arkivum

Bringing archived data to life

Digital preservation as a way to make content more accessible and usable for the long-term

Matthew Addis

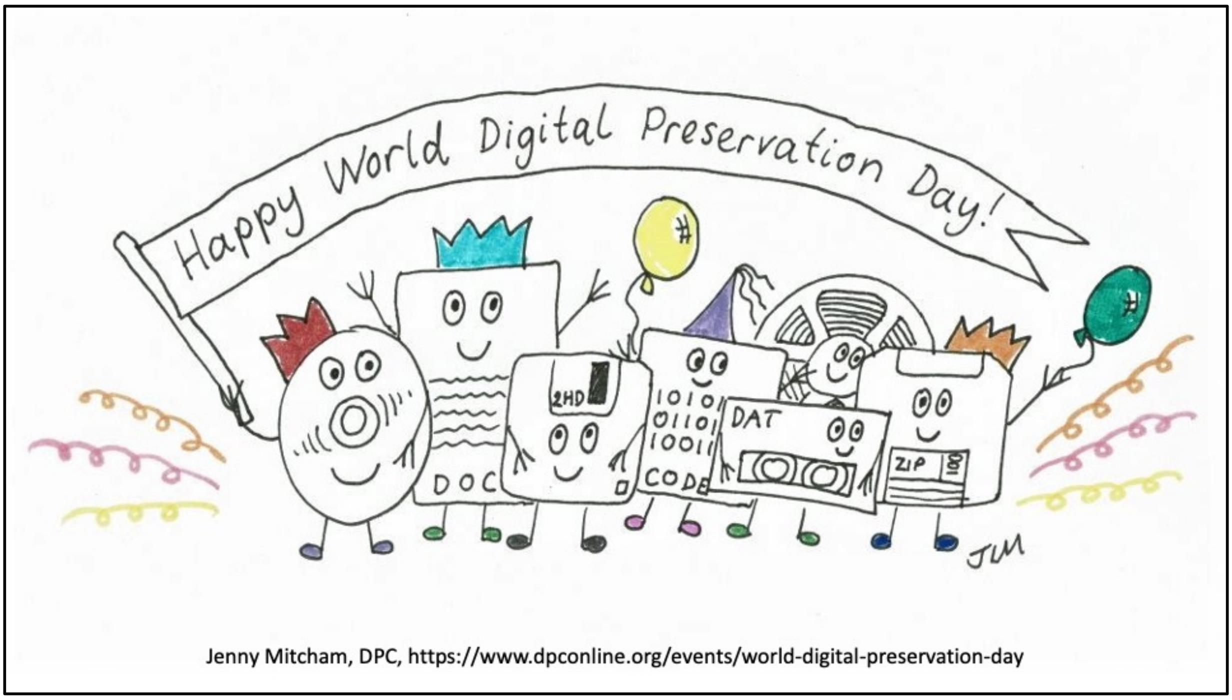
Arkivum

UCT WDPD 2021

Hi, I’m Matthew Addis. I’m the CTO and one of the founders of Arkivum.

If you don’t know Arkivum, we provide a managed software solution for Digital Preservation and online access to digital content.

This talk isn’t about Arkivum and our products and services, this talk is about Digital Preservation.



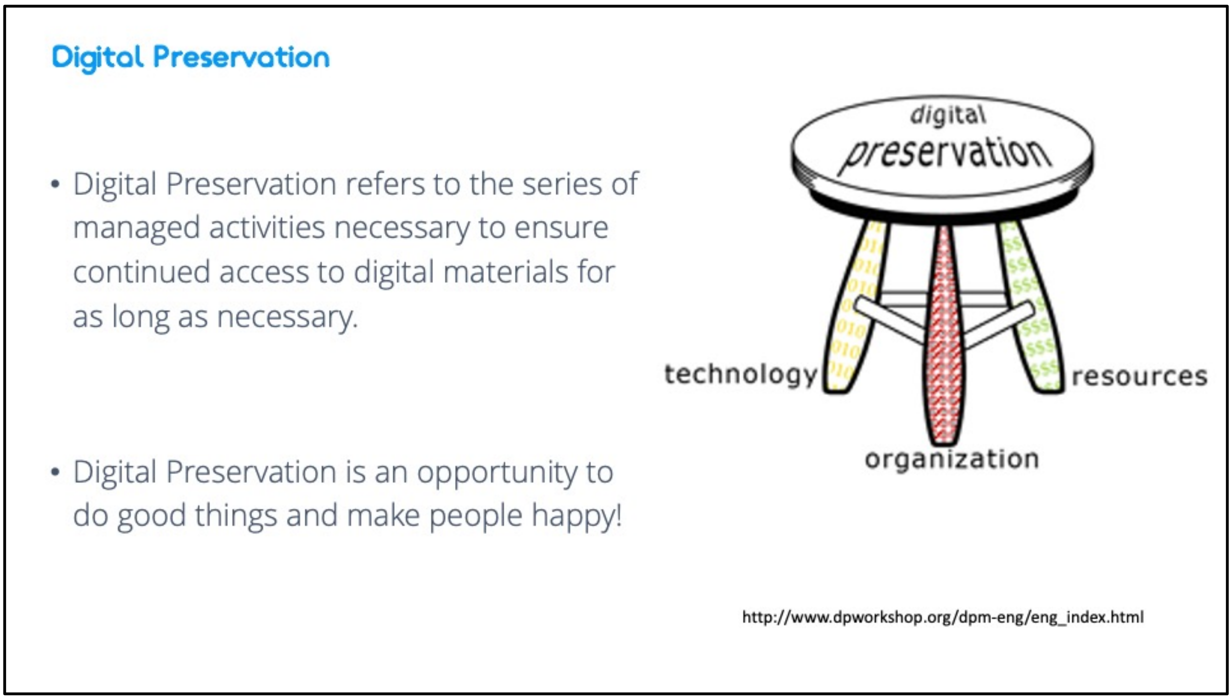
Jenny Mitcham, DPC, https://www.dpconline.org/events/world-digital-preservation-day

