TAP: A Tiered Preservation Model for Digital Resources

Umar Qasim
University of Alberta
Alberta, Canada
umar.qasim@ualberta.ca

Sharon Farnel
University of Alberta
Alberta, Canada
sharon.farnel@ualberta.ca

John Huck
University of Alberta
Alberta, Canada
john.huck@ualberta.ca

ABSTRACT

Rapid changes in the field of technology and exponential increase in the volume of digital content makes long-term preservation of institutional resources a challenging task. Digital preservation requires a commitment for applying preservation actions along with continuous monitoring and management of the preserved resources. The expense of these actions mean that a memory institution needs to make choices about what level of preservation it can afford to provide for a resource when it makes a commitment to preserve it. This paper presents a tiered model to determine preservation levels for digital content, based on an assessment that considers three factors: type of resource, archival responsibility, and projected preservability of the resource (TAP). The paper presents a practical, flexible approach to a complex set of factors and includes examples of how the model can be applied at an academic library.

1. INTRODUCTION

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. Given this situation, common strategies used for preservation, such as emulation, normalization, and migration, may become very expensive to apply across the board.

In this paper we present a tiered assessment model for preserving digital resources at memory institutions. The TAP model assesses digital resources based on three factors: **T**ype of resource, **A**rchival responsibility, and projected **P**reservability level. 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. The model is described in the section 2 and an implementation of the model is discussed in the section 3.

2. THE TAP MODEL

Digital preservation requires a set of processes and activities to ensure long term access to digital resources but 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, whereas medium and long term preservation may only be needed for some specific resources. The tiered preservation model helps in assessming resources and is based on three factors: type of resource, archival responsibility, and projected preservability, as detailed below.

2.1 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 rest on the foundation of acquisition and digitization selection policies [14]. This is especially true when an institution is primarily acquiring digital resources [3][2]. However, other factors also merit consideration when selecting for preservation. An institution will wish to safeguard the investment it has already made in a resource [6][4]. Institutions are often stewards of digital resources acquired or created through diverse means, beyond local digitization, and that range must be taken into account [10]. When institutions hold unique material of enduring value, they have a special relationship to that material, as it unlikely to be preserved elsewhere [13][15]. In particular, we suggest a set of five resource types with different scores as shown in table 1.

The first type of resource is Collections of strength. These are resources that the institution has designated as signature collections according to internal defined criteria. These types of resources 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. These are resources that are comprised of unique content created in the context of the parent institution's core activities. They represent a significant investment by the institution, and would not necessarily be preserved elsewhere, but lack the profile or focus to be a Collection of strength. An example is a campus institutional repository.

The third type of resource is Other locally digitized or purchased resources. 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. 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

Table 1: Types of resources

Type of Resource	Score
Collection of Strength	5
Local Born Digital Resources	4
Purchased / Digitized Resources	3
Licensed Resources	2
External Resources	1

Table 2: Scoring Levels for Archival Responsibility

Archival Responsibility	Score
Sole Responsibility	2
Shared Responsibility	1
Third-party Responsibility	0

may be key resources that are heavily used by local users.

The fifth type of resource is Externally created, digital resources that are of great value and significance to the institution. These are resources that the institution has assumed stewardship of, though they originated elsewhere. Responsibility to preserve these resources may be the result of strategic decisions made by the institution or its parent organization. An example is an at-risk collection of digital resources created in the local community.

2.2 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 [17][19][8][7]. In our model, we use three types of archival settings as described below.

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. Table 2 shows scores for different categories of archival responsibility.

2.3 Projected Preservability

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. Researchers and practitioners have identified a number of factors that can help to project the preservability of a file format or in other words to determine the level of projected preservability of a resource. TAP model uses five different determinants, i.e. adoption, openness, transparency, stability and interoperability to measure the projected preservability of a resource as discussed below.

2.3.1 Adoption

Adoption is the extent to which a file format has been widely adopted and formally selected for preservation by memory institutions [18]. This information is captured from other memory institutions' published resources when their local registry of file formats is publicly available. 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.

2.3.2 Openness

Openness is the extent to which a file format specification is in the public domain [16][9]. 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.

2.3.3 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 [18]. 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.

2.3.4 Stability

Stability of a file format is determined by the format's backward compatibility and its frequency of releases [5]. 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 a high number of version releases have a low stability level, whereas resources which are backward compatible or have a low number of version releases are considered at a medium level of stability and resources which are both backward compatible and have a low number of version releases are highly stable.

2.3.5 *Interoperability*

Interoperability is the ability of a file format to be accessible on multiple hardware and software platforms [18]. 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 a low level of interoperability, software interoperable file formats are at a medium level whereas highly interoperable file formats are both software and hardware interoperable.

Scores obtained from each of these factors are aggregated to obtain an overall score. The TAP model considers an aggregated score of 90% and above as a high level of projected preservability and promote such files as recommended file formats, aggregated score of 60% to 90% as a medium level of projected preservability and consider these files as acceptable file formats, and resources below 60% are at the low level of projected preservability and are considered as bit-level file formats. Table 4 shows projected preservability of several file formats.

3. IMPLEMENTATION

Organizations may bundle their preservation strategies based on the preservation level of a resource. There is a lack of agreement on the appropriate number of levels of preservation; the literature contains examples of two [12], three [11], and four [1], to list a few. At the University of Alberta Libraries (UAL), 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 and medium term so we have chosen to use three levels of preservation: 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. The value matrix described in the next paragraph helps to determine the required preservation level of a resource.

