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

A risk driven approach to Bitstream Preservation

Paul Wheatley

DPC Technology Watch Guidance Note

December 2022

DigitalPreservationCoalition

ISSN: 2048-7916

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

1 Introduction

Storing digital content, our binary data, 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" illustrates, it is a crucial one. The DPC Rapid Assessment Model, or DPC RAM, (DPC, 2021) states that in order to achieve Level 3 "Managed" for the "Bitstream Preservation" criterion, "Decisions on the frequency of integrity checking and the number of copies held take into consideration risks, value of the content and costs (both financial and environmental)". 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 risks faced by content stored over the long-term and concludes with a simple approach to assess and document the risks and mitigating actions put in place to address them.

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 storage risks. 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.

2 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 bitstreams. 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 (known as the distinct but related "Content Preservation" within DPC RAM).

3 Understanding digital preservation storage requirements

Although the risks facing the storage efforts of 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 for users now, with some facility for backup and recovery over the short-term. Long-term digital preservation is usually less concerned with minor interruptions in immediate operation or access to content (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). Preservation needs to be delivered in a transparent and documented manner in order to demonstrate the provenance and authenticity of preserved content for future users. Effective long-term digital preservation therefore requires consistent process, rigorous working and independent verification.

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 and scalability. These are considered to be outside the scope of this Guidance Note but are described in detail in the NDSA Digital Preservation Storage Criteria (Schaefer et al, 2015).

4 A risk-based approach to digital preservation storage

The following text is a simple guide to assessing the storage risks faced in a particular instance, and considering an appropriate set of mitigations for these risks. It should be noted that there is not a one size fits all solution to this problem. One organization may have a different risk profile to another. The appetite for risk, the value of the content and the resources available for mitigation may also vary.

The answer to the obvious question of "How many copies should I keep?" might typically be 3 (NDSA 2019), but of course depends on many other factors such as the frequency of integrity checking (Addis, 2020). As such, a holistic approach to considering storage risk and mitigation is usually more meaningful and more helpful than the consideration of specific issues in isolation.

This Guidance Note is therefore designed to guide a practitioner through the process of reflecting on and assessing their storage risks in order to evaluate and possibly amend their risk mitigation in a manner appropriate for their organization and their unique requirements.

4.1 When do losses occur?

Few major incidents of loss are reported on in detail, but a common theme appears to be the presence of multiple issues occurring at the same time. This might include human error, power outage, natural/human made disaster, or unexpected software behaviour due to software bugs (The Register, 2017a; The Register, 2017b). 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 be monitored, validated and ideally assessed independently to ensure their continuing effectiveness.

The NDSA 2021 Fixity Survey (NDSA, 2021) surveyed organizations operating in the long-term preservation community. It appears to suggest that few organizations 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 are of a small scale. This provides some confidence that risk 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 contributing to future surveys is to be encouraged.

4.2 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 risks are identified, understood and appropriately managed. However, there are additional benefits to this approach. A formal risk assessment will result in documentation of the process and the result, providing evidence of preservation planning activities that 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). It can also act as useful evidence

when making the case for resources to implement additional risk mitigation. Communicating the risks and reasons behind the somewhat unique requirements for long-term digital preservation remains a significant organizational challenge. Perhaps most importantly, assessing risks and documenting the process is widely acknowledged as good practice in this domain as it results in a record of what the risks are, what actions have been taken and why. Subsequent actions or modifications to policy and procedure can then be informed by prior decision making.

4.3 Steps in applying a risk assessment for digital preservation storage

A simple risk assessment process will be sufficient to guide consideration of acceptable (or unacceptable) preservation risk, but it must be comprehensive in scope and an honest assessment of the risks, their likelihood and their impacts. The ISO 27001 standard on information security (Wikipedia, 2022) may provide useful guidance in defining and documenting a risk assessment approach, but the key steps of a simple risk assessment process are 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 risks relevant to your defined scope.
3. Score the likelihood of each risk occurring and the impact that each risk will have if it occurs. These scores can be multiplied, to generate an initial risk score for each risk.
4. Document existing risk mitigation actions in place at your organization, and provide an adjusted score for each risk 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 risks.

There is no one correct answer as to the question of what risk 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), in order to identify an acceptable level of risk. Consequently it may be useful to develop risk profiles for different collections or document levels of preservation commitment as in this example from 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 risks and some typical mitigation actions that might be associated with them, but other risks may be relevant to your situation:

Storage risks/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

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

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. This may be a useful approach for developing a broader risk assessment.

4.4 A clouded picture of storage diversity, replication and service provision

A variety of storage risks lie somewhat hidden within the applications, middleware and 3rd party services we depend on to manage our digital content. This section considers some of what we know – and don't know – about these evolving technologies and services, and what related risks might need to be considered.

Cloud storage services offer a host of potential benefits in storing content for the long term. Outsourcing the management of storage hardware and systems can be convenient, can introduce geographical separation and can introduce some diversity into a storage architecture. 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) notes 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." Most cloud storage providers include integrity checking with their storage services, but don't always expose the details or results. Around half of respondents to the NDSA 2021 Fixity Survey (NDSA, 2021, p.44) who used cloud storage reported receiving 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. Is it sufficient to rely on cloud services if no further information is provided, or reports are impractical to utilise?

The ubiquity and scale of cloud storage beyond the digital preservation domain suggests a resilient technology, and one that is likely to have been far more rigorously tested than many other points of risk in a digital preservation architecture. The understandably risk-averse digital preservation community has so far however adopted a variety of approaches. The Wellcome Library has utilised more than one cloud provider to introduce further resilience, whilst trusting 3rd party integrity checking (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." 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). 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).

Hockx-Yu and Brewer (2021) note concern about potential risks present in 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, regardless of replication and other mitigations that are employed at the storage level. Examples of loss as a result of software bugs within preservation systems have been experienced. 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.

5 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 document the key information required to answer this question and identify where further risk mitigation may be required. It's vital that we continue to learn from experiences involving the loss of digital content, so

that we can justify the resources required to mitigate preservation risks and design those mitigations to be effective with a minimum of economic and carbon cost (Stokes, 2022).

6 References

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

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/

DPC (2021) *DPC Rapid Assessment Model*. Available at: https://web.archive.org/web/20220411235248/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

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

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

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

Schaefer et al. (2015) *Digital Preservation Storage Criteria*. Available at: https://doi.org/10.17605/OSF.IO/SJC6U

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

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/

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

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