Today is of course world digital preservation day!

I'd recommend that you head over to the Digital Preservation Coalition web site and read the all the blog posts that will be published today.

The DPC have lined up an amazing group of people to share their insight and expertise today – and it's always well worth a read.

https://www.dpconline.org/events/world-digital-preservation-day



Digital Preservation

- Digital Preservation refers to the series of managed activities necessary to ensure continued access to digital materials for as long as necessary.
- Digital Preservation is an opportunity to do good things and make people happy!

digital
preservation
technology
resources
organization

http://www.dpworkshop.org/dpm-eng/eng_index.html

It’s common in digital preservation talks to start with a definition of digital preservation and then talk about one or more aspects, such as tools and technology, policies and procedures, or funding and sustainability.

But I want to start with the end-result of digital preservation.

Digital preservation helps ensure that digital content is accessible and usable for people in the future.

And that means digital preservation is opportunity to do good things and to make those people in the future very happy!

http://www.dpworkshop.org/dpm-eng/eng_index.html

GLAM

Findable Accessible Interoperable Reusable

CORE TRUST SEAL

TRUST

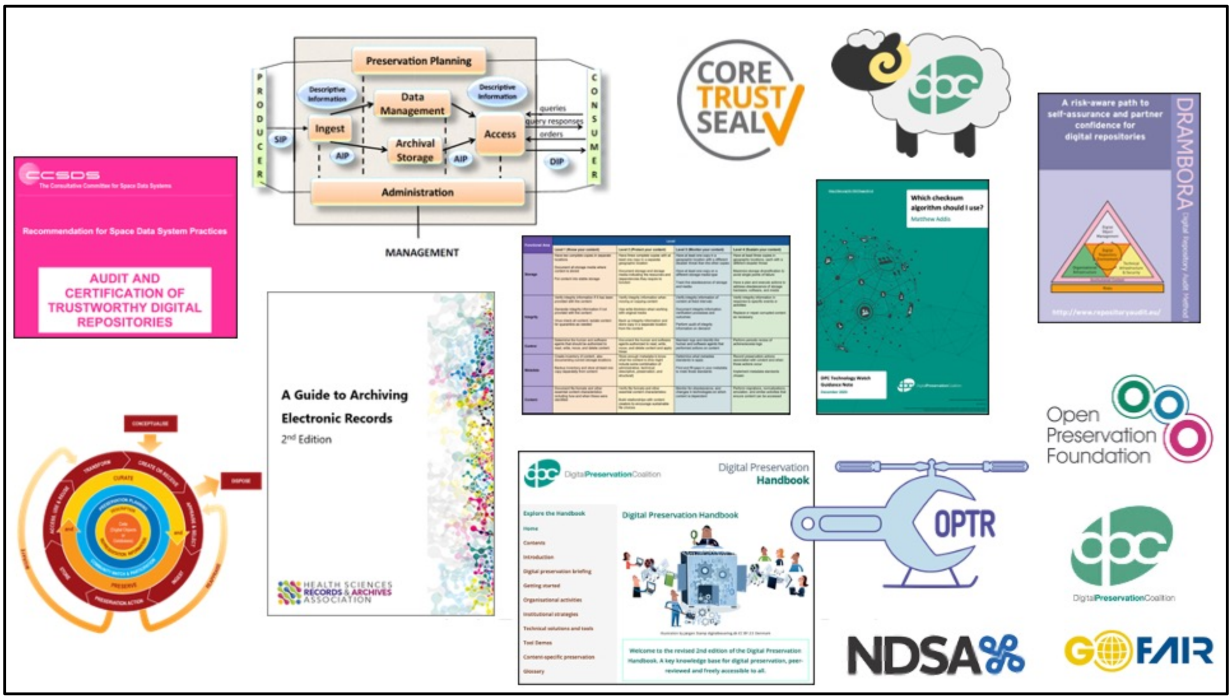
Transparency Responsibility User focus Sustainability Technology

It could be research data and other scholarly outputs.   Digital preservation can help ensure research data remains findable, accessible, interoperable and reusable – otherwise known as FAIR – not just today and tomorrow but next year, next decade and potentially forever.

It could be special collections and archives of historical, cultural and societal value and making sure that digital content in Galleries, Libraries, Archives and Museums – otherwise known as GLAM – is available for future generations to enjoy.

Digital Preservation has a major role to play.

Not just to ensure that the content is accessible in years to come, but to ensure it maintains its authenticity, integrity and provenance.   For example, digital preservation is an essential part of Trusted Digital Repositories.

