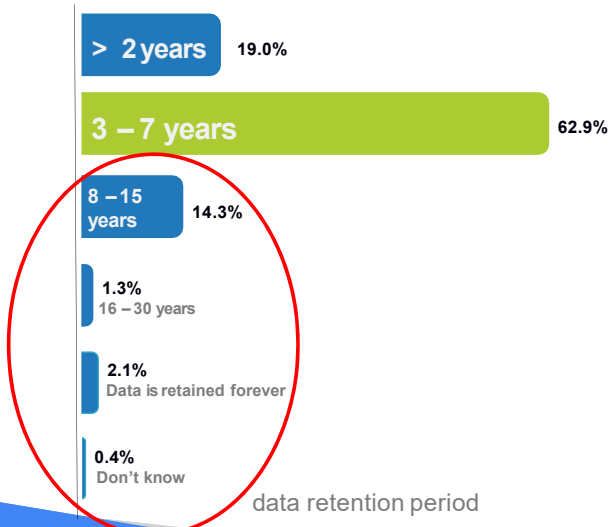
Born Digital versus **Digitized**

Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

Digitized—analog materials converted to digital

You might end up with some of the same formats you will find with Born Digital----PDF, TIFF, JPEG, JPEG2K

How Much Do You Have to Store and Preserve?



Does every retention period require the same type of storage and stewardship?

Storage Terminology: Powers of 10

Kilo	10^3	1 kilobyte (KB) = 1000 bytes
Mega	10^6	1 megabyte (MB) = 1000 KB
Giga	10^9	1 gigabyte (GB) = 1000 MB
Tera	10^{12}	1 terabyte (TB) = 1000 GB
Peta	10^{15}	1 petabyte (PB) = 1000 TB
Exa	10^{18}	1 exabyte (EB) = 1000 PB
Zetta	10^{21}	1 zettabyte (ZB) = 1000 EB
Yotta	10^{24}	1 yotta (YB) = 1000 ZB
Google	10^{100}	

Hitachi Sponsored Object Storage Survey 2019, n=532

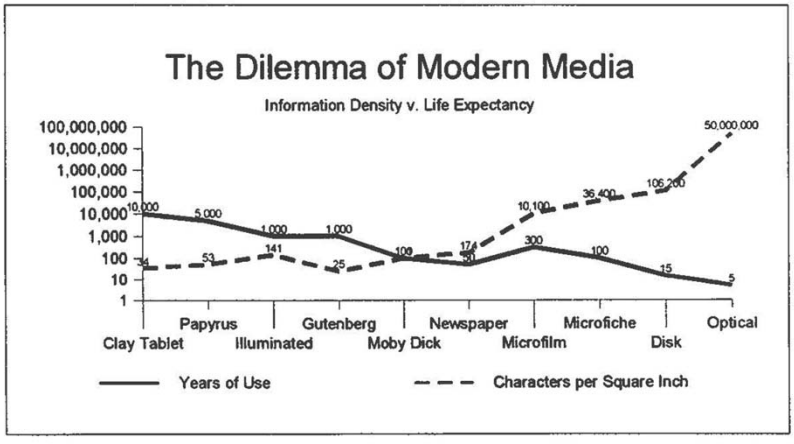
Media as Storage

Using Media for Storage

Any and all media types will fail over time:

Deteriorate
Machine dependency
Mechanical failure
Software dependency
Environmental factors
Lose track of their whereabouts
(not an exhaustive list)

	LTO-5	LTO-6	LTO-7	LTO-7 Type M	LTO-8	LTO-9	LTO-10	LTO-11	LTO-12
Native Capacity	1.5TB	2.5TB	6TB	9TB	12TB	Up to 24TB	Up to 48TB	Up to 96TB	Up to 192TB
Compressed Capacity	3.0TB	6.25TB	15TB	22.5TB	30TB	Up to 60TB	Up to 120TB	Up to 240TB	Up to 480TB
Native Transfer Rate	140 MB/s	160 MB/s	300 MB/s	300 MB/s	360 MB/s	TBD	TBD	TBD	TBD
Compressed Transfer Rate	280 MB/s	400 MB/s	700 MB/s	700 MB/s	750 MB/s	TBD	TBD	TBD	TBD



Our capacity to record information has increased exponentially over time while the longevity of the media used to store the information has decreased equivalently.

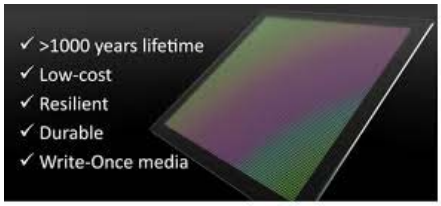
Washington, DC: Council on Library and Information Resources, 1996
https://deepblue.lib.umich.edu/handle/2027.42/150192

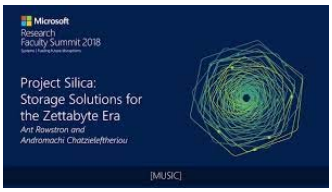
Solving this dilemma—a variety of ideas exist

1. Use more than 1 type of technology (hard drive/LTO/Cloud)
2. Keep it "spinning" (includes the cloud) and check fixity
3. High density media (storage) that lasts longer

Solving this Dilemma—Media That Lasts Longer?