3.1 The Value Matrix

The Value Matrix helps to determine the level of preservation for a resource and is based on the three factors mentioned above: type of resource, archival responsibility and projected preservability. Scores obtained from each of these factors are aggregated to obtain an overall score as a guideline to determine the level of preservation appropriate for a resource. UAL suggests an aggregated score of 90% and above to preserve a resource at gold level, 60% to 90% for resources at silver level, and resources below 60% for bronze level. These scores are only used as a guideline; the final decision about the level of preservation for a resource is made by the stewards, curators and technical experts at UAL. Table 3 provides an example of a value matrix.

Preservation Strategies	Gold Plan	Silver Plan	Bronze Plan
Bit Preservation	✓	✓	✓
Core Metadata	✓	✓	✓
Virus Checks	✓	✓	✓
Multiple Copies	✓	✓	✓
Integrity Checks / Checksum	✓	✓	✓
Bagit File Packaging	✓	✓	✓
Unique and Persistent Identifiers	✓	✓	
Normalization	✓	✓	
Characterization and Validation Checks	✓	✓	
Extended Metadata	✓	✓	
Migration	✓		
Full Metadata	✓		
Media Refresh	✓		

Figure 1: Preservation strategies at various levels.

3.1.1 *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.

3.1.2 *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.

3.1.3 *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 full metadata.

UAL uses varying levels of preservation strategies for its gold, silver and bronze resources as shown in Figure 1.

A UAL collection of strength example is the Western Canadiana material held in the Special Collections Library. Much

Table 3: Example of Projected Preservability

File Format	Adoption	Openness	Transparency	Stability	Interoperability	%	PP	Score
xml	2	2	2	2	2	100%	High	3
pdfa	2	2	2	2	2	100%	High	3
rtf	1	0	1	2	2	60%	Medium	2
bmp	1	0	2	0	0	30%	Low	1

Table 4: Example of a Value Matrix

Type of Resource	Archival Responsibility	Projected Preservability	%	Level
5	2	3	100%	Gold
4	2	2	80%	Silver
2	1	1	40%	Bronze

of this material is digitized to the highest possible standards (e.g. jpeg2000, METS/ALTO metadata), and is preserved locally. This type of collection will receive higher scores for all three of the factors considered and therefore could be preserved at the gold level.

UAL's institutional repository contains several collections of locally-created born digital resources, such as photographs and field notes. UAL has less control over file format specifications and so the score for projected preservability could be at acceptable level. The score for archival responsibility remains the same as preservation is local only. This type of collection could be preserved at the silver level.

UAL provides licensed access to a multitude of datasets in support of the research and teaching of faculty and students. Many of these do not fall within our collections of strength, and therefore receive a lower score in terms of type of resource. Because these datasets are created by outside individuals or organizations, file formats vary, with many falling into the 'bit-level' category; projected preservability is therefore lower. Because other institutions (likely including the creator/vendor) also archive these datasets, the score for archival responsibility is lower. As these resources receive an overall lower score hence could be preserved at the bronze level.

4. 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 then bundle preservation strategies accordingly. We believe the 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.

5. REFERENCES

[1] N. D. S. Alliance. Ndsa levels of digital preservation: Release candidate one., 2012.

[2] O. I. D. Archive. Digital preservation policies. Technical report, Odum Institute, 2011.

[3] U. D. Archive. Preservation policy, 2011.

[4] A. Bia, R. Munoz, and J. Gomez. Dicomo: the digitization cost model. *International Journal On Digital Libraries*, 11(2):141–153, 2010.

[5] A. Brown. Digital preservation guidance note: Selecting file formats for long-term preservation. the national archives, 2008.

[6] R. Davies, P. Ayris, R. Mcleod, H. Shenton, and P. Wheatley. 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, 2007.

[7] M. Day. Toward distributed infrastructures for digital preservation: The roles of collaboration and trust. *The International Journal of Digital Curation*, 1(3), 2008.

[8] B. Lavoie and L. Dempsey. Thirteen ways of looking at digital preservation. *D-Lib Magazine*, 10(7-8), 2004.

[9] Library and A. Canada. Local digital format registry(ldfr). file format guidelines for preservation and long-term access, 2013.

[10] Y. U. Library. Yale university library policy for the digital preservation. online, 2007.

[11] U. of Minnesota Digital Conservancy. University digital conservancy preservation policy, 2009.

[12] O. C. of University Libraries. Preservation implementation plan., 2011.

[13] U. of Utah J. WIllard Marriott Library. Digital preservation program: Digital preservation policy. Online, 2012.

[14] B. Ooghe and D. Moreels. Analysing selection for digitisation: Current practices and common incentives. In *D-Lib Magazine*, pages 9–10, 2009.

[15] A. Prochaska. Digital special collections: the big picture. *RBM: A Journal of Rare Books, Manuscripts, and Cultural Heritage*, 10(1):13–24, 2009.

[16] J. Rog and C. van Wijk. Evaluating file formats for long-term preservation. national library of the netherlands; the hague, the netherlands, 2008.

[17] K. Skinner and M. Schultz. A guide to distributed digital preservation. Educopia Institute, 2010.

[18] M. Todd. File formats for preservation. dpc technology watch series report 2009. Report, DPC, 2009.

[19] C. Webb. Digital preservation: A many-layered thing: Experience at the national library of australia. In *In Proceedings of The State of Digital Preservation: An International Perspective Conference*, 2002.