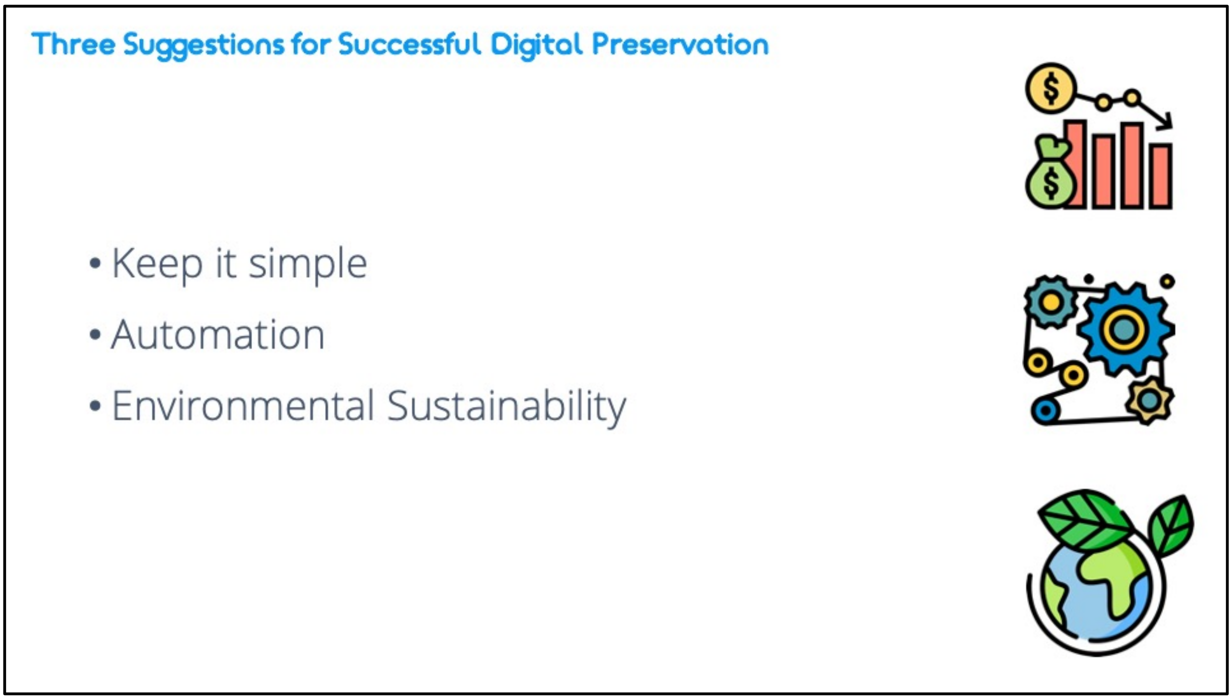

CCSDS
The Consultative Committee for Space Data Systems
Recommendation for Space Data System Practices
AUDIT AND CERTIFICATION OF TRUSTWORTHY DIGITAL REPOSITORIES
Preservation Planning
Descriptive Information
Data Management
Descriptive Information
Ingest
Archival Storage
Access
Administration
queries
query responses
orders
SIP
AIP
AIP
DIP
PRODUCER
CONSUMER
MANAGEMENT
CORE TRUST SEAL
dpc
A risk-aware path to self-assurance and partner confidence for digital repositories
DRAMBORA
http://www.repositoryaudit.eu/
Which checksum algorithm should I use?
Matthew Addis
DPC Technology Watch Guidance Note
A Guide to Archiving Electronic Records
2nd Edition
HEALTH SCIENCES RECORDS & ARCHIVES ASSOCIATION
CONCEPTUALISE
DISPOSE
Open Preservation Foundation
DigitalPreservationCoalition
Digital Preservation Handbook
Explore the Handbook
Home
Contents
Introduction
Digital preservation briefing
Getting started
Organisational activities
Institutional strategies
Technical solutions and tools
Tool Demos
Content-specific preservation
Glossary
Welcome to the revised 2nd edition of the Digital Preservation Handbook. A key knowledge base for digital preservation, peer-reviewed and freely accessible to all.
OPTR
dpc
DigitalPreservationCoalition
NDSA
GO FAIR

Digital Preservation isn’t an easy job though.

Thankfully, there’s a huge amount of experience and good practice from the community to draw upon.

But it’s an ever-evolving field due to new types of content, new techniques and new ways of using that content.

This can make getting started in digital preservation quite hard.



Three Suggestions for Successful Digital Preservation

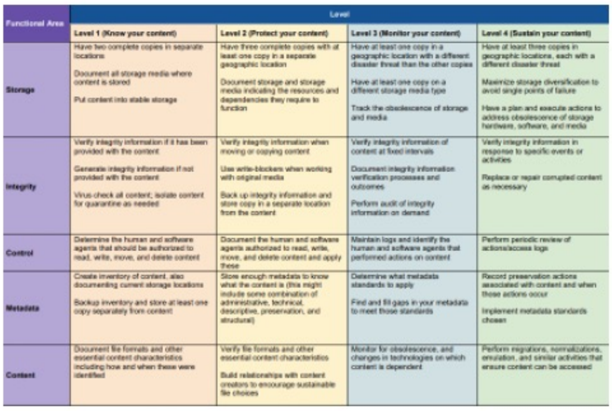
- Keep it simple
- Automation
- Environmental Sustainability

So, I wanted to offer three suggestions for a successful digital preservation journey.

