The Cost of Keeping It: Towards Effective Cost-Modeling for Digital Preservation at the University of Maryland

Kate Dohe
Manager, Digital Programs and Initiatives
University of Maryland Libraries
College Park, Maryland
katedohe@umd.edu

David Durden
Data Services Librarian
University of Maryland Libraries
College Park, Maryland
durden@umd.edu

ABSTRACT

With the introduction of tools like the DLF's Digitization Cost Calculator, forecasting and fundraising for digitization projects can be achieved with transparency and clarity. However, estimating and articulating the considerable long-term expenses of digital preservation lags behind. The surfeit of digital materials entering cultural heritage institutions introduce significant costs that rapidly outstrip the costs of digitization, and these costs are challenging to represent clearly at the outset of a project–either due to obscure technical details, the array of pricing options for storage and preservation systems, and the impossibility of predicting the price of "keeping it forever."

In our library, we are in the early stages of developing a cost model for digital preservation systems loosely aligned to the costs of systems and activities within the NDSA Levels of Digital Preservation framework. This work is intended to articulate the ongoing costs of desirable and essential digital curation activities to digital project stakeholders, as well as administrators–with the ultimate goal of sustainable funding for responsible digital preservation. Our "Digital Preservation Cost Calculator" has been successfully used to estimate project expenditures in preparation for grant applications and philanthropic financing requests.

We are exploring prospective features that can transition this tool from a local budgeting tool to a full-fledged digital preservation application. This paper will introduce our use case and requirements, current development challenges, and propose a prospective roadmap and options for community engagement.

CONFERENCE THEME

Mapping out sustainable digital preservation approaches and communities: What business and economic models can facilitate digital preservation feasibility?

KEYWORDS

Cost and time estimates, digital preservation, digital libraries, digital curation.

iPRES2018, September 24–27, 2018, Boston, Massachusetts USA

1 INTRODUCTION

The University of Maryland (UMD) Libraries is at a transitional moment in its digital preservation practices and initiatives. Beginning in 2014, the Libraries scaled up digitization projects significantly with the assistance of expanded fundraising capacity, including private philanthropic fundraising as well as several successful grant applications.[1] Over three years, the capacity and throughput of digital content from digitization projects jumped from gigabytes to terabyte scale, and has stabilized at approximately 30 terabytes per year. This significant uptick in scale is a direct byproduct of an institutional emphasis on reformatting at-risk audiovisual materials in the Libraries' special collections, combined with more readily available funding. In this time period, UMD also joined numerous digital preservation consortia or services, including Academic Preservation Trust (APTrust) and the Digital Preservation Network (DPN).

2 DIGITAL PRESERVATION AT UMD

UMD Libraries has had a digital preservation policy in place since 2013 [7], which was last revised in 2014 to address the proliferation of content and service options available to the institution. This policy mandates some level of digital preservation for all unique digital assets at risk of obsolescence or loss, with treatment aligned to the National Digital Stewardship Alliance (NDSA) Levels of Digital Preservation [9]. The policy does not mandate specific activities for certain content forms, and recommends developing a comprehensive digital preservation program over time.

However, the rapid growth of digital reformatting, combined with the cost-prohibitive nature of digital preservation through membership services, staffing and leadership changes, and a paucity of reliable funding, has led to a period of decision paralysis within the Libraries with regard to digital preservation activities. As a result, digital preservation activities have maintained a holding pattern with a simple workflow that has been in place for several years: assets are archived to LTO tape by the main campus IT division, with one copy stored in a campus data center and the second going to an offsite facility. This arrangement with the UMD Division of Information Technology (DIT) lacks ongoing monitoring and fixity checking, description, inventory management, or consistent description, and thus does not qualify as even a minimally viable digital preservation strategy.

Addressing this problem at the local level required a combination of policy review, workflow revisions, and ultimately justification for

[1] This information is currently unpublished in an internal report. Documentation pending

an increase in stable funding. While this work is inherently iterative, one of the initial challenges faced by digital preservation stakeholders was an absence of compelling evidence to clearly articulate the return on investment in the use of more sophisticated systems. Invoices from the DIT department responsible for archiving and backup services reflected frequently inconsistent and obscure service charges to our User Systems and Support department; rapid reorganization and leadership changes in DIT served to compound institutional communication silos regarding the unique support needs of the Libraries. Essentially, no employee within the Libraries could reliably articulate or forecast our digital archiving costs, or make assurances about the fixity and stability of our content, but we could operate under the assumption that the bills were being paid, so the solution must therefore be affordable. Changing these assumptions to advocate for a more robust and transparent solution required development of a service and platform matrix, with costs clearly delineated.