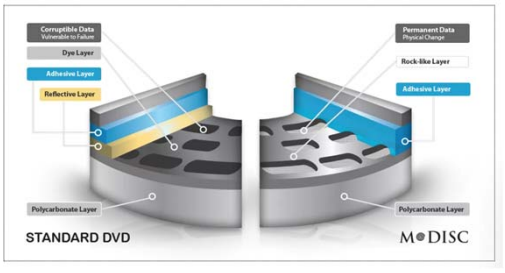
Silica (quartz glass) aims to replace both tape and optical archival discs as the media of choice for large-scale, (very) long duration cold storage. **Microsoft** Research is partnering with film giant Warner Bros., which is directly interested in reducing costs and increasing reliability in its own **cold storage programs.**





Solving this Dilemma—Media That Lasts Longer?

Millenniata M-Disk: The possibility of permanent data archival (2011)
https://www.cnet.com/news/millenniata-m-disk-the-possibility-of-permanent-data-archival/

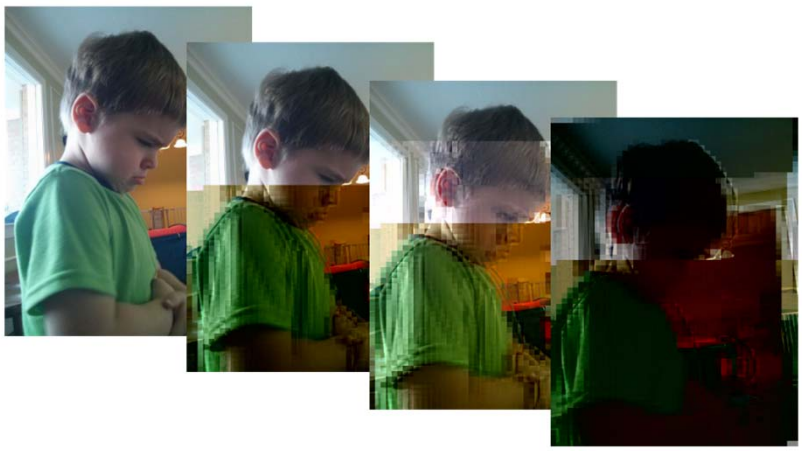


- Media is not likely to be compatible with future computer HW/SW
- Its hard to gain adoption, at scale
- Millenniata, Inc. officially went bankrupt in December 2016.

Storage Risks

Bit Rot and Bit Flips

Individual bits flip or become corrupt, turning a 0 to a 1 or vice versa.



A bit flip or bit rot in the heading of some file formats will prevent you from even opening the file.

From left to right, there is continuing degradation of a JPEG image as bits are flipped. The left image is the original; each image to the right has one bit flipped from the last.

Managing Storage with Fixity

Fixity: the state of being unchanging or permanent.

It's a tool that you can utilize to manage files that are on a storage device AND when moving files from one location to the next.

Checksums and Hashing: Checksums (Cyclic Redundancy Checks, or CRCs) and cryptographic hashes (MD5 and SHA algorithms).

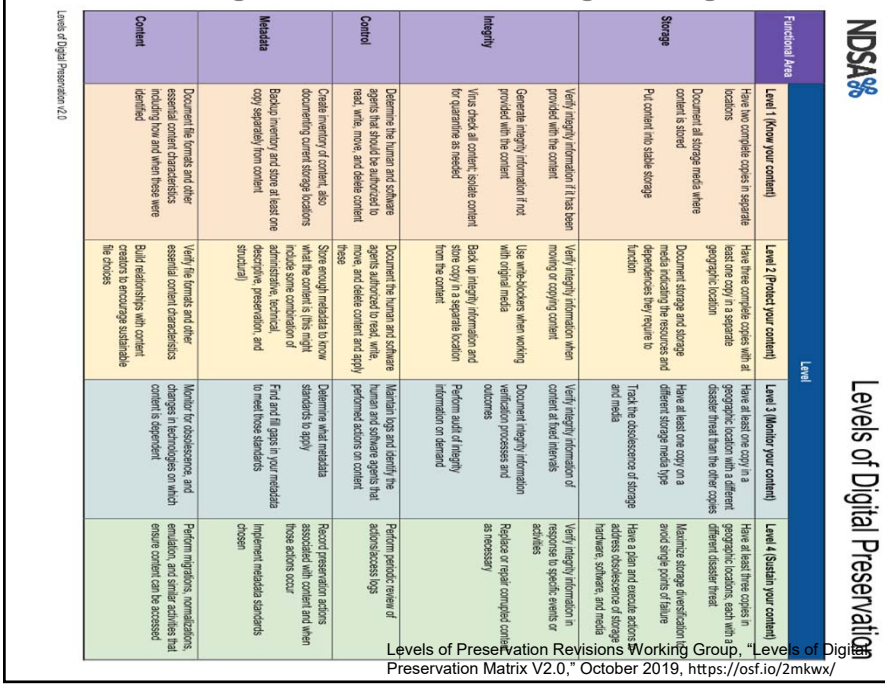
Md5: 6d5b04d33455ac13a2291216e5b552a2
SHA-1: 1a26f9ce33857a5c742877aa8de982968d87f67b
SHA-256: 06a67229b29321064ab6b83cd3fce40bc8079666a1197d324e8f2ce28dd24dff