The first thing is to keep things simple – and don't think just because someone else, e.g. a national library or archive appears to be doing a lot more, that you are somehow not doing a good job or not looking after your content properly.

The second thing is to use automation where possible – this helps cope with ever increasing volumes of data, to keep costs down, and to reduce errors.

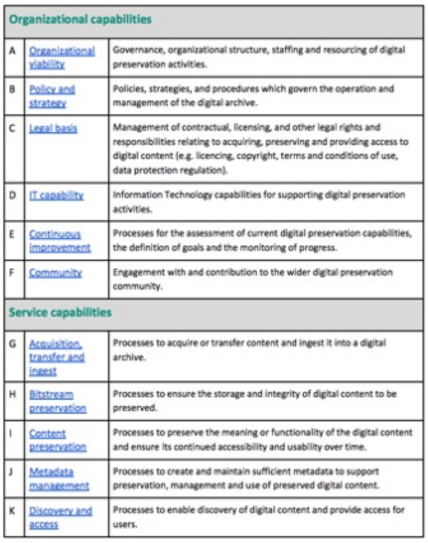
And the third thing is to have an eye on environmental sustainability –this will become an essential part of digital preservation, so thinking about this now will prepare for the future.



Functional Area	Level			
	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 Document all storage media where content is stored Put content into stable storage	Have three complete copies with at least one copy in a separate geographic location Document storage and storage media indicating the resources and dependencies they require to function	Have at least one copy in a geographic location with a different disaster threat than the other copies Have at least one copy on a different storage media type Track the obsolescence of storage and media	Have at least three copies in geographic locations, each with a different disaster threat Maximize storage diversification to avoid single points of failure 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 Generate integrity information if not provided with the content Virus-check all content; isolate content for quarantine as needed	Verify integrity information when moving or copying content Use write-blockers when working with original media Back up integrity information and store copy in a separate location from the content	Verify integrity information of content at fixed intervals Document integrity information verification processes and outcomes Perform audit of integrity information on demand	Verify integrity information in response to specific events or activities Replace or repair corrupted content as necessary
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 Backup inventory and store at least one copy separately from content	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 Find and fill gaps in your metadata to meet those standards	Record preservation actions associated with content and when those actions occur 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 Build relationships with content creators to encourage sustainable file choices	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

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

NDSA



Organizational capabilities		
A	Organizational viability	Governance, organizational structure, staffing and resourcing of digital preservation activities.
B	Policy and strategy	Policies, strategies, and procedures which govern the operation and management of the digital archive.
C	Legal basis	Management of contractual, licensing, and other legal rights and responsibilities relating to acquiring, preserving and providing access to digital content (e.g. licencing, copyright, terms and conditions of use, data protection regulation).
D	IT capability	Information Technology capabilities for supporting digital preservation activities.
E	Continuous improvement	Processes for the assessment of current digital preservation capabilities, the definition of goals and the monitoring of progress.
F	Community	Engagement with and contribution to the wider digital preservation community.
Service capabilities		
G	Acquisition, transfer and ingest	Processes to acquire or transfer content and ingest it into a digital archive.
H	Bitstream preservation	Processes to ensure the storage and integrity of digital content to be preserved.
I	Content preservation	Processes to preserve the meaning or functionality of the digital content and ensure its continued accessibility and usability over time.
J	Metadata management	Processes to create and maintain sufficient metadata to support preservation, management and use of preserved digital content.
K	Discovery and access	Processes to enable discovery of digital content and provide access for users.

https://www.dpconline.org/our-work/dpc-ram

Starting with keep it simple, consider using a maturity model, such as the NDSA levels or preservation of DPC Rapid Assessment Model.

These are simple and practical guidelines to digital preservation.

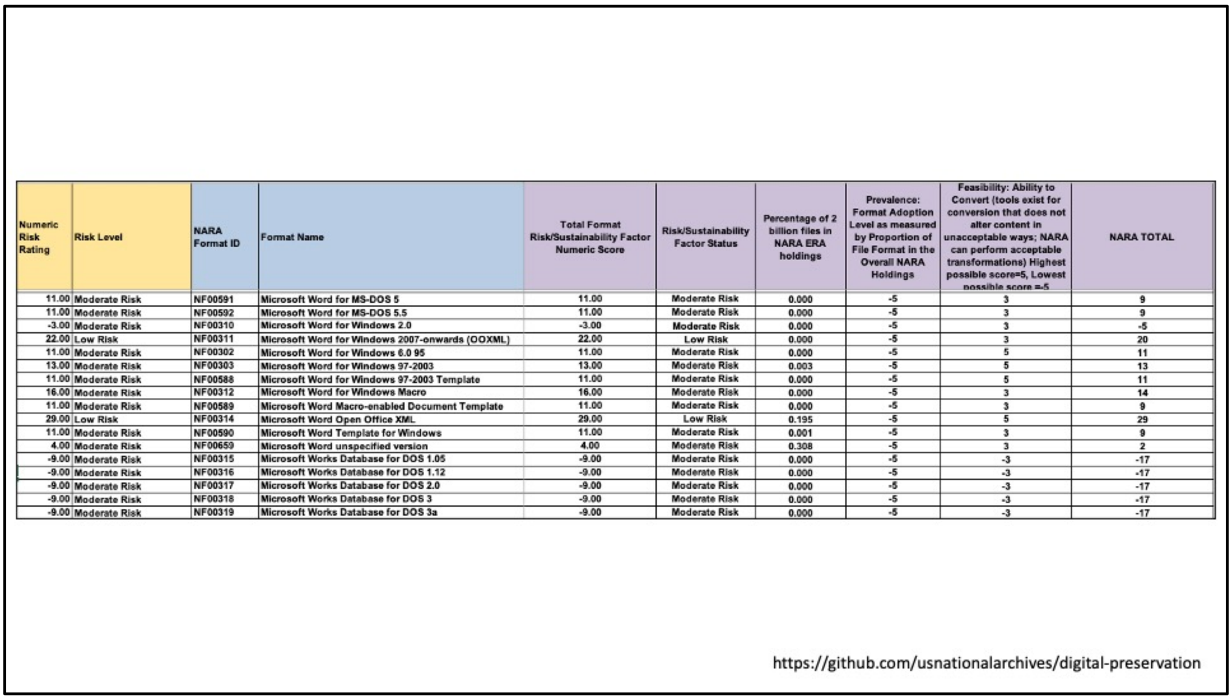
The important thing is that they have a progression of levels.

