http://doi.org/10.7207/twgn22-02

A Risk Driven Approach to Bitstream Preservation

Paul Wheatley

DPC Technology Watch Guidance Note

Updated January 2025

DigitalPreservationCoalition

Version 1.3, 2025

First published in Great Britain in 2022 by the Digital Preservation Coalition.

This Guidance Note is published by the Digital Preservation Coalition (DPC). The DPC is an international charitable foundation which supports digital preservation and helps its members around the world to deliver resilient long-term access to digital content and services. In addition to the publication of reports on a range of themes which cover the state of the art in digital preservation, the DPC also supports its members through community engagement, targeted advocacy work, training and workforce development, identification of good practice and standards, and through good management and governance. Its vision is a secure digital legacy.

Discover the DPC's publications, including the latest updates and revisions, at:
https://www.dpconline.org/digipres/discover-good-practice/tech-watch-reports

Find out more about the DPC, the support it offers and how to become a member at:
https://www.dpconline.org/about/join-us

Version 1.0 first published in 2022.

Version 1.3, January 2025. Revised references, some restructuring and rewording, additional examples and expansion of references on cloud storage challenges and mitigations.

1 Introduction

Storing digital content without loss is far from the only consideration in achieving long term digital preservation. But as the DPC's unofficial tag line of "Keep the bits" highlights, it is a crucial one. The DPC Rapid Assessment Model, or DPC RAM, (DPC, 2024) states that in order to achieve Level 3 "Managed" for the "Bitstream Preservation" criterion, "A process of risk assessment is used to evaluate storage risks and appropriate mitigations (such as the number of copies, location, technologies used, frequency of integrity checking)". But what does this really look like in practice?

This Guidance Note explores some of the challenges that must be addressed by a storage architecture designed for long-term digital preservation. It considers the threats faced by content stored over the long-term and describes a simple approach to assess and document the level of risk and mitigating actions put in place.

Digital preservation practitioners may find this Guidance Note useful when seeking to establish appropriate preservation storage or when verifying that their existing storage is fit for purpose. It may assist with making the case within an organisation for the resources to provide further mitigation to address identified threats to stored digital content. In addition to the more broadly scoped "Core requirements for a digital preservation system" (DPC, 2022) this Guidance Note might be helpful in communicating the somewhat unique requirements of long-term digital preservation when engaging with IT staff.

Note that when considering an appropriate storage architecture for preservation it is important to consider and document many other storage requirements, such as access, interoperability, portability and scalability. These are considered to be outside the scope of this Guidance Note but are described in detail in the "Preservation Storage Criteria, Version 4" (Schaefer et al, 2024).

1.1 What is "Bitstream Preservation"?

The DPC RAM describes Bitstream Preservation as "Processes to ensure the storage and integrity of digital content to be preserved." The information we are interested in preserving is encoded in various ways, often utilising specific file formats. Ultimately it is represented by a series of zeros and ones. These are known as binary digits or bits. A series of bits, perhaps representing a file, is often referred to as a bitstream. Bitstream Preservation is concerned with the ongoing survival of these bits. It does not address how information is encoded in these bits, or more crucially for preservation, how a bitstream can be decoded into useful information (referred to as the distinct but related "Content Preservation" within DPC RAM).

1.2 Understanding digital preservation storage requirements

Although the threats to data storage faced by the digital preservation world are not entirely dissimilar to those in a more typical IT setting, the differing requirements of preservationists might require a different approach be taken. Rosenthal et al (2005) note that "many of these threats are not unique to digital preservation systems, but their specific mission and very long time horizons incline such systems to view the threats differently from more conventional systems."

A typical IT storage regime might be designed to deliver resilient services (i.e. with little downtime) for users now, with some facility for backup and recovery over the short-term. Long-term digital preservation is usually less concerned with interruptions to immediate operations or access to content for users (downtime), but must be able to ensure that no (or little) content will be lost over the genuine long-term (often defined as 100 years or more) (Prater, 2018).

A likely IT response to a substantial ransomware attack might be to establish new business services from the ground up, often leveraging cloud-based technology. But this disaster recovery approach that is focused on business services isn't as relevant to the world of digital preservation where keeping data is king. To put it bluntly, if all the digital content that has been patiently collected for decades is lost, there is no recovery from disaster. There is just disaster.