2017 Fixity Survey Report (how practitioners are using Fixity)
https://ndsa.org/documents/Report_2017NDSAFixitySurvery.pdf

Fixity Definition Level of Effort and Return on Investment Instrument
Expected File Size File size that differs from the expected can be an indicator of problems, for example [by highlighting] zero byte files **Low level of effort and low level detail.** File size is auto-generated technical metadata that can be viewed in Windows Explorer or other common tools.
Expected File Count File count that differs from the expected can be an indicator that files are either added or dropped from the package. **Low level of effort and low level detail.** File count is auto-generated technical metadata that can be viewed in Windows Explorer or other common tools.
CRC Error detection code, typically used during network transfers. **Low level of effort and moderate level of detail.** CRC function values, which are variable but typically 32 or 64 bit, are relatively easy to implement and analyze.
MD5 Cryptographic hash function **Moderate level of effort and high level of detail.** CPU and processing requirements to compute the hash values are low to moderate depending on the size of the file. The output size of this hash value is the lowest of the cryptographic hash values at 128 bits.
SHA1 Cryptographic hash function **Moderate level of effort, high level of detail, and added security assurance.** Due to its higher 160-bit output hash value, SHA-1 requires more relative time to compute for a given number of processing cycles CPU and processing time than MD5.
SHA256 More secure cryptographic hash function **High level of effort, very high level of detail, and added security assurance.** With an output hash value of 256 bits, SHA-256 requires more relative time to compute for a given number of processing cycles CPU and processing time than SHA-1.

Storage Within the Levels of Digital Preservation

There is No Digital Preservation without Digital Storage?



NDSA

Levels of Digital Preservation

Functional Area	Level 1 (Know your content)	Level 2 (Protect your content)	Level 3 (Monitor your content)	Level 4 (Sustain your content)
Storage	Have two complete copies in separate locations	Have three complete copies with at least one copy in a separate geographic location	Have at least one copy in a geographic location with a different disaster threat than the other copies	Have at least three copies in geographic locations, each with a different disaster threat
	Document all storage media where content is stored	Document storage and storage media indicating the resources and dependences they require to function	Have at least one copy on a different storage media type	Maximize storage diversification to avoid single points of failure
	Put content into stable storage		Track the obsolescence of storage and media	Have a plan and execute actions to address obsolescence of storage hardware, software, and media
Integrity	Verify integrity information if it has been provided with the content	Verify integrity information when moving or copying content	Verify integrity information of content at fixed intervals	Verify integrity information in response to specific events or activities
	Generate integrity information if not provided with the content	Use write-blockers when working with original media	Document integrity information verification processes and outcomes	Replace or repair corrupted content as necessary
	Virus check all content; isolate content for quarantine as needed	Back up integrity information and store copy in a separate location from the content	Perform audit of integrity information on demand	
Control	Determine the human and software agents that should be authorized to read, write, move, and delete content	Document the human and software agents authorized to read, write, move, and delete content and apply these	Maintain logs and identify the human and software agents that performed actions on content	Perform periodic review of actions/access logs
Metadata	Create inventory of content, also documenting current storage locations	Store enough metadata to know what the content is (this might include some combination of administrative, technical, descriptive, preservation, and structural)	Determine what metadata standards to apply	Record preservation actions associated with content and when those actions occur
	Backup inventory and store at least one copy separately from content		Find and fill gaps in your metadata to meet those standards	Implement metadata standards chosen
Content	Document file formats and other essential content characteristics including how and when these were identified	Verify file formats and other essential content characteristics	Monitor for obsolescence, and changes in technologies on which content is dependent	Perform migrations, normalizations, emulation, and similar activities that ensure content can be accessed
		Build relationships with content creators to encourage sustainable file choices		

Levels of Digital Preservation v2.0

Levels of Preservation Revisions Working Group, "Levels of Digital Preservation Matrix V2.0," October 2019, https://osf.io/2mkwx/

Digital Storage **alone** has **not** solved our digital preservation requirements.

There are storage related requirements in the top row and throughout the Document

Does every collection require the same type of storage and stewardship?

Build versus Buy

Build versus Buy?

Analyze and Decide Early in your program planning

- Assess your organization
- Assess resources available-technical, financial
- Understand Current Technology in use-compute, network, storage, cloud

Assessment will feed into your decision about **Open Source versus a Vended Solution** and decisions regarding **on-Premise versus SaaS systems and storage.**

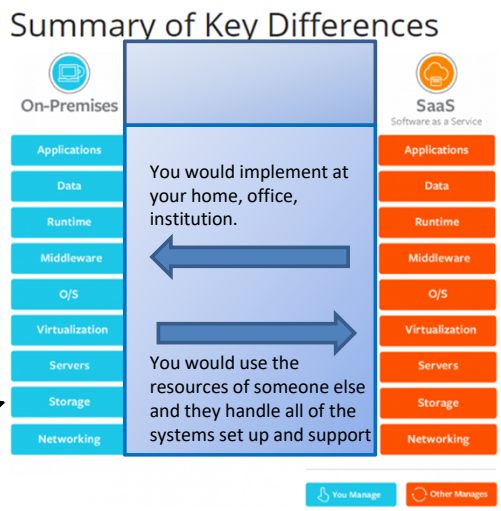
Deployment Models

Where is my system and storage?