It is perfectly legitimate to start at level 1 and work upwards as and when you need to.

This helps get the basics of digital preservation in place first without worrying about everything that you might possibly need or want to do in the future.

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

https://www.dpconline.org/our-work/dpc-ram



Numeric Risk Rating	Risk Level	NARA Format ID	Format Name	Total Format Risk/Sustainability Factor Numeric Score	Risk/Sustainability Factor Status	Percentage of 2 billion files in NARA ERA holdings	Prevalence: Format Adoption Level as measured by Proportion of File Format in the Overall NARA Holdings	Feasibility: Ability to Convert (tools exist for conversion that does not alter content in unacceptable ways; NARA can perform acceptable transformations) Highest possible score=5, Lowest possible score =-5	NARA TOTAL
11.00	Moderate Risk	NF00591	Microsoft Word for MS-DOS 5	11.00	Moderate Risk	0.000	-5	3	9
11.00	Moderate Risk	NF00592	Microsoft Word for MS-DOS 5.5	11.00	Moderate Risk	0.000	-5	3	9
-3.00	Moderate Risk	NF00310	Microsoft Word for Windows 2.0	-3.00	Moderate Risk	0.000	-5	3	-5
22.00	Low Risk	NF00311	Microsoft Word for Windows 2007-onwards (OOXML)	22.00	Low Risk	0.000	-5	3	20
11.00	Moderate Risk	NF00302	Microsoft Word for Windows 6.0 95	11.00	Moderate Risk	0.000	-5	5	11
13.00	Moderate Risk	NF00303	Microsoft Word for Windows 97-2003	13.00	Moderate Risk	0.003	-5	5	13
11.00	Moderate Risk	NF00588	Microsoft Word for Windows 97-2003 Template	11.00	Moderate Risk	0.000	-5	5	11
16.00	Moderate Risk	NF00312	Microsoft Word for Windows Macro	16.00	Moderate Risk	0.000	-5	3	14
11.00	Moderate Risk	NF00589	Microsoft Word Macro-enabled Document Template	11.00	Moderate Risk	0.000	-5	3	9
29.00	Low Risk	NF00314	Microsoft Word Open Office XML	29.00	Low Risk	0.195	-5	5	29
11.00	Moderate Risk	NF00590	Microsoft Word Template for Windows	11.00	Moderate Risk	0.001	-5	3	9
4.00	Moderate Risk	NF00659	Microsoft Word unspecified version	4.00	Moderate Risk	0.308	-5	3	2
-9.00	Moderate Risk	NF00315	Microsoft Works Database for DOS 1.05	-9.00	Moderate Risk	0.000	-5	-3	-17
-9.00	Moderate Risk	NF00316	Microsoft Works Database for DOS 1.12	-9.00	Moderate Risk	0.000	-5	-3	-17
-9.00	Moderate Risk	NF00317	Microsoft Works Database for DOS 2.0	-9.00	Moderate Risk	0.000	-5	-3	-17
-9.00	Moderate Risk	NF00318	Microsoft Works Database for DOS 3	-9.00	Moderate Risk	0.000	-5	-3	-17
-9.00	Moderate Risk	NF00319	Microsoft Works Database for DOS 3a	-9.00	Moderate Risk	0.000	-5	-3	-17

https://github.com/usnationalarchives/digital-preservation

And when you start looking at the details of digital preservation, e.g. what to do with different file formats that you might have, then again try to keep it simple.