2 A risk-based approach

The following text is a simple guide to assessing the threats that digital content faces within a particular organization, and consideration of an appropriate set of mitigations to address them.

The obvious but simplistic question of "How many copies should I keep?" might typically but perhaps crudely be answered as "3" (NDSA 2019). More detail is required to understand the effectiveness of risk mitigation offered by the multi-copy redundancy in question. For example, keeping 3 copies (where most users are prevented from accessing all 3 copies) will reduce the likelihood of data loss from human error. But if all 3 of those copies sit within the same building, they provide little mitigation to the threat of fire.

There is no single solution that will suit all contexts. One organization will likely have a different threat profile to another. Each organization's appetite for risk, the value of its content and the resources it has available for mitigation will vary.

This Guidance Note is therefore designed to guide a practitioner through a process of assessing and reflecting on the threats to the storage of their own digital content in a manner that is appropriate for *their* organization and *their* requirements.

2.1 What causes data loss?

Organizations are often reluctant to share the details of data losses, but a common theme in those that have been publicized appears to be the realisation of multiple threats at the same time. This might include human error, power outage, natural/human made disaster, or unexpected software behaviour due to bugs (The Register, 2017a; The Register, 2017b). Data comes under greater risk when it is in motion, such as management processes, data transfer or migration due to media refreshment. Other common factors include failing to fully complete or verify processes such as integrity checking, patching software or backing up content. An important lesson to learn from this is that establishing processes to mitigate storage risks is, on its own, insufficient to ensure preservation. The mitigation processes must themselves be carefully monitored, validated and ideally assessed independently to ensure their continuing effectiveness.

Ransomware continues to be a critical risk to the survival of digital content. The high-profile attack on the British Library (Wikipedia, 2023) and the lack of consensus on how to tackle the burgeoning wave of ransomware attacks (The Register, 2024) highlights the continuing need for digital preservation to mitigate against this threat in ways that go above and beyond typical cyber security measures.

The advent of outsourcing storage has reduced the risk of loss due to hardware failure, but has led to new threats to data. In 2024 Unisuper, an Australian fund manager, discovered that their "entire infrastructure subscription was deleted" (Mellor 2024). Both copies of their data in Google Cloud were lost, although most data was ultimately recovered from a third copy stored with a different cloud provider.

The NDSA 2021 Fixity Survey (NDSA, 2021) surveyed organizations operating in the long-term preservation community. It appears to suggest that few organizations who responded suffered frequent integrity checking failures and only some of these were associated with storage dedicated to digital preservation. Figures are not provided on the scale or impact of these failures, but detail on the nature of the cause and rectification implies that many of these were of a small scale. This provides some confidence that mitigations are functioning with a degree of effectiveness. Continuing to gather richer data in this area is likely to be invaluable in informing appropriate preservation approaches and justifying investment in their implementation, with hopefully a minimum of economic and carbon cost (Stokes, 2022).

2.2 Why use a risk-based approach to storage?

Bitstream Preservation is a fundamental building block for ensuring digital information can be preserved for the long-term and ultimately accessed with its value realised. It is therefore critical to ensure that threats are identified, understood and appropriately managed. However, there are additional benefits to this approach.

Documenting threats and mitigations is widely acknowledged as being good practice. A formal risk assessment will result in documentation of the process and the result, providing evidence of appropriate preservation planning activities. This might be necessary for archives/preservation certification. The Core Trust Seal certification standard asks the question "Are risk management techniques used to inform the strategy?" and requires documentary evidence in order to reach compliance (Core Trust Seal, 2022).

A risk assessment can also act as useful evidence when making the case for resources to implement additional risk mitigation. Communicating the reasoning behind the somewhat unique requirements for long-term digital preservation remains a significant organizational challenge. Highlighting gaps in preservation capability and the impact of possible of loss of content, financial cost and reputational damage can be a powerful way to get senior managers onside.

