Beyond one size fits all. A tiered model for digital preservation

Umar Qasim, Sharon Farnel, John Huck - University of Alberta Libraries

Introduction

Long-term preservation of repositories and their contents is a challenging task due to the pace of changes in the field of technology. Technological obsolescence is a well known phenomenon and organizations require enormous amounts of resources, both human and financial, to deal with this challenge. This issue becomes even more challenging for memory institutions which are dealing with a wide range of digital resources. A resource can be very large in size, e.g., a single file in terabytes, or it can be a dataset with thousands of small files that can only be accessed with a particular hardware and software platform. Given this situation, common strategies used for preservation, such as emulation, normalization, and migration, may become very expensive to apply across the board.

The size, number and variety of digital resources under the stewardship of memory institutions pose human, technical and financial challenges with respect to their long-term preservation. Institutions cannot develop or maintain a sustainable digital preservation program alone, and so partnering with others to share in the responsibility is an effective approach. Stewards, curators and technical experts specify the value of a resource and whether or not it should be preserved locally, in a shared arrangement with others, or by a trusted third party. And while all resources selected for local preservation receive some sort of preservation treatment, attempting to carry out the same level of preservation actions on all of these resources is unrealistic and creates unacceptable levels of risk.

In this paper we present a tiered model for preserving digital content at memory institutions. Institutions can use this model to separate digital resources of enduring value that require rigorous preservation actions from those that require only minimal preservation operations and are intended to be preserved for a short period of time. This model is built on an assessment that considers three factors: resource type, archival responsibility, and level of projected preservability. We will first describe the model and then outline a local implementation.

Proposed Model

Digital preservation is a set of processes and activities to ensure long term access to digital information resources. Preserving digital content for long term access is a challenging task, and tremendous opportunities can be realized with effective strategies and planning. A range of digital preservation strategies is possible (not limited to: normalisation, migration on request and emulation) to ensure long-term access to a digital source. In addition, with techniques like media refresh and media transfer, longevity problems can be addressed with continuous efforts. All of these and many more strategies are possible to implement but require resources from organizations to be put aside for the implementation.

Digital resources at memory institutions do not require the same strategies for every single object. In some cases, resources might only need to be preserved for a short time period of time and a bit-level preservation model could be good enough for such resources. In other cases, short to medium term preservation is sought for certain types of resources, and long term preservation might only be required for specific resources. The tiered preservation model we present in this paper takes into consideration this requirement for differentiated preservation strategies. Evaluation is based on three factors: type of resource, archival responsibility, and

projected preservability, as detailed below.

Type of resource

The first evaluation factor, type of resource, considers the nature of the resource from a variety of perspectives, and bears similarities to acquisition or digitization selection policies. In fact, preservation selection criteria sit on the foundation of acquisition and digitization selection policies(1). This is especially true when an institution is primarily acquiring digital resources(2,3). However, other factors also merit consideration. An institution will wish to safeguard the investment it has already made in a resource(4,5). Institutions are often stewards of digital resources acquired or created through diverse means, beyond local digitization, and that range must be taken into account(6). When institutions hold unique material of enduring value, they have a special relationship to that material, as it unlikely to be preserved elsewhere(7,8).

This model proposes five resource types, which may be referenced by the core characteristic of each. Given in order from greatest to least priority, they are: *excellence, mission, ownership, investment,* and *mandate*. These characteristics are not mutually exclusive; in fact, each type implicitly includes the characteristics of the types lower than itself. For instance, a resource that relates to *excellence* is likely to reflect a core aspect of *mission*, involve *ownership* or *investment*, and therefore include an implicit *mandate* to preserve. At the bottom of the scale, a resource for which there is only a *mandate* to preserve will not be a resource of *excellence*, and will not have originated from core *mission* activities, nor from other activities that would entail *ownership* or *investment*; rather, the *mandate* to preserve would either come from an explicit decision or be assigned from a higher authority. It may therefore be said that the model prioritizes resources where a greater number of factors occur together.

The first type of resource is *Collections of Strength*, and relates to **excellence**. These are resources that have been designated part of collections of strength within the institution according to specific policies and criteria. They are promoted at a strategic level and reflect the identity and reputation of the institution. They are the result of a significant investment in time and money, and their content is significant and unique. They may be flagship digitization projects based on special collections holdings or the research focus of the parent institution.