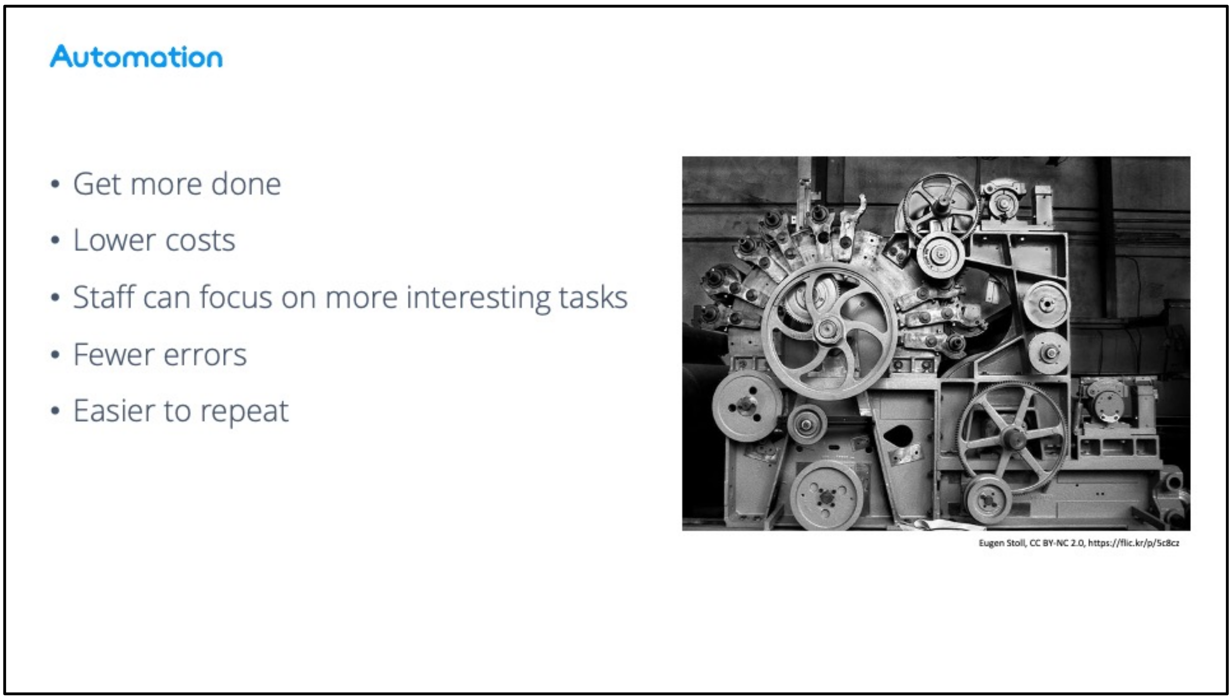
People worry about things like technical obsolescence of file formats, but the reality is not always as bad as some people think or have predicted.

This is an example of NARA's risk assessment of just some of the Microsoft file formats. Even old Word formats aren't that much of a risk from a preservation perspective.

NARA has done assessments of a huge range of formats covering images, audio, video, documents, email and more – which makes it a great starting point for your own risk assessment.

The point is that resources like this can help you avoid investing a lot of time, effort and money into file format conversions that could be unnecessary.

https://github.com/usnationalarchives/digital-preservation



Automation

- Get more done
- Lower costs
- Staff can focus on more interesting tasks
- Fewer errors
- Easier to repeat

Eugen Stoll, CC BY-NC 2.0, https://flic.kr/p/5c8cz

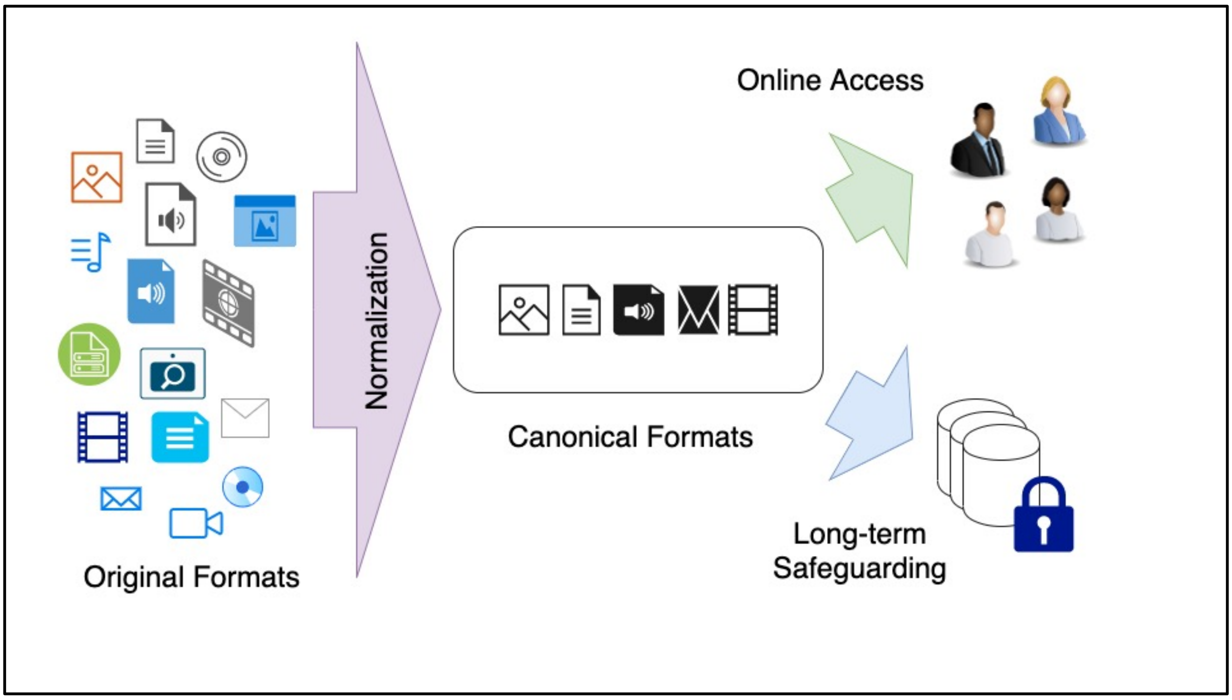
When you have decided what digital preservation actions you need to take, then my next suggestion is to automate the preservation process as much as you can.

