BUILDING SUSTAINABLE DIGITAL PRESERVATION

A Typology of Digital Collections

Helen Hockx-Yu
University of Notre Dame
USA
hyu3@nd.edu

Donald Brower
University of Notre Dame
USA
dbrower@nd.edu
https://orcid.org/0000-0001-7347-546X

Abstract – The digital preservation function in an academic research library is chiefly concerned with the ongoing access to its digital collections. This paper describes the work in developing a typology of digital collections at Hesburgh Libraries, University of Notre Dame, which serves as a framework for planning and building a sustainable Digital Preservation program.

Keywords – Libraries, Digital Collections, Digital Preservation, Typology, Framework

Conference Topics – Designing and Delivering Sustainable Digital Preservation

I. INTRODUCTION

The Hesburgh Library at the University of Notre Dame has a digitization program and an institutional repository with bit-level preservation commitment. [1] Yet, some of library's digital collections fall outside the scope of these. We desired a holistic digital preservation strategy mindful of our complete holdings.

The Digital Collections Typology Project took place between April and December 2018. The project developed a Typology of Digital Collections that can be used as a framework to plan and guide digital preservation.

The nature of libraries' collections has evolved over the last twenty years or so, from physically owned and locally stored collections to what Dempsey coined "facilitated collections", which include a broad range of local, external and collaborative resources organized around user needs. [2] As many academic and research libraries, Collections of the Hesburgh Library spread across the spectrum as illustrated in the diagram below. Each collection in the spectrum may require a different approach to long term preservation.

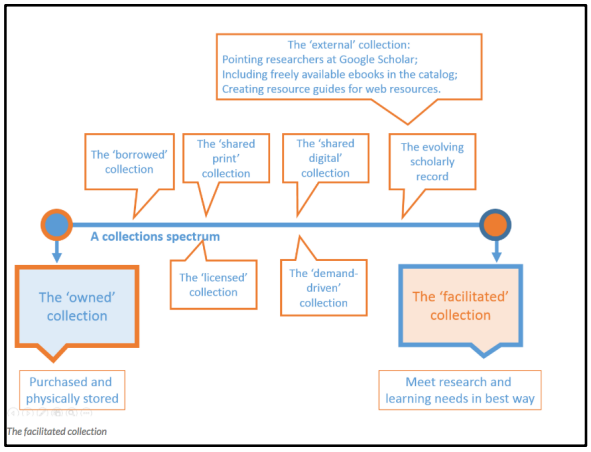

Figure 1 Collections Spectrum. [3]

Hesburgh Libraries' Digital Collections can be in any of the formats below:

- Entire collections in digital format
- Items in digital format that are a part of a collection which also has analog items
- Stand-alone items in digital format
- Digital surrogates of analog collection items
- Digitally encoded content on physical media

The scope of the project also extends to physical items that require digitization to preserve access, for example VHS tape and audio cassettes (which are both analog encoded content on physical media).

II. RELEVANT WORKS

While Dempsey's Collections Spectrum provided a useful context for understanding the evolution of collections, we needed something more detailed to approach digital preservation holistically based on common characteristics.

The field of digital preservation and curation has developed a number of additional frameworks that provide useful context:

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

DOI: 10.1145/nnnnnnn.nnnnnnn

iPRES 2019

- The DCC Lifecycle, an iterative, high-level overview of the stages required for successful curation and preservation of data. [4] While our work focuses on data (collections) and preservation action, the DCC model provides the "big picture" and reminds us that actions are required across all stages of the lifecycle, and cannot only the stage of "preservation action."
- The generic, appraisal/selection framework for digital curation by Jinfang Niu. [5]. Although our work mainly concerns items that have already been appraised, selected and ingested as part of our collections, the framework offers criteria (i.e. mission alignment, value and cost) that we can apply in future to decide which type of collection items should be prioritized for preservation.
- The SPOT model for risk assessment. [6] identifies threats to digital preservation. This is relevant to our work and the typology of preservation threat can be usefully mapped to our typology of digital collections to help with prioritization.
- The Digital Content Review Process developed by the Digital Preservation Management Workshop [7]. Our work shares the same goals but is much less comprehensive as it is intended as a starting point to understand the high-level, broad scope of digital preservation at the Hesburgh Library. We expect to include many of the useful details and even adopt parts the framework when carrying out the next level of assessment and prioritization.

We also considered inventory management, a common method that libraries use to assess, examine and track the condition of their collections. This would be a valid starting point for digital preservation, too, but over a certain size inventories become hard to work