2.3 Steps in applying a risk assessment for digital preservation storage

A simple risk assessment process will likely be sufficient to guide consideration of acceptable (or unacceptable) preservation risk, but it must be comprehensive in scope and an honest assessment of the threats, the likelihood of their realisation and their potential impact. The ISO 27001 standard on information security (Wikipedia, 2022) may provide useful guidance in defining and documenting a risk assessment approach. Many organisations will have their own risk management process that might usefully be applied for this task. Otherwise, a simple risk assessment process is outlined below:

1. Identify and record the scope of your risk assessment, detailing in particular the digital content to which it will apply.
2. Identify significant threats to your digital content, relevant to your defined scope.
3. Score the likelihood of each threat occurring and the scale of the impact it will have if it is realised. These scores can be multiplied together, to generate an initial score for each threat.
4. Document existing mitigation actions that are in place at your organization, and provide an adjusted score for each threat that takes into account the mitigation.
5. Consider your organization's appetite for the adjusted risk scores that have been generated. It may be useful to consult with a range of internal stakeholders, senior management and possibly external advisors such as the DPC or peer organizations.

6. Document any additional mitigation actions that are deemed necessary to further address outstanding levels of risk.

There is no one correct answer as to the question of what mitigations are appropriate for a particular organization. Any particular mitigation action may lower preservation risk, but will likely also result in a financial cost and possibly also an environmental cost. It may be necessary to consider the uniqueness and value of content to be preserved (or conversely the financial or reputational cost of losing the content) and what level of loss might be acceptable (Pendergrass et al, 2019). Consequently it may be useful to develop risk assessments for different collections or document levels of preservation commitment as at Penn State University Libraries (2021). Digital preservation systems are increasingly providing facilities for users to tailor storage profiles to particular holdings.

The following table provides a summary of common storage threats and some typical mitigation actions that might be associated with them, but other threats and mitigations may be relevant to your situation:

Storage threats	**Potential mitigation actions**
Bit rot / loss or damage to content	• Replicate content to create redundant copies • Implement integrity checking and repair
Storage hardware failure	• Monitor, manage and repair/replace storage hardware • Implement integrity checking and repair
Storage media/hardware obsolescence	• Plan and implement refreshment/replacement of storage media/hardware before end of life
Accidental deletion / human error / malicious damage by staff	• Replicate content to create redundant copies • Ensure rigorous write-access control and principle of least privilege • Implement integrity checking and repair • Establish process for managing legitimate content change/disposal • Document/audit all actions resulting in content alteration
Malicious damage by external party	• Implement cyber security measures • Replicate content to create redundant copies • Retain copies of content in different management regimes • Establish offline copy of content • Implement integrity checking and repair
Common mode failure (single point of hardware / software failure affecting all replicated copies)	• Use a mix of hardware / software technologies
Natural / human made disaster	• Replicate content to geographically separated locations with differing risk profiles • Establish disaster recovery policy and procedure
Failure or closure of third-party storage provider	• Establish plan of action in event of unexpected closure • Avoid dependence on a single third-party vendor (eg. cloud provider) • Ensure independent access to cloud storage resold by third-party provider • Utilise escrow facilities

Failure to implement risk mitigation processes (above) or verify they are functioning effectively	• Document storage management procedures • Test and validate mitigation actions • Provide clear reporting on the implementation of risk mitigation processes to active governance body • Establish independent audit/certification of processes and procedures for long-term preservation

An overriding theme of these mitigations is the need for *diversity* within a storage architecture. Diversity seeks to remove or at least minimize single points of failure including people, software, hardware, geographical location or service provider.

Anecdotally, the digital preservation community has often identified human error as the most significant threat to digital content. Consider where human error might play a role in the likelihood and impact of all of the threats outlined above. The growing threat of ransomware attack is also considered amongst the most critical risks faced.

The following provide valuable further reading on risk management in the context of preservation storage:

- The "Usage Guide for the Preservation Storage Criteria" describes a useful categorisation of risk types as well as more in-depth discussion of risk management, bit integrity and independence of storage copies (Schaefer et al, 2019).
- An in-depth exploration of risk management in digital preservation is provided by Pennock (Pennock, 2024).
- The "Digital Archiving Graphical Risk Assessment Model", or DiAGRAM, tool (National Archives, 2020) uses a statistical method called a Bayesian network to produce a graphical model of digital preservation risks, which includes a focus on digital preservation storage.