This helps you preserve content more quickly - and most importantly make your content available sooner for people to use.

Automation helps lower costs.

Automation means software and machines can do the boring repetitive work.   This means people have more time to do interesting things like curation, publication, supporting users or doing new research.

And automation means DP is more repeatable and less error prone, which again helps get more done.



For example, if you are going to do file format conversions, for example to create access copies to support online reuse of content, then automate the process.

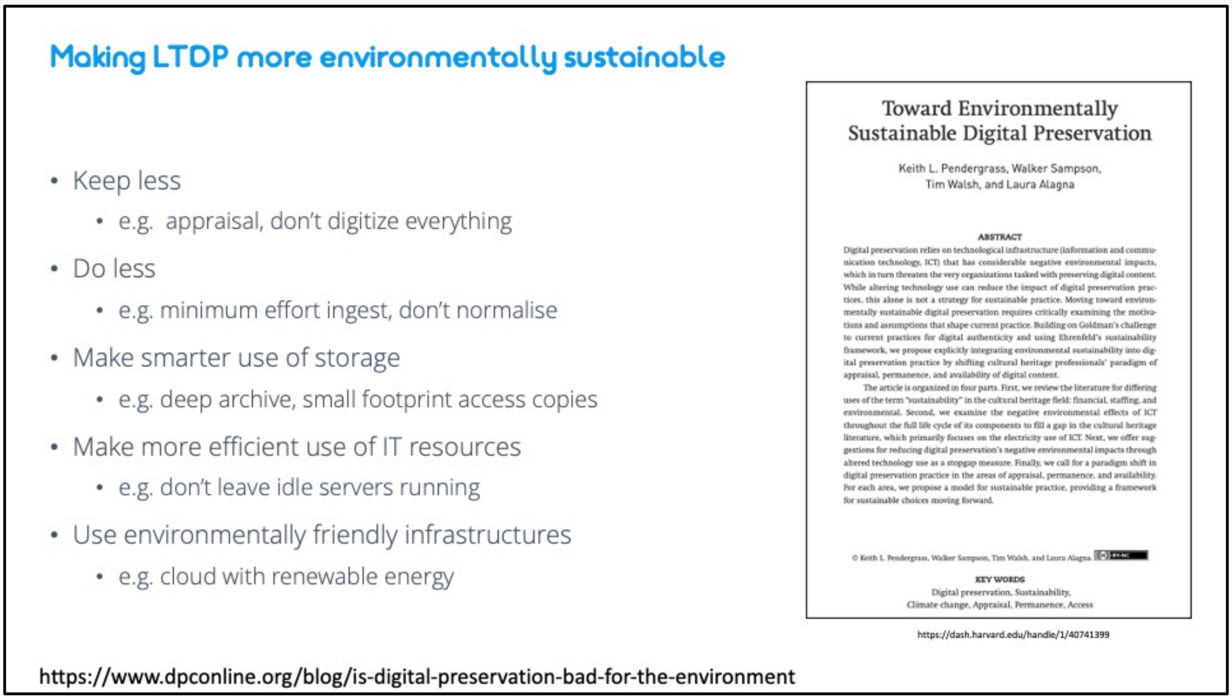
This means automating file format identification, characterization, normalization and validation as much as you can.

This is what digital preservation system do for you – and Arkivum’s solution is just one example.

Automation can bring issues of quality control and validation – which sometimes needs manual processes, for example people checking the results of file format conversions.

But there is a big risk in digital preservation of striving for perfection one-file-at-a time and just not getting enough done – which means content gets left behind and never gets preserved at all.

So put all your content into a simple and automated digital preservation system right at the start – and then do detailed QC afterwards.



Making LTDP more environmentally sustainable

- Keep less
  - e.g. appraisal, don't digitize everything
- Do less
  - e.g. minimum effort ingest, don't normalise
- Make smarter use of storage
  - e.g. deep archive, small footprint access copies
- Make more efficient use of IT resources
  - e.g. don't leave idle servers running
- Use environmentally friendly infrastructures
  - e.g. cloud with renewable energy

Toward Environmentally Sustainable Digital Preservation

Keith L. Pendergrass, Walker Sampson,
Tim Walsh, and Laura Alagna

ABSTRACT

Digital preservation relies on technological infrastructure (information and communication technology, ICT) that has considerable negative environmental impacts, which in turn threaten the very organizations tasked with preserving digital content. While altering technology use can reduce the impact of digital preservation practices, this alone is not a strategy for sustainable practice. Moving toward environmentally sustainable digital preservation requires critically examining the motivations and assumptions that shape current practice. Building on Goldman's challenge to current practices for digital authenticity and using Ehrenfeld's sustainability framework, we propose explicitly integrating environmental sustainability into digital preservation practice by shifting cultural heritage professionals' paradigm of appraisal, permanence, and availability of digital content.