The second type of resource is *Locally created, born digital resources*, and relates to **mission**. These are resources that have received significant investment because they represent unique content created in the context of the parent institution's core mission of research and teaching activities, and which would not necessarily be preserved elsewhere. An example would be a campus institutional repository.

The third type of resource is *Other locally digitized or purchased resources*, and relates to **ownership**. These are resources that the institution has digitized or has had digitized, and therefore owns, but which are not necessarily unique holdings or closely related to core mission. Digitization may have been a result of convenient opportunity. Retrospective scanning of microfilm series or newspapers are examples.

The fourth type of resource is *Licensed resources with perpetual access rights*, and relates to **investment**. These are resources that the institution has invested funds in to ensure perpetual access, but which it does not own or bear exclusive responsibility for. They may be key resources that are heavily used or critical for local users.

The fifth type of resource is *Externally created, born digital resources that the institution or*

parent institution is mandated to preserve, and relates to a **mandate**. These are resources that the institution has assumed stewardship of, though they were created elsewhere. Responsibility to preserve these resources may be the result of strategic decisions made by the institution or its parent organization. An example is at-risk digital resources that originated in the local community.

Archival responsibility

The number and types of resources that are either born digital or digitized is vast and continues to grow at an increasing rate. For this reason, memory institutions have for some time understood that no single organization can be responsible for preserving them all, nor can, or should, any memory institution preserve its own digital content without engaging in collaborations and partnerships(9,10,11,12).

In order to ensure effective, efficient and sustainable digital preservation programs, memory institutions become involved in preservation activities locally as well as in collaboration with other trusted partners. Decisions are made as to whether digital resources selected for preservation should be preserved locally, preserved in collaboration with other organization/s or entrusted to a third party organization.

The first category of archival responsibility is *sole*, which indicates that the resource is being preserved only by the institution itself. An example may be locally digitized content. The second category of archival responsibility is *shared*, which indicates that an institution is engaged in a collaborative preservation effort. An example might be Open Journal System content preserved as part of a LOCKSS network. The third category of archival responsibility is *third-party* responsibility, which indicates that an institution has determined that a third party is more suitable for ensuring the long term accessibility of a digital resource, and so has outsourced preservation responsibilities. An example might be partner resources digitized and available through the Internet Archive.

Projected preservability

More and more resources at memory institutions are becoming available in digital formats, but the long term accessibility to these resources is questionable. A digital resource or a file format is typically accessible and usable through a particular software and hardware. Technological changes have made a number of resources from the past inaccessible and unusable due to the lack of support for these resourcer on newer software/hardware platforms. Preserving resources for long term access requires careful decisions about the selection of file formats for preservation. Researchers and practitioners have identified a number of factors that can help to project the preservability of a file format - the Projected Preservability construct.

Projected preservability is a measure to determine the likelihood that a digital resource will be accessible and usable in the long run. Resources at a higher level of projected preservability indicate a higher degree of confidence in providing preservation commitments and are more likely to be accessible in the future. Projected preservability is measured through five different determinants, i.e. adoption, openness, transparency, stability and interoperability.

Adoption: Adoption is the extent to which a file format has been widely adopted and formally selected for preservation by memory institutions(13). This information is captured from other memory institutions' published resources when their local registry of file formats is publicly available. With this information, any newly ingested file format is assessed for the level of adoption according to the following scale: Low adoption means no one else is using this file

format for preservation, medium adoption is if less than 50% of the recorded institutions are recommending this file format for preservation and high means 50% or more of the recorded institutions are recommending this file format for preservation.

Openness: Openness is the extent to which a file format specification is in the public domain(14,15). An open file format has a published specification for encoding information, usually maintained by a standards organization, and can be used and implemented by anyone. Open file formats are expected to have less chance of being locked in by a specific technology and/or vendor than proprietary formats. Since the specifications are known and open, other institutions are likely to implement the same solution adhering to the same standard. Hence, openness offers better protection of the digital files against obsolescence of their applications. Proprietary file formats are considered at a low level of openness, whereas Non-proprietary file formats are considered at a medium level and non-proprietary and standardized file formats are considered at a high level of openness.