Your systems and storage will likely be implemented in one of these models.

Note the TCO of your project spans from Tools/Applications to Servers and Storage and potentially Networking

Storage



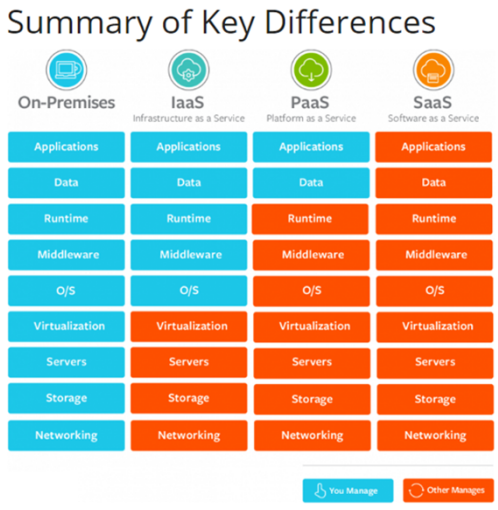
Where is my system and storage?

There are other models that IT may talk about

Storage of course is constant throughout but responsibilities change

How do these approaches compare to Pizza?

Storage



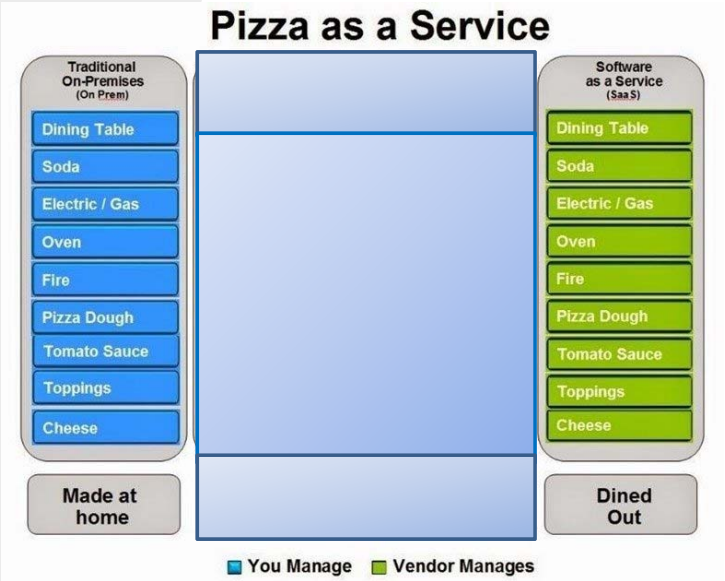
Where is my system and storage?

The Simplified Pizza Making Model

https://dachou.github.io/2018/09/28/cloud-service-models.html

https://www.linkedin.com/pulse/three-monkeys-cloud-iaas-paas-saas-manu-gupta/

Microsoft AZ-900/Denver



Albert Baron, https://www.linkedin.com/pulse/20140730172610-9679881-pizza-as-a-service/

Where is my system and storage?

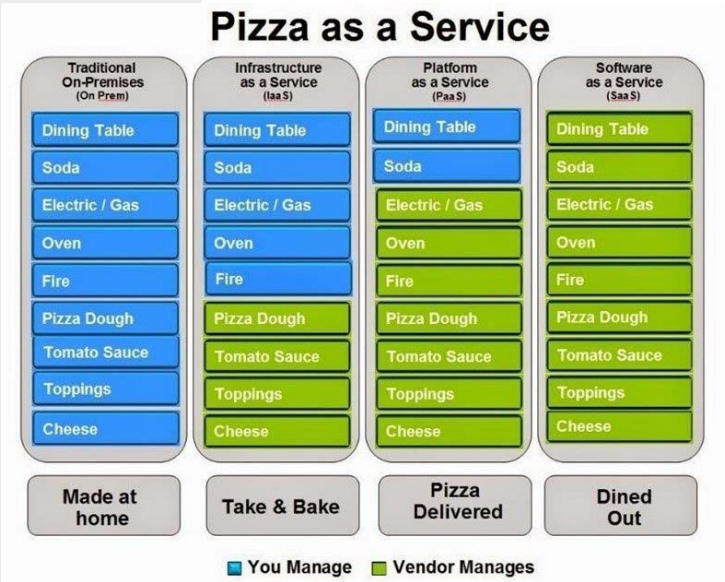
The Simplified Pizza Making Model

Helps explain IaaS and PaaS in case you want to talk tech or work in a hybrid configuration

https://dachou.github.io/2018/09/28/cloud-service-models.html

https://www.linkedin.com/pulse/three-monkeys-cloud-iaas-paas-saas-manu-gupta/

Microsoft AZ-900/Denver



Albert Baron, https://www.linkedin.com/pulse/20140730172610-9679881-pizza-as-a-service/

Sample on-Premise Storage Solutions

On-prem storage systems

Direct Attach Storage-1:1 relationship, computer to storage. For example, Solid State Drives.