The article is organized in four parts. First, we review the literature for differing uses of the term "sustainability" in the cultural heritage field: financial, staffing, and environmental. Second, we examine the negative environmental effects of ICT throughout the full life cycle of its components to fill a gap in the cultural heritage literature, which primarily focuses on the electricity use of ICT. Next, we offer suggestions for reducing digital preservation's negative environmental impacts through altered technology use as a stopgap measure. Finally, we call for a paradigm shift in digital preservation practice in the areas of appraisal, permanence, and availability. For each area, we propose a model for sustainable practice, providing a framework for sustainable choices moving forward.

© Keith L. Pendergrass, Walker Sampson, Tim Walsh, and Laura Alagna. BY-NC

KEY WORDS

Digital preservation, Sustainability,
Climate change, Appraisal, Permanence, Access

https://dash.harvard.edu/handle/1/40741399

https://www.dpconline.org/blog/is-digital-preservation-bad-for-the-environment

And finally, my third point is around environmental sustainability.

LTDP means use of IT resources: every time digital content is stored, processed, moved, downloaded, viewed, migrated or fixity checked.

Use of IT resources means manufacturing and disposal of computing equipment, e.g. servers.

IT resources means constructing buildings to house and protect that equipment.

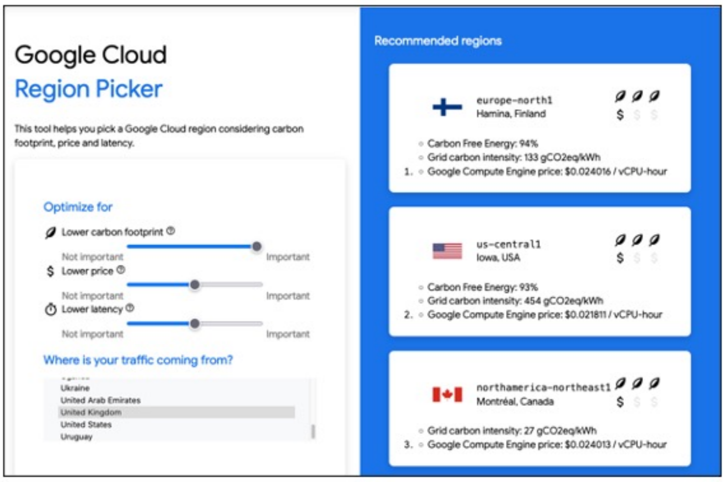
That all has an environmental impact, which is called the 'embodied carbon footprint'

Then comes use of those IT resources. Which means power and cooling – and an additional environmental impact depending on where the energy comes from, for example fossil fuels or renewables.

There has been quite a lot of discussion recently in the DP community on ways to make LTDP more environmentally sustainable with some of the possibilities shown here.

https://www.dpconline.org/blog/is-digital-preservation-bad-for-the-environment
https://dash.harvard.edu/handle/1/40741399

Digital Preservation in the Cloud



Google Cloud
Region Picker

This tool helps you pick a Google Cloud region considering carbon footprint, price and latency.

Optimize for

Lower carbon footprint — Not important / Important

Lower price — Not important / Important

Lower latency — Not important / Important

Where is your traffic coming from?

Ukraine
United Arab Emirates
United Kingdom
United States
Uruguay

Recommended regions

1. europe-north1 — Hamina, Finland — $
   - Carbon Free Energy: 94%
   - Grid carbon intensity: 133 gCO2eq/kWh
   - Google Compute Engine price: $0.024016 / vCPU-hour
2. us-central1 — Iowa, USA — $
   - Carbon Free Energy: 93%
   - Grid carbon intensity: 454 gCO2eq/kWh
   - Google Compute Engine price: $0.021811 / vCPU-hour
3. northamerica-northeast1 — Montréal, Canada — $
   - Grid carbon intensity: 27 gCO2eq/kWh
   - Google Compute Engine price: $0.024013 / vCPU-hour

https://cloud.withgoogle.com/region-picker/

https://www.google.com/about/datacenters/gallery/#st-ghislain-solar-panels

https://www.google.com/about/datacenters/gallery/#hamina-exterior-landscape

But there is hope too. For example, digital preservation in the cloud has a lot of potential.

It is possible to choose datacentres that use almost entirely carbon free energy.

For example, Arkivum has been involved in a large-scale digital preservation project recently called ARCHIVER and we've deployed our solution into Google Cloud.

This means we can do digital preservation with minimal environmental impact.

But it's not just a case of 'lift and shift' to a green datacentre or cloud provider.

It's a combination of things:

(i) Minimize the need for resources, which goes back to my point of keeping it simple.
(ii) Use automation so resources are used efficiently and only when necessary
(iii) Host your system somewhere that has a low carbon footprint, for example using carbon free energy.

arkivum
Bringing archived data to life

QUESTIONS?

matthew.addis@arkivum.com orcid.org/0000-0002-3837-2526 www.arkivum.com

www.arkivum.com | hello@arkivum.com

Hopefully, I’ve provided some useful suggestions and pointers on digital preservation.

I’m sure that these will be revisited in various ways during the rest of this event. I can see that there are some really interesting talks coming up on topics such as automation, file format migrations, quality assurance and trusted digital repositories. I’m very much looking forward to those.

Thank you.