



IPRES 2023

Digital Preservation in Disruptive Times

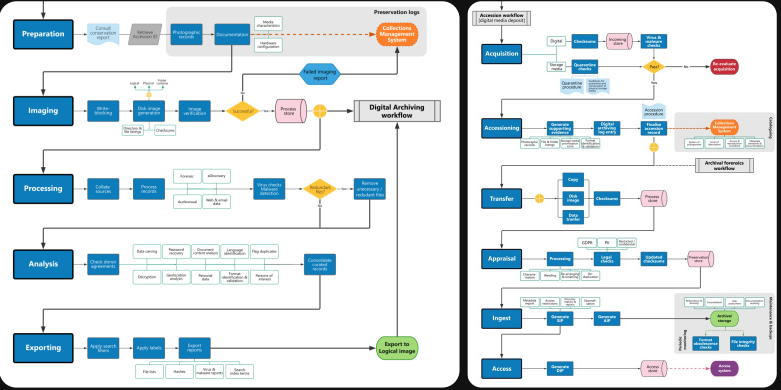
19th International Conference ■ Champaign-Urbana, Illinois ■ September 19–22, 2023


Prioritizing Storage Media for Digital Archiving and Preservation

Leo Konstantelos & Emma Yan, U of Glasgow

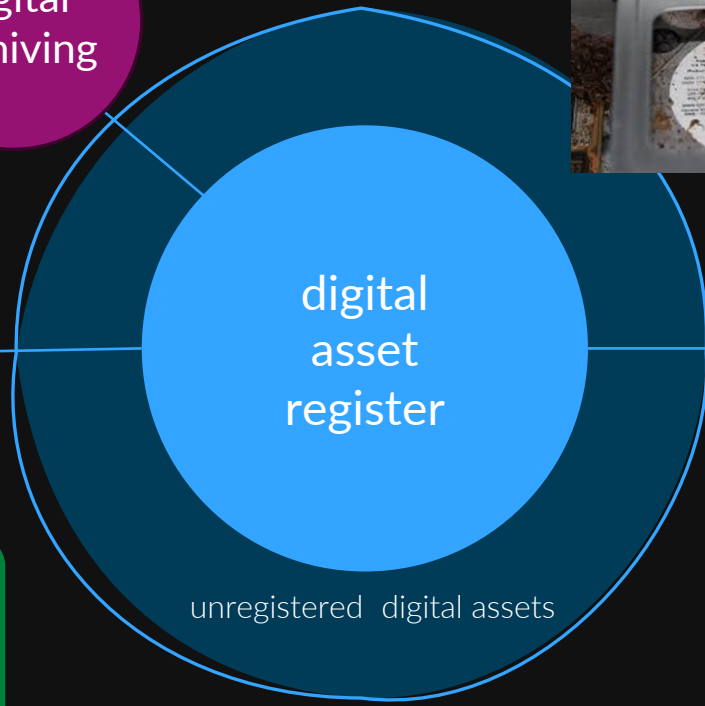
Thu 21st Sep 2023, 09:00-09:15


PROBLEM STATEMENT



 prioritizing collections

has media?



 **prioritizing storage media:**

- degradation while in store
- legacy and obsolete media
- conditions of storage and handling pre-acquisition
- quality, lifespan and age of media

ARCHIVES & SPECIAL COLLECTIONS

EXISTING WORK

Methodology:

- identify criteria for prioritization
- generate score scales per criterion
- define priority score and action
- collate knowledge on average lifespan and level of endangerment
- wrap it all up in a nifty tool!

The Global 'Bit List' of Endangered Digital Species

Risk Classifications

Lower Risk: Digital materials are listed as Lower Risk when it is clear that the requirements for their ongoing preservation are met, and there is a strong likelihood of their ongoing preservation.

Vulnerable: Digital materials are listed as Vulnerable when the requirements for their ongoing preservation are not fully met, and there is a significant risk of their ongoing endangerment.

Endangered: Digital materials are listed as Endangered when the requirements for their ongoing preservation are not met, and there is a high risk of their ongoing endangerment.

Critically Endangered: Digital materials are listed as Critically Endangered when they are at the highest risk of becoming unavailable for future generations.

Practically Extinct: Digital materials are listed as Practically Extinct when they are no longer accessible, and there is no realistic prospect of their recovery.

Concern: Digital materials are listed as Concern when an active member of the Digital Preservation community has been identified as being responsible for their ongoing preservation, and there is a high risk of their ongoing endangerment.