Network Attach Storage-a single storage device that serves files over Ethernet. Relatively inexpensive and easy to set up. Drives arranged in a RAID configuration typically. Step up from a home config.

Storage Array Network-more challenging to set up and administer. Makes shared storage available for mission critical and high-performance applications, like video editing. Requires administration by an IT staff

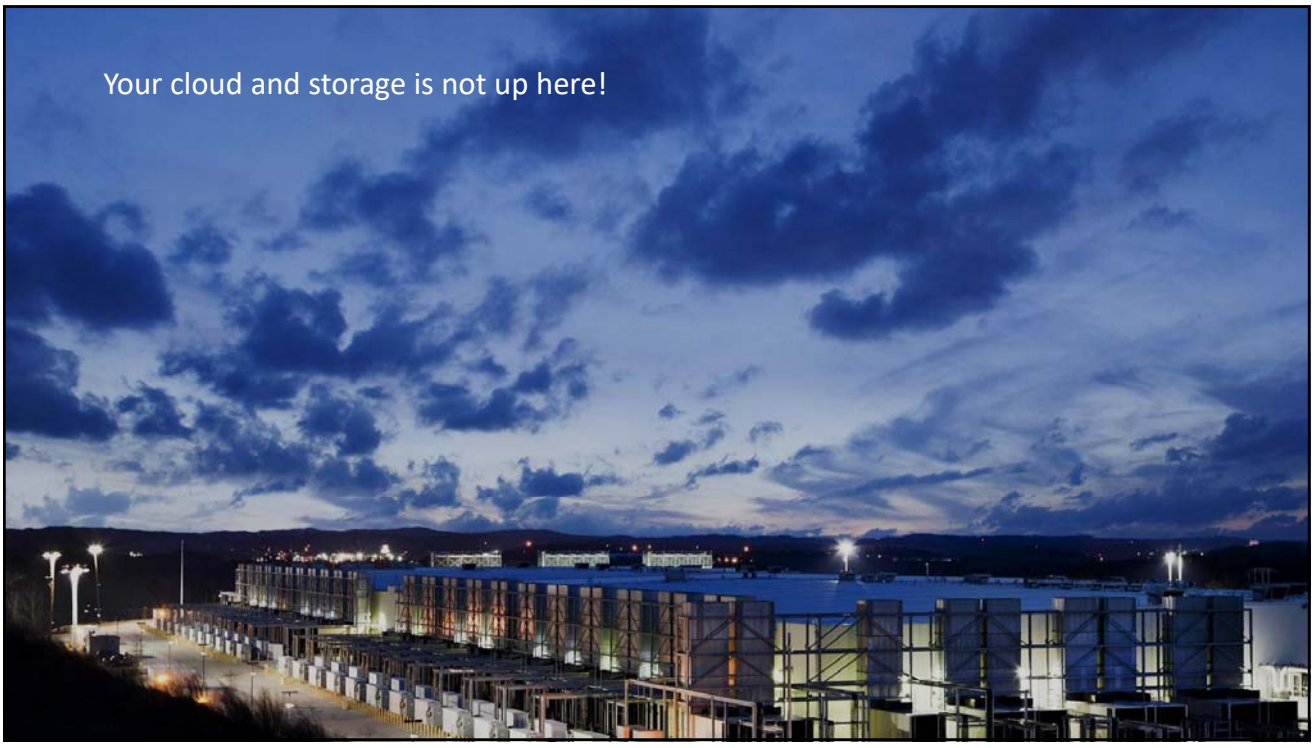
Your role is to know the details of your collections and the use cases.

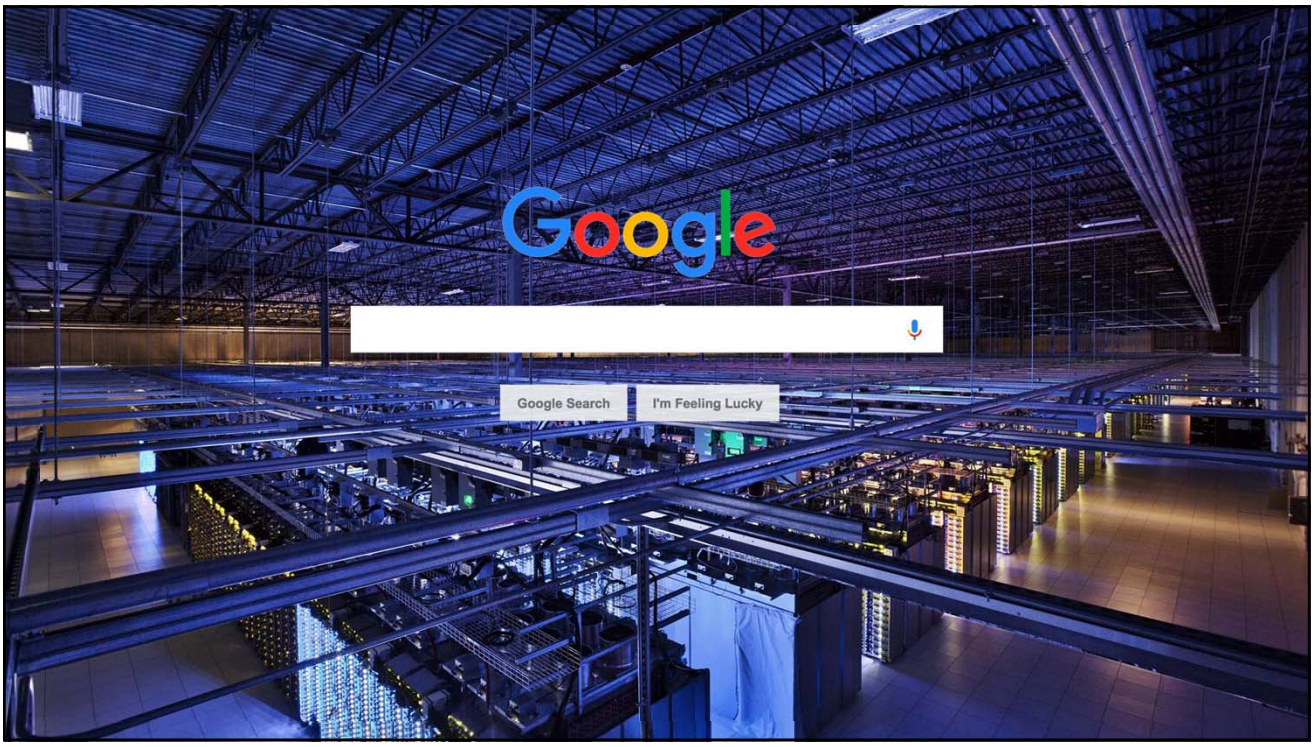
You don't need to become a storage wizard, but familiarity with the options will help in your collaborations with IT and technical staff.