Given these environmental realities, the UMD Libraries struggled to clearly articulate the costs and benefits of a given solution or system for digital materials. The UMD Libraries stewarded a significant amount of unique digital material–approaching 100TB in 2017–and it was clear that preservation of reformatted AV materials would rapidly outstrip the costs of reformatting that content within a matter of years. The priority for strategic communication then became defining the preservation costs and required treatments at the outset of project fundraising, with buy-in from curators, development officers, and Libraries leadership.

3 BUILDING A DIGITAL PRESERVATION COST CALCULATOR

The digital preservation cost calculator at the UMD Libraries began as a complex series of spreadsheets–an origin similar to other digital preservation planning tools and cost models [6]. The original spreadsheet version of the digital preservation cost calculator tracked digital preservation costs and services used at the UMD Libraries. In its initial phase, these calculation tools focused expressly on preservation in terms of its digital storage requirements–how many terabytes of content to budget for in a given system. Given the complexities of predicting even these costs, grappling with complex issues regarding labor and service was clearly above and beyond the abilities of spreadsheet formulas. However, it should be made explicitly clear that this approach is fundamentally like claiming a library has the same expenses as a warehouse–it elides the substantial level of human expertise required to manage, maintain, and steward materials. In the interest of devising a simple minimum viable product to address the urgent and tangible issue of suitable preservation storage, these expenses were not included in the initial stages of the cost calculator, but must absolutely be accounted for in future iterations of the calculator.

3.1 Motivations and Initial Work

Initial work on expanding the calculator tool began with an environmental scan of preservation systems, which included local, vended, and consortial approaches, and included data for services the Libraries did not use. Each system or service was evaluated using the NDSA Levels of Digital Preservation and Preservica Digital

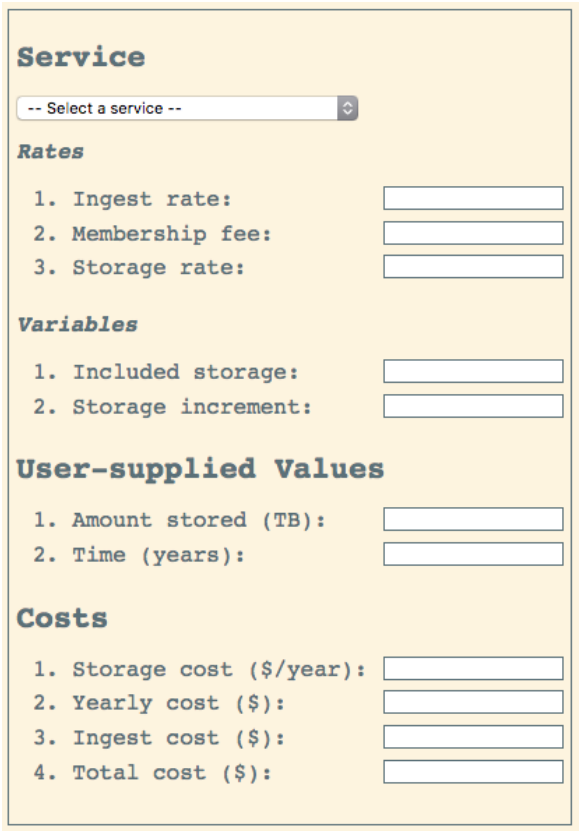

Figure 1: The digital preservation cost calculator interface.

Preservation Maturity Model [10] with the goal of contextualizing cost in relation to level of digital preservation service. In the original calculator spreadsheet, these systems were organized by common variables, such as units of data measurement and dollars, which allowed for direct cost-comparison across services. The current instantiation of the cost calculator primarily serves as a forecasting tool to assist with financial planning with regards to current growth of digital assets and for new digitization projects.