3 Understanding an evolving landscape of risk mitigation

A variety of threats to our digital content lie somewhat hidden within the applications, middleware and 3rd party services we depend on to manage our digital content. Outsourcing our preservation functions can have great benefits, but it also changes the threat profile to our digital content. This section considers some of what we know – and don't know – about these evolving technologies and services, and what approaches this community is beginning to implement to mitigate these new and emerging threats.

3.1 A clouded picture of storage diversity, redundancy and service provision

Cloud storage services offer a host of potential benefits in storing content for the long term and in convenience. However, a number of potential risks and concerns have been raised with cloud storage, despite its rapid uptake within the digital preservation community. How much trust can safely be placed in outsourcing not only storage, but also the integrity checking of the storage and other processes such as media refreshment? Rosenthal (2019) noted in 2019 that "Verifying the integrity of data stored in a cloud service without trusting the service to some extent is a difficult problem to which no wholly satisfactory solution has been published."

The ubiquity and scale of cloud storage beyond the digital preservation domain suggests a resilient technology, and one that is likely to be far more rigorously tested than many potential points of failure in an onsite digital preservation architecture. The understandably risk-averse digital preservation community has so far however adopted a variety of ways of employing cloud storage:

- The Wellcome Library has moved to a cloud only approach, but introduces diversity into its storage by using two different cloud providers (Chan, 2021). Chan notes that "The level of data integrity and safety they (the cloud) provide goes far beyond anything we could build in-house. We trust their verification process, and don't do any additional checking of data at rest." Wellcome does however read back data moved to cloud, in order to double check successful deposit.
- The National Archives (UK) has outsourced storage and its preservation platform to a third party, with two copies held in the cloud. Crucially, they also maintain a third "custodial" copy on site (Daly, 2024), in a form based the Oxford Common File Layout (Jefferies and Woods, 2024) that is independently understandable. This adds diversity to mitigate storage threats and decoupling from the managed service provides agility in light of a rapidly changing picture of technologies and service providers.
- The National Library of Scotland uses a mix of onsite and cloud storage, whilst applying full or sampled integrity checking of different storage nodes (Hibberd, 2020).
- The Natural Environment Research Council has avoided cloud services altogether, with the advantage of having complete control over their integrity checking functions (NDSA, 2021, p.68).

3.2 How to verify preservation effectiveness when storage and integrity checking have been outsourced?

Most cloud storage providers include integrity checking with their storage services, but they typically don't expose how integrity checking at rest is performed. Claims of high durability remain difficult to verify. Around half of respondents to the NDSA 2021 Fixity Survey (NDSA, 2021, p.44) who used cloud storage reported receiving some integrity information from their providers. Around a third of respondents who received integrity information were unable to make use of it, for a variety of reasons. There is some evidence that 3rd parties are listening to feedback from users and are enhancing support for integrity checking, at least while it is in transit to and from the cloud (Stormacq, 2024).

Independently checking the integrity of data at rest in the cloud by recalculating checksums has been successfully demonstrated by the University of Illinois at Urbana-Champaign, despite practical and economical challenges (Rimkus and Schmitt, 2024). In an ideal world, organizations would not need to pay to duplicate the integrity checking claimed by cloud service providers and effectively pay twice for a function they have outsourced. Having gained some confidence in their providers with full and independent integrity checks in previous years, the University of Illinois has since implemented a policy of checking new content and random sampling of content at rest instead of further full integrity checks. "The goal here is to trust that this level of file durability will continue, but to verify this durability at a modest pace in order to uncover flaws in storage and file management, should these arise."

The community has begun to discuss, and in some cases implement, alternative and/or complimentary approaches to full integrity checks. This has in part been prompted by the identification of threats that are associated in particular with outsourced cloud storage such as the account deletion example referenced in 2.12.1 above. These include:

- Monitoring changes in cloud storage item inventories or performing full file manifest checks that do not go as far as computationally intensive (and therefore cloud compute costly) integrity checks. This is a good way to spot problems introduced by human error or software

problems like errors in automated workflows, while trusting the storage platform to maintain data at rest.