NAS with eight drive bays for 3.5" disk drives.

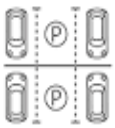
https://www.backblaze.com/blog/whats-the-diff-nas-vs-san/

Leveraging the Cloud

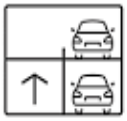




One Slide Summary on Block versus File versus Object Storage



Block storage
'Parking lot' metaphor—data stored in rigidly defined blocks—access by specific "space" location



File storage
'Parking garage' metaphor—data arranged in hierarchical levels—retrace path to access

Object storage
'Valet parking' metaphor—no need to worry about storage details—easy to store and access data

https://www.youtube.com/watch?v=JbPw56Wk3Jk

Object Storage in the Cloud will be our predominant and "go to" technology.

Object storage software optimizes the storage and management of object files, metadata, media, static content and unstructured information.

Compare the best Object Storage solutions currently available-**all 51 of them**

https://sourceforge.net/software/object-storage/?page=1

Storage Classes (AWS Example)

Storage class	Features
S3 Standard	• ≥3 availability zones
S3 Standard - Infrequent Access (IA)	• Retrieval fee associated with objects • Most suitable for infrequently accessed data
S3 Intelligent- Tiering	• Automatically moves objects between tiers based on access patterns • ≥3 availability zones
S3 One Zone-IA	• 1 availability zone • Costs 20% less than S3 Standard-IA
S3 Glacier	• Not available for real-time access • Must restore objects before you can access them • Restoring objects can take 1 minute - 12 hours
S3 Glacier Deep Archive	• Lowest cost storage for long term retention (7-10 years) • ≥3 availability zones • Retrieval time within 12 hours

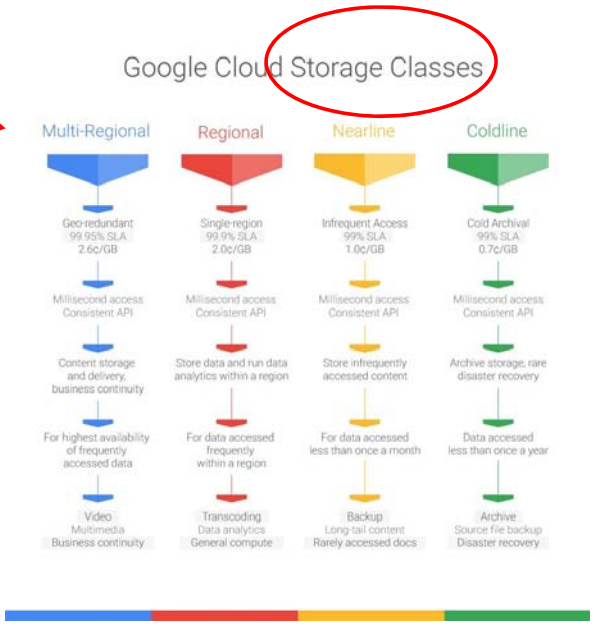
GCS: Object/Blob store

Google Cloud Storage is a scalable object storage service suitable for all kinds of unstructured data

Cloud Storage vs Perst. Disk:

- Scales to exabytes
- Accessible from anywhere; REST interface
- Higher latency than PD
- Write semantics include insert and overwrite file only
- Offers versioning
- Cheaper - put your data here until you need it

Lots of guidelines on picking storage on our site



Everyone Needs to Have a Cloud Strategy

TOP CHALLENGES TO IMPLEMENTING CLOUD

47%	Vendor lock-in
34%	Security concerns
34%	Concerns about where data is stored
31%	Lack of the right skill sets to manage and derive the maximum value from cloud investments
29%	Concerns surrounding integration