Digital Species

Sound and Video	Research Outputs	Portable Media	Digital Legal Records	Social Media
Gaming	Integrated Storage	Formats	Media Art	Public Records
Sensitive Data	Web	Community Archives	Museums Data & Collections	Political Data
Apps	Engineering Data	Orphaned Works	Personal Archives	

Digital Preservation Guidance Note: 2

The National Archives

Selecting Storage Media for Long-Term Preservation

Digital Preservation Guidance Note 2: Selecting storage media for long-term preservation

Document Control

Author: Adrian Brown, Head of Digital Preservation Research

Document Reference: DPGN-02

Issue: 2

Issue Date: August 2008

Digital Preservation Guidance Note: 3

The National Archives

Care, Handling and Storage of Removable media

Digital Preservation Guidance Note 3: Care, handling and storage of removable media

Document Control

Author: Adrian Brown, Head of Digital Preservation Research

Document Reference: DPGN-03

Issue: 2

Issue Date: August 2008



You've Got to Walk Before You Can Run:

First Steps for Managing Born-Digital Content Received on Physical Media

By Ricky Enay

OCLC Research

WORLD LIBRARY AND INFORMATION CONGRESS: 74th GENERAL CONFERENCE AND CONGRESS - 01-06/2008

Quebec

Date: 01/07/2008

Media Matters: developing processes for preserving digital objects on physical carriers at the National Library of Australia

Authors: Douglas Elford, Nicholas Del Pozo, Snezana Mihajlovic, David Pearson, Gerard Clifton, Colin Webb (presenter), National Library of Australia, Canberra, Australia

IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING (PREPRINT)

Lifespan and Failures of SSDs and HDDs: Similarities, Differences, and Prediction Models

Riccardo Pincirolli, Lishan Yang, Jacob Alter, and Evgenia Smirni

Abstract—Data center downtime typically centers around IT equipment failure. Storage devices are the most frequently failing components in data centers. We present a comparative study of hard disk drives (HDDs) and solid state drives (SSDs) that constitute the typical storage in data centers. Using six-year field data of 100,000 HDDs of different models from the same manufacturer from the Backblaze dataset and six-year field data of 30,000 SSDs of three models from a Google data center, we characterize the workload conditions that lead to failures. We illustrate that their root failure causes differ from common expectations and that they remain difficult to discern. For the case of HDDs we observe that young and old drives do not present many differences in their failures. Instead, failures may be distinguished by discriminating drives based on the time spent for head positioning. For SSDs, we observe high levels of silent mortality and characterize the differences between infant and non-infant failures. We develop several machine learning failure prediction models that are shown to be surprisingly accurate, achieving high recall and low false positive rates. These models are used beyond simple prediction as they act as a surrogate for the complex interaction of workload characteristics that lead to failures and identify failure root causes from monitored symptoms.

Index Terms—Supervised learning, Classification, Data centers, Storage devices, SSD, HDD

Digital Preservation Coalition

Digital Preservation Handbook

Explore the Handbook

- Home
- Contents
- Introduction
- Digital preservation briefing
- Getting started
- Organisational activities
 - Creating digital materials
 - Acquisition and appraisal
 - Decision tree

Legacy media




Illustration by Jergen Stamp digitalbevaring.dk. CC BY 2.5 Denmark

arcserve

Data Storage Lifespans: How Long Will Media Really Last?

APRIL 18TH, 2013



NATIONAL PARK SERVICE

Conserve O Gram

October 2010 Number 22/5

Digital Storage Media

iPRES 2015

CHAPEL HILL - NOVEMBER 2-6

Alternatives for Long-Term Storage Of Digital Information.

Chris L. Erickson
Brigham Young University
HBLL 2217, Provo, UT 84602
1-801-422-1851
chris_erickson@byu.edu

Barry M. Lunt
Brigham Young University
265 CTB, Provo, UT 84602
1-801-422-2264
luntb@byu.edu

Archiving 2011

May 16-19, 2011
Salt Lake City, Utah

How Long is Long-Term Data Storage? (Focal), Barry M. Lunt, Brigham Young University, and Douglas Hansen, Wayne Rust, and Mark Worthington, Millenniata, Inc. (USA)



Assumptions / constraints

- only include media we can process in archival forensics lab
- no other copy of digital assets other than in storage media → **unique records**
- storage conditions pre-acquisition will be considered as “aggravating”
- for hybrid collections, digital assets in storage media will be considered as valuable and important as the rest of the items in the collection