Transparency: Transparency is the extent to which the contents of a file are open to the direct analysis using basic tools such as, human readable text editors(13). Additionally, audio/video file formats concealed with compression and wrappers are less transparent and prone to higher preservation complexities. Both of these characteristics, human readability and compression, indicate how complicated a file format can be to decipher. If a lot of effort has to be put into deciphering a format, and with the chance it will not completely be understood, the format can represent a danger to digital preservation and long-term accessibility. Textual file formats which use simple and direct representation will be easier to migrate to new formats and are preservation friendly. The level of transparency is measured as follows: Compressed and/or non readable file format (where applicable) are at a low level of transparency, Lossless compressed and/or human readable file format(where applicable) are considered at a medium level whereas Uncompressed and/or human readable file format (where applicable) are considered at a high level of transparency.

Stability: Stability of a file format is determined by the format's backward compatibility and its frequency of releases(16). A file format is backward compatible if it provides all of the functionality of a previous version of the format. Frequency of version/extension releases is another indicator of the stability of a file format. A format with more than one release in the last five years is less stable than a format with one or fewer releases in the same period. The level of stability is an indication that the development of the format follows a managed release cycle. Resources which are not backward compatible and have high number of version releases have a low stability level, whereas resources which are backward compatible or have low number of version releases are considered at a medium level of stability and resources which are both backward compatible and have low number of version releases are highly stable.

Interoperability: Interoperability is the ability of a file format to be accessible on multiple hardware and software platforms(13). Formats that are supported by a wide range of software or hardware are highly desirable in many situations. This feature also tends to support the long-term sustainability of data by facilitating the possibility of migration of the data from one technical environment to another. Following is the assessment criteria for interoperability: Platform dependent resources are at low level of interoperability, software interoperable file formats are at a medium level whereas highly interoperable file formats are both software and hardware interoperable.

Implementing the Tiered Model

Once digital resources have been assessed and ranked based on the criteria of type of resource, archival responsibility and projected preservability, organizations can then bundle their preservation strategies based on the preservation level of a resource. Hence for a resource which only requires bit-level preservation, very limited and selective preservation strategies can work effectively. On the other hand, a resource selected for long term preservation needs all possible preservation strategies to be implemented.

There is no single agreed upon most appropriate number of levels of preservation; the literature contains examples of two(17), three(18), and four(19), to list a few. Organizations will determine the most appropriate number of levels based on their particular context. At the University of Alberta Libraries we have resources that we intend to preserve over the long term as well as others that we intend to preserve only over the short or medium term so we have chosen to bundle our preservation strategies into three levels: gold, silver and bronze. Digital resources at the gold level are subject to more rigorous preservation actions than those at the silver or bronze level.

Gold Level Preservation: Resources preserved at this level are subject to a rich set of preservation actions for long-term accessibility. Upon ingest, a resource will go through virus checking, fixity checking, file validation, format normalization and archival packaging processes. Gold level resources are archived with *full metadata* to capture information about the resource, provenance, authenticity, preservation activity, technical environment and rights. To prevent a loss of access to files due to file format obsolescence, all resources at Gold level are subject to a file format migration strategy, which helps to keep the content stored in formats that are readable by the current technology.

Silver Level Preservation: Silver level preservation is intended for resources that require medium to long-term preservation but are currently being preserved elsewhere and/or have lower projected preservability. Resources within this plan undergo virus checks, integrity checks, and file format normalization, and include *extended metadata*. The file format normalization process helps to store resources in UAL recommended archival file formats. Active monitoring is not part of this plan, and it also lacks any migration strategies. Multiple copies help to encounter the problem of media decay and ensure bit-level preservation.

Bronze Level Preservation: Resources preserved at this level are subject only to bit-level preservation activities. Under this level, a resource will be subject to virus checks and fixity checking. Only *core metadata* is archived along with the resource. This is a basic level of preservation which ensures the integrity of each bit over time**.** Multiple copies of a resource are retained to encounter the perils of media decay and help to replace any corrupted bits with a valid copy. This level of preservation lacks advanced preservation activities like format normalization, format migration, validation checks and preservation metadata.

Conclusion

In this paper we have proposed a tiered model for preserving digital content at memory institutions that is built on an assessment which considers three factors: resource type, archival responsibility, and level of projected preservability. This model allows institutions to assess and rank digital resources in terms of preservation needs and helps institutions to bundle preservation strategies accordingly. This model is simple to apply and flexible enough to be usable by a variety of memory institutions. Although we have described the way in which we

have implemented the model at the University of Alberta Libraries, the model does not dictate the method of implementation or the specific preservation strategies to be employed.