Digital Preservation Storage

Community Collaboration (Trending)

Motivation: Need for a guiding document for (digital) preservation storage …

- because all preservation activities rely on storage
- for those who consume or provide preservation storage
- adaptable to any institutional context by combining with local policies, needs, regulations, preferences
- informed by community feedback

https://osf.io/8pj3g/

Digital Preservation Storage Criteria
https://osf.io/sjc6u/

AN OVERVIEW OF THE DIGITAL PRESERVATION STORAGE CRITERIA AND USAGE GUIDE

Eld Zierau
Royal Danish Library
Denmark
elz@kb.dk
0000-0003-3406-3555

Sibyl Schaefer
University of California, San Diego
USA
sschaefer@ucsd.edu
0000-0002-7292-9287

Nancy Y McGovern
Massachusetts Institute of Technology
USA
nancymcg@mit.edu
0000-0002-7733-1516

Andrea Goethals
National Library of New Zealand
New Zealand
Andrea.Goethals@dia.govt.nz
0000-0002-5254-9818

Abstract – The Digital Preservation Storage Criteria (or "Criteria") resulted from a community discussion at iPres 2015 on providing guidance to organizations that either use or provide digital preservation storage. First developed in 2016, they have been refined in iterative versions over the last three years based on feedback gathered at conference sessions and through a survey. The Criteria are intended to help with developing requirements for, or evaluations of, preservation storage solutions; to seed discussions about preservation storage; or to use within digital preservation instructional material. The latest version of the Criteria contains sixty-one criteria grouped into eight categories: content integrity, cost considerations, flexibility, information security, resilience, scalability & performance, support, and transparency.

The key new development since the Criteria was presented at the iPRES 2018 workshop is a usage guide, developed to accompany the Criteria. It includes sections on key topics to consider for preservation storage in addition to the Criteria: risk management, independence, elements in establishing bit safety, and cost considerations. The usage guide will be released publicly for review as one of the next steps in the project, along with developing version 4 of the Criteria and taking steps to further build the community around the Criteria.

Keywords – digital preservation storage, archival storage, criteria, risk management

Conference Topics – Designing and delivering sustainable digital preservation; The cutting edge: technical infrastructure and implementation;

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands,

DOI: 10.1145/nnnnnnn.nnnnnnn

Collaboration: a Necessity, an Opportunity or a Luxury?

I. INTRODUCTION AND BACKGROUND

The Digital Preservation Storage Criteria (or "Criteria") are a result of a collaborative process based within the digital preservation community. This paper provides some context that traces the development and implementation of the Criteria and looks ahead to current and possible future developments. The development of the Criteria has involved iterative cycles of definition and elaboration by a working group, followed by opportunities for community review and feedback, and then finally the integration of community feedback into a series of versions that are publicly available on a project website [1]. Since the advent of computers, storage and processing capacity have framed the development and evolution of preservation strategies; the Criteria are meant to address evolving organizational requirements as digital preservation programs mature, as technological options emerge and evolve, and as opportunities and challenges become clearer.

A. Definition of Digital Preservation Storage

One of the prerequisites for identifying and elaborating the Criteria was developing a working definition of Preservation Storage, absent a shared and authoritative definition within the digital

iPRES 2019

Structure of the criteria

No.	Criteria	Category	Description	Related Criteria & References
1	Integrity checking	Content Integrity	Performs verifiable and/or auditable checks to detect changes or loss in or across copies …	
2	…	…	…	
…				
61	…	…	…	

Some numbers:

- 61 Criteria
- 8 Categories
- 4 Versions
- 1 Usage Guide
- Lots of input!

Complete List: https://osf.io/sjc6u/

Structure of the criteria

No.	Criteria	Category	Description	Related Criteria & References
1	Integrity checking	Content Integrity	Performs verifiable and/or auditable checks to detect changes or loss in or across copies …	
2	…	…	…	
…				
61	…	…	…	

Categories:

- Content Integrity
- Cost Considerations
- Flexibility
- Information Security
- Resilience
- Scalability & Performance
- Support
- Transparency

Complete List: https://osf.io/sjc6u/

Community Collaboration

Motivation: Using the DP Storage Criteria requires context that is grounded in digital preservation principles

- Open Science Framework (OSF): https://osf.io/sjc6u/
- dpstorage Google group: https://groups.google.com/forum/#!forum/dpstorage