- Testing if a 3rd party service is still live by requesting a small number of files on a daily basis (Altman and Landau, 2024). Quickly identifying that a storage service is no longer functioning would enable rectification to be raised or alternatives put in place.
- Separating contract management of multiple cloud instances held with different providers to different organizational departments and staff owners to reduce the possibility of accountancy or billing errors resulting in the accidental closure of all copies at the same time.
- Automated checking of cloud provider billing via API. A sudden and significant change in regular billing for cloud services might indicate a catastrophic change as to what is (or isn't) being stored.

Ultimately, the scale of usage of cloud storage services and lack of publicised bit rot issues suggests that replicating full integrity checks performed by cloud providers is unnecessary. This community is however yet to settle on an agreed level of trust and risk mitigation for the use of cloud services for long term preservation.

3.3 Hidden single points of failure?

Hockx-Yu and Brewer (2021) note concern about storage intermediaries such as cloud gateways like the AWS Storage Gateway, and tape storage gateways such as Spectra Logic's BlackPearl Converged Storage System. Storage intermediaries can provide convenient access to multiple and seemingly diverse storage locations but at the same time can introduce single points of failure and single points of external attack. They state "Storage intermediaries directly challenge the notion of redundancy…" They recommend "…to raise awareness and deepen understanding of them, especially how they could become the single point of failure leading to data loss or digital preservation".

Commercial or open-source digital preservation system applications are increasingly used to manage and deliver storage, integrity checking and a variety of other services of relevance to this Guidance Note. Consideration should be given to the potential of these systems in presenting single points of failure, especially where they are used to manage otherwise diverse storage locations. Examples of loss as a result of software bugs within preservation systems have been encountered. Preservationists should continue to challenge preservation system vendors in this area, and work with them in reporting and addressing any potential issues that might be identified.

4 Conclusion

The design and implementation of storage architectures for long-term digital preservation is often influenced or led by a host of factors unrelated to the concerns of keeping data for long periods. Issues such as limited resourcing, practicality, organizational policies to outsource IT and many more can distract attention from a key question – is a particular storage architecture sufficient to preserve data for the long-term? A simple risk assessment process can be a useful approach to answer this question and make the case for any additional mitigation work to be employed. This approach is equally valid amidst the rapid move towards outsourcing preservation functions. Risk assessment continues to reveal that storage diversity is hugely desirable. It is this diversity that is the key to mitigating the wide variety of threats that our content will face over the long-term.

5 References

Addis, M. (2020) *Which checksum algorithm should I use?* Available at: http://doi.org/10.7207/twgn20-12

Altman, M and Landau, R. (2024) *Selecting Efficient and Reliable Preservation Strategies:*

Modeling Long-term Information Integrity Using Large-scale Hierarchical Discrete Event Simulation. IJDC. Available at: https://doi.org/10.2218/ijdc.v18i1.743

Chan, A. (2021) *Our approach to digital verification*. Available at: https://web.archive.org/web/20220809134815/https://stacks.wellcomecollection.org/our-approach-to-digital-verification-79da59da4ab7?gi=f955d8d0d5c1

Core Trust Seal (2022) *Core Trust Seal*. Available at: https://web.archive.org/web/20220802114709/https://www.coretrustseal.org/

Daly, S (2024) *A decoupled Custodial Copy for cloud-based Digital Preservation Systems.* Available at: https://doi.org/10.5281/zenodo.13647419

DPC (2021) *DPC Rapid Assessment Model Version 3*. Available at: https://web.archive.org/web/20250131231437/https://www.dpconline.org/digipres/implement-digipres/dpc-ram

DPC (2022) *Core requirements for a digital preservation system*. Available at: https://web.archive.org/web/20220810111732/https://www.dpconline.org/digipres/implement-digipres/core-requirements-for-a-digital-preservation-system

Jefferies, N and Woods, A. (2024) *The Oxford Common File Layout*, Github, Available at: https://github.com/OCFL *