Storage media-specific criteria

- Average lifespan (from existing literature)
- Year of production → measure of longevity and obsolescence
- Environmental conditions of storage **post-acquisition**
- 'Bit List' of Digitally Endangered Species classification



METHODOLOGY

CRITERIA AND SCORES

Bit List' of Digitally Endangered Species	
Classification	Score
Lower risk	1
Vulnerable	2
Endangered	3
Critically Endangered	4
Practically extinct	5
Average lifespan	
Lifespan	Score
1-3 years	5
3-5 years	4
5-10 years	3
10-20 years	2
More than 20 years	1
Conditions	
Conditions	Score
Optimal conditions	1
Good conservation practice	2
Minimal conservation practice	3
Some aggravating conditions	4
Mostly aggravating conditions	5
Year of production	
Produced	Score
Within the last 5 years	1
More than 5 years ago	5

KNOWLEDGE

Medium	Produced	Bit list status	Average lifespan (years)
Current internal HDD	Within the last 5 years	Vulnerable	3-5 years
Current internal SSD	Within the last 5 years	Vulnerable	3-5 years
Non-current internal HDD	More than 5 years ago	Critically Endangered	3-5 years
Non-current internal SSD	More than 5 years ago	Critically Endangered	3-5 years
Current portable HDD	Within the last 5 years	Endangered	3-5 years
Current portable SSD	Within the last 5 years	Endangered	3-5 years
Current optical media (CD, DVD, BlueRay)	Within the last 5 years	Endangered	5-10 years
Current magnetic tape	Within the last 5 years	Endangered	10-20 years
Current Flash storage (USB stick, SD card)	Within the last 5 years	Vulnerable	3-5 years
Floppy disk	More than 5 years ago	Critically Endangered	1-3 years
Non-current magnetic tape	More than 5 years ago	Critically Endangered	10-20 years
Cassette tape	More than 5 years ago	Critically Endangered	10-20 years

PRIORITY SCORE

Score	Priority level
1	Low priority - action within 3 years
2	Low priority - action within 1 year
3	Medium priority - action within 6 months
4	High priority - action within 3 months
5	Extreme priority - immediate action



$$(bit-list \times 0.25) + (lifespan \times 0.25) + (YoP \times 0.25) + (conditions \times 0.25) = priority\ score$$

$$(4 \times 0.25) + (4 \times 0.25) + (5 \times 0.25) + (4 \times 0.25) = 4.25$$



NIFTY TOOL!

STORAGE MEDIA PRIORITISATION

Select storage medium:

Non-current internal HDD

Storage medium details:

Produced

Bit List Status

More than 5 years ago

Critically Endangered

What conditions has the storage medium been kept in?

Some aggravating conditions

Priority score: **4**

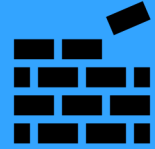
Priority action: **High priority - action within 3 months**

Currently developed as:

- an AppSmith app (internal use only)
→ <https://www.appsmith.com/>
- an Excel spreadsheet app
- web version forthcoming



FUTURE WORK



Work in progress

Open to discussion and community feedback



Ongoing piece of work

Bound to change as knowledge on storage media evolves



Current feedback and planned changes

- storage conditions should be given a higher weight
- integration of storage media prioritization with other selection and appraisal practices / decision-making
- automated score generation as part of other tools / processing (e.g. forensics tools)
- Community review and updates of knowledge base, potential for integration into Bit List

DISSEMINATION & FEEDBACK

Current version of the methodology:

- available for download Q1 2024
- Excel spreadsheet with formulae to allow score weighting customisation
- Includes criteria, scales and collated knowledge
- Licenced under a [CC BY-NC-SA](#)

Feedback, questions, suggestions, corrections:

- Email us! Contacts in next slide

CONTACTS

ARCHIVES & SPECIAL COLLECTIONS

Dr. Leo Konstantelos, Senior Assistant Archivist (Digital):



leo.konstantelos@glasgow.ac.uk



[@lkonstantelos](https://twitter.com/lkonstantelos)

Emma Yan, Assistant Archivist (Accessions):



emma.yan@glasgow.ac.uk



[@eswyan](https://twitter.com/eswyan)



University
of Glasgow

THANK YOU!

#UofGWorldChangers

f   **@UofGlasgowASC**