The spreadsheets used to calculate cost and levels of digital preservation of each selected service quickly became cumbersome to use. Drawing upon inspiration from DLF's Digitization Cost Calculator [5], we set about building our own web-based cost calculator. The first web version of the digital preservation cost calculator was built using static HTML and JavaScript resulting in a simple web form that would return costs calculated from three basic inputs: service provider, size of data, and length of storage term. Initial development of the calculator began as an internal Coding Group project, rather than a formal development initiative, and as such it is much closer to a proof of concept than a production-level application. The cost calculator website has a significantly improved user experience and functionality compared to the early spreadsheet. However, the current calculator does not provide a way to directly compare multiple services side-by-side on the same webpage; instead, a user would need to perform their calculations and manually record the results.

3.2 Local Implementation

Even in its infancy, the web-based cost calculator has already proven its utility for financial forecasting of digital reformatting and infrastructure projects. Its presentation of information is designed to provide prompt, accurate information for non-technical stakeholders, though the tool is exclusively used by the Digital Programs and Initiatives department. The most frequent use case for the calculator at this juncture is swiftly providing data to the Libraries' development officer, who may present prospective funders with an array of treatment options and finite time periods to fund and sustain a digital project. Thus far, the information provided by the calculator has served to communicate "where rubber meets the road" project costs to donors earlier in the conversation, and with time we are optimistic that it can be used to support endowments or other sustainable funding models.

The cost calculator has already been used to provide accurate cost forecasts in multiple grant applications and project proposals. This has proven to be essential to the development of accurate grant budgets and project costs for Libraries-driven projects that are expected to generate at least a terabyte of digital assets. As the Libraries become progressively involved as partners in enterprise-level research data management as well as electronic records curation and retention, the cost calculator may have significant business implications for other campus units, as well.

In practice, the calculator is a communication tool that serves to improve transparency and accountability around a set of services and processes that can often seem abstract or shrouded in mystery to non-technical library staff or prospective donors. Given the simplicity of the minimum viable product benchmark for the calculator, the tool does not currently incorporate the complex and unverifiable service structure of UMD's DIT archiving and backup service, but does enable Libraries staff to communicate the costs of straightforward and reliable services such as APTrust and DPN. Over time, our goal is to use the tool as a mechanism to advance collaboration with UMD's DIT and align local services with Libraries' requirements, but it currently serves to highlight the delta between options to internal stakeholders.

3.3 Future Development

Future development of the cost calculator tool focuses on these main areas: 1) enhance the data model to include a levels of digital preservation assessment, 2) add direct service comparison and budget reporting tools to the webpage, 3) implement a historical cost feature to demonstrate annual fluctuations in service rates, and 4) improve model to include staff and student labor costs for digital curation and other preservation activities.

As the cost calculator evolves, so too will our local cost model. Our long-term goal is to enable informed decision making for digital preservation in terms of cost per unit, cost over time, and cost by level of preservation. At the onset this model will include cost of storage only,, but in time we plan to include digitization costs and staff time to improve the accuracy of our forecasts. This approach to forecasting will enhance digital preservation workflows by providing the means to determine when to move digital assets from local or cloud storage (which may have a lower cost per unit of storage, but lacks more robust preservation features) to distributed node-based services (which come with a higher cost per unit of storage, but feature long-term stability and ongoing preservation activities).

Future cost models for digital preservation at the UMD Libraries should provide methods for comparing various services based on similar metrics and options. It is simple enough at the moment to compare vended services using such variables as length of time, included storage, additional storage, etc. However, value and impact for vended and consortial services will not be apparent until the cost of performing similar levels of service using internal infrastructure and staff can be demonstrated in equitable terms. Using cost and price models such as those presented by the University of California Curation Center at the California Digital Library [1], CERN's Data Preservation in High Energy Physics [2, 4], the Royal Danish Library/Danish National Archives [8], and NASA's Goddard Space Flight Center [3] will prove useful for normalizing the cost of digital preservation activities at the UMD to facilitate comparative financial planning.