Mellor, C. (2024) Google Cloud deleted a large customer's infrastructure. Blocks and Files. Available at: https://web.archive.org/web/20250131234033/https://blocksandfiles.com/2024/05/14/google-cloud-unisuper/

National Archives (2020) *DiAGRAM - The Digital Archiving Graphical Risk Assessment Model*. Available at: https://web.archive.org/web/20220809153306/https://nationalarchives.shinyapps.io/DiAGRAM/

NDSA (2019) *2019 Storage Infrastructure Survey*. Available at: https://doi.org/10.17605/OSF.IO/UWSG7

NDSA (2021) *2021 Fixity Survey*. Available at: https://doi.org/10.17605/OSF.IO/2QKEA

Pendergrass, K. L. Sampson, W. Walsh, T. and Alagna, L. (2019) *Toward Environmentally Sustainable Digital Preservation*, American Archivist, Volume 82, Issue 1. Available at: https://web.archive.org/web/20220804114356/https://meridian.allenpress.com/american-archivist/article/82/1/165/432804/Toward-Environmentally-Sustainable-Digital

Penn State University Libraries (2021) *Policy UL-AD19 Digital Preservation Policy*. Available at: https://web.archive.org/web/20220804123736/https://libraries.psu.edu/policies/ulad-19

Pennock, M. (2024) *Disentangling Digital Preservation Risk: An Interdisciplinary Exploration and Solution.* Available at: https://dx.doi.org/10.15132/20000457

Prater S. (2018) *How to Talk to IT about Digital Preservation*, Journal of Archival Organization, Available at: https://web.archive.org/web/20210520101609/https://minds.wisconsin.edu/bitstream/handle/1793/78844/How%20to%20Talk%20to%20IT%20about%20Digital%20Preservation.pdf?sequence=3&isAllowed=y

The Register (2017) *GitLab.com melts down after wrong directory deleted, backups fail*. Available at: https://web.archive.org/web/20220729152515/https://www.theregister.com/2017/02/01/gitlab_data_loss/

The Register (2017) *KCL external review blames whole IT team for mega-outage, leaves managers unshamed*. Available at: https://web.archive.org/web/20220729152543/https://www.theregister.com/2017/02/23/kcl_external_review/

The Register (2024), *What do ransomware and Jesus have in common? A birth month and an unwillingness to die*. Available at: https://www.theregister.com/2024/12/24/ransomware_in_2024/

Rimkus, K.R. and Schmitt, G. (2024). *File Fixity in the Cloud: Policy, Business, and Technical Considerations*. iPRES 2024. Available at: https://doi.org/10.21428/5676bf2d.3d7946ac

Rosenthal et al. (2005) *Requirements for Digital Preservation Systems*, DLib November 2005, Volume 11, Number 11. Available at: https://web.archive.org/web/20220423182212/http://www.dlib.org/dlib/november05/rosenthal/11rosenthal.html

Rosenthal D. (2019) *DSHR's Blog: Cloud for Preservation*. Available at: https://web.archive.org/web/20220804151256/https://blog.dshr.org/2019/02/cloud-for-preservation.html

Schaefer et al. (2019) *Usage Guide for the Preservation Storage Criteria*. Available at: https://osf.io/4cvqa

Schaefer et al. (2024) *Digital Preservation Storage Criteria, Version 4*. Available at: https://osf.io/ym6ua

Stokes, P. (2022). *Catastrophic data loss is going to cost us how much….?!.* Available at: https://web.archive.org/web/20220830114430/https://www.dpconline.org/blog/stokes-cost-of-catastophic-data-loss

Stormacq, S. (2024). *Introducing default data integrity protections for new objects in Amazon S3*. Amazon. Available at: https://aws.amazon.com/blogs/aws/introducing-default-data-integrity-protections-for-new-objects-in-amazon-s3/

Wikipedia (2022) *ISO/IEC 27001*, Available at: https://web.archive.org/web/20220804122211/https://en.wikipedia.org/wiki/ISO/IEC_27001

*Entries marked with an asterisk have not been web archived due to service unavailability

Wikipedia (2023) *British Library cyberattack*, Available at: https://en.wikipedia.org/wiki/British_Library_cyberattack