References

1) Ooghe, B., & Moreels, D. (2009). Analysing Selection for Digitisation: Current Practices and Common Incentives. *D-Lib Magazine 15*(9-10). http://www.dlib.org/dlib/september09/ooghe/09ooghe.html

2) UK Data Archive. (2011). Preservation policy. http://www.data-archive.ac.uk/curate/preservation-policy

3) Odum Institute Data Archive. (2011). Digital preservation policies. www.irss.unc.edu/odum/contentSubpage.jsp?nodeid=629

4) Davies, R., Ayris, P., McLeod, R., Shenton, H., & Wheatley, P. (2007). How much does it cost? The LIFE Project -- Costing Models for Digital Curation and Preservation. *Liber Quarterly: The Journal Of European Research Libraries, 17*(1-4), 233-241.

5) Bia, A., Muñoz, R., & Gómez, J. (2010). DiCoMo: the digitization cost model. *International Journal On Digital Libraries, 11*(2), 141-153.

6) Yale University Library. (2007). Yale University Library policy for the digital preservation. http://www.library.yale.edu/iac/dpc/final1.html

7) University of Utah J. WIllard Marriott Library. (2012). Digital preservation program: Digital preservation policy, Appendix B: Digital preservation decision flowchart. http://www.lib.utah.edu/collections/digital/digital-preservation.php

8) Prochaska, A. (2009). Digital special collections: the big picture. *RBM: A Journal of Rare Books, Manuscripts, & Cultural Heritage 10*(1), 13-24.

9) Skinner, K., & Schultz, M. (2010). *A Guide to Distributed Digital Preservation*. Educopia Institute. http://www.metaarchive.org/sites/metaarchive.org/files/GDDP_Educopia.pdf

10) Webb, C. (2002). Digital Preservation: A Many-Layered Thing: Experience at the National Library of Australia. In *Proceedings of The State of Digital Preservation: An International Perspective Conference*. http://www.clir.org/pubs/reports/pub107/webb.html

11) Lavoie, B., & Dempsey, L. (2004). Thirteen Ways of Looking At … Digital Preservation. *D-Lib Magazine, 10 (7/8)*. http://www.dlib.org/dlib/july04/lavoie/07lavoie.html

12) Day, M. (2008).Toward Distributed Infrastructures for Digital Preservation: The Roles of Collaboration and Trust. *The International Journal of Digital Curation* 1(3). http://www.ijdc.net/index.php/ijdc/article/view/60

13) Todd, M. (2009). File formats for preservation. DPC Technology watch series report 2009. http://www.nationalarchives.gov.uk%2Fdocuments%2Fselecting-file-formats.pdf&ei=MD8tUZbUNeOVjALW9oFo&usg=AFQjCNF7OxJVAaRvZiroR04hes2-ZU6nxQ

14) Rog, J., van Wijk, C. (2008). Evaluating File Formats for Long-term Preservation. National Library of the Netherlands; The Hague, The Netherlands http://www.kb.nl/sites/default/files/docs/KB_file_format_evaluation_method_27022008.pdf

15) Library and Archives Canada. Local Digital Format Registry(LDFR). File Format Guidelines for Preservation and Long-term Access http://www.collectionscanada.gc.ca/obj/012018/f2/012018-2200-e.pdf

16) Brown, A. (2008). Digital Preservation Guidance Note: Selecting File Formats for Long-Term Preservation. The National Archives. http://www.nationalarchives.gov.uk/documents/selecting-file-formats.pdf

17) OCUL (Ontario Council of University Libraries). (2011). *Preservation Implementation Plan*. http://www.ocul.on.ca/sites/default/files/TDRPolicy_Implementation.pdf

18) University of Minnesota Digital Conservancy. (2009). *University Digital Conservancy Preservation Policy*. http://conservancy.umn.edu/pol-preservation.jsp

19) National Digital Stewardship Alliance. (2012). *NDSA Levels of Digital Preservation: Release Candidate One*. http://blogs.loc.gov/digitalpreservation/2012/11/ndsa-levels-of-digital-preservation-release-candidate-one/