Designing Storage Architectures for Digital Collections 201

AUSTRALASIA PRESERVES

DESIGNING STORAGE ARCHITECTURES FOR DIGITAL COLLECTIONS
September 17-18 2018

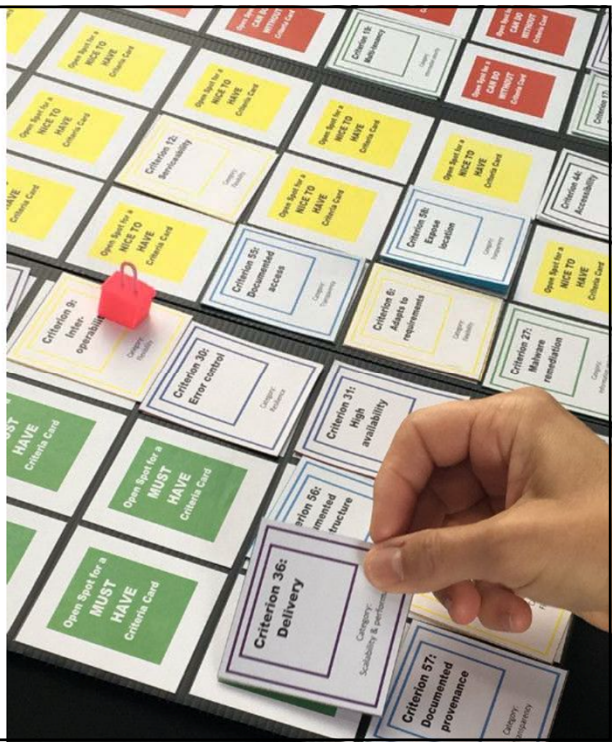
Mike Thuman/Digital Enduro, LLC

Community Collaboration

Motivation: Using the DP Storage Criteria requires context that is grounded in digital preservation principles AND A BOARD GAME

The goal of the game is to learn and think about the most important characteristics of preservation storage for digital material, and to practice explaining your reasons why you think particular characteristics are or are not important in particular contexts. This is not a competitive game per se so there is not a 'winner' of the game. The game is over when all of the 61 Criteria Cards have been placed in one of three sections on the board:

Must Have, Nice To Have, or Can Do Without:



Planning Notes From Today

Get a detailed inventory started/updated

- Your **Use Cases** drive the type of **storage criteria** that will be most important for your collections (**use cases**---open for access?, private, limited to institution only, etc.).
- **Retention periods**-closed, embargoed?

Assess your organizational readiness or current capabilities

- Technical resources, current infrastructure etc.

Build versus Buy

- Follow the path for your selected strategy. On-prem, Cloud, Hybrid etc.

Develop a comprehensive storage strategy

- **NDSA Levels**, multiple copies, geographically separate.
- Use different storage technologies, online, nearline, offline.
- Use storage that is actively monitored.

Its not one size fits all

Have a systems mindset and approach

Planning Notes From Today (2)

Read Scott Prater, How to talk to IT about Digital Preservation:
https://minds.wisconsin.edu/handle/1793/78844
- Covers storage, fixity and other constructs

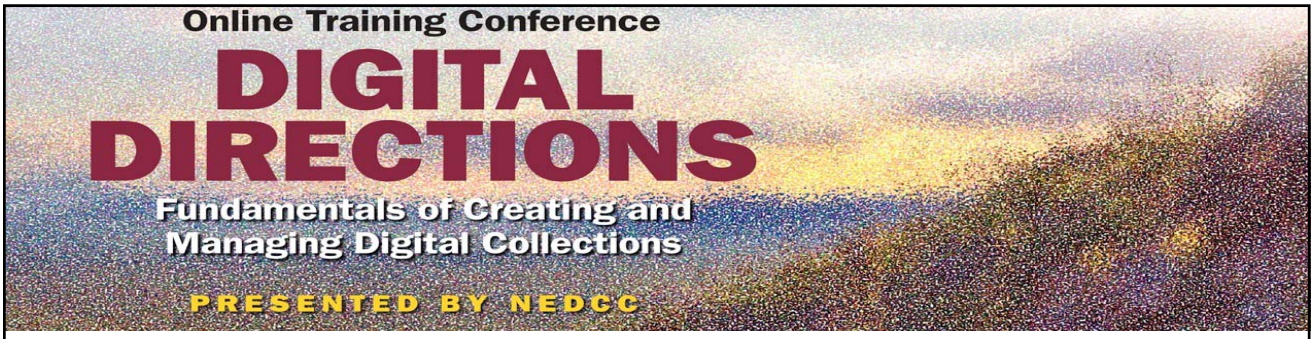
Read the DPC Handbook/Storage Section,
https://www.dpconline.org/handbook/organisational-activities/storage

Have a cloud strategy or technology watch.

Start/update a Digital Preservation policy.

Plan for [rapid] growth of born digital data.

Plan for lifelong learning of formats, tools, transformations, and the potential for emulation.



Thank you!

Please complete the online evaluation form.
A link was sent to your email address.
You can edit your responses until you click submit.

Additional Reference Material

Back Ups versus Digital Preservation

Backups and digital preservation are not the same thing and many IT departments or experts may not appreciate this.

Preservation storage systems require a higher level of geographic redundancy, stronger disaster recovery, longer-term planning, and most importantly active monitoring of data integrity in order to detect unwanted changes such as file corruption or loss.

https://www.dpconline.org/handbook

Two predominant back up destinations- MEDIA or CLOUD

Back up software would be coupled with RESTORE features.

Digital Preservation Coalition Handbook:
https://www.dpconline.org/handbook

NDSA Levels of Preservation:
https://ndsa.org/publications/levels-of-digital-preservation/

Digital Preservation Storage Criteria:
https://osf.io/sjc6u/

NAS vs. SAN
https://www.backblaze.com/blog/whats-the-diff-nas-vs-san/

2017 Fixity Survey Report (how practitioners are using Fixity)
https://ndsa.org/documents/Report_2017NDSAFixitySurvery.pdf