3.4 Impact Beyond UMD

We intend to release our cost calculator to the the digital preservation community at-large once we have completed initial feature development and implementation. Given the rapidly shifting landscape of digital preservation initiatives, cloud-based storage options, and evolving community standards of activity, our intent is to seek partners in the practitioner community for reliable and consistent costs, as well as domain expertise regarding economic modeling for longer-term preservation treatment. The code for the cost calculator will be open source and available for anyone to reconfigure and adapt for use at their own institution.

Forecasting and planning are the primary use cases for the our cost calculator; however, we believe that building a levels of digital preservation assessment feature into the tool will have considerable value for the community. The data model for this feature would rely upon both vendor supplied variables, where available, and local variables provided by institutional IT departments. This feature couples well with forecasting and planning as materials selected for preservation can be evaluated against appropriate preservation solutions based on multiple factors. The digital preservation cost calculator will allow repository managers and decision makers to demonstrate the value of the cost of digital preservation in terms of services and standards.

4 CONCLUSION

Ultimately, "the cost of keeping" our unique materials is an open question, fraught with implicit assumptions and values. As the practitioner community at a variety of institutions transitions from "hard drives in a box in the server room" to operationalized and technological solutions, it will be essential for digital preservationists to adopt unambiguous, evidence-based approaches to communicating these costs and benefits to administrators and stakeholders that act as the gatekeepers for program funding. The cost calculator at UMD represents a single component of a comprehensive digital preservation program, and in order to unlock its full potential it must be located within a community of practice.

ACKNOWLEDGEMENTS

The authors would like to thank Peter Eichman, Joseph Koivisto, and Joshua Westgard or their ongoing contributions to the development of the Digital Preservation Cost Calculator.

REFERENCES

[1] Stephen Abrams, Patricia Cruse, and John Kunze. 2015. Total Cost of Preservation (TCP): Cost and Price Modeling for Sustainable Services. https://confluence.ucop.edu/display/Curation/Cost+Modeling?preview=/163610649/173539758/TCP-cost-price-modeling-for-sustainable-services-v2_2_2.pdf

[2] Frank Berghaus, Jakob Blomer, Germán Cancio Melia, Sünje Dallmeier Tiessen, Gerardo Ganis, Jamie Shiers, and Tibor Simko. 2016. CERN Services for Long Term Data Preservation. In *Proceedings of the 13th International Conference on Digital Preservation (iPRES16)*. iPRES, Swiss National Library, Bern Switzerland, 168–176. http://www.ipres2016.ch/frontend/organizers/media/iPRES2016/_PDF/IPR16.Proceedings_4_Web_Broschuere_Link.pdf

[3] NASA: Goddard Space Flight Center. 2015. Cost Estimation Toolkit. https://opensource.gsfc.nasa.gov/projects/CET/index.php#software

[4] DPHEP Collaboration. 2015. *Status Report of the DPHEP Collaboration: A Global Effort for Sustainable Data Preservation in High Energy Physics.* Technical Report DPHEP-2015-001. CERN, Geneva Switzerland.

[5] Digital Library Federation. 2016. Digitization Cost Calculator. http://dashboard.diglib.org/

[6] Ulla Bøgvad Kejser, Anders Bo Nielsen, and Alex Thirifays. 2011. Cost Model for Digital Preservation: Cost of Digital Migration. *The International Journal of Digital Curation* 6, 1 (2011), 255–267. https://doi.org/10.2218/ijdc.v6i1.186

[7] Jennie Levine Knies, Robin C. Pike, Joanne Archer, Vincent J. Novara, and Carla Montori. 2014. University of Maryland Libraries: Digital Preservation Policy. http://hdl.handle.net/1903/14745

[8] Royal Danish Library and Danish National Archives. 2012. Cost Model for Digital Preservation. http://www.costmodelfordigitalpreservation.dk/contact

[9] Megan Phillips, Jefferson Bailey, Andrea Goethals, and Trevor Owens. 2013. The NDSA Levels of Digital Preservation: An Explanation and Uses. In *Proceedings of the Archiving Conference (Archiving 2013)*. IS&T, Society for Imaging Science and Technology, Springfield, VA, 216–222. http://ndsa.org/documents/NDSA_Levels_Archiving_2013.pdf

[10] Preservica. 2014. Digital Preservation Maturity Model. https://preservica.com/uploads/resources/Preservica-White-Paper-Maturity-Model-2014_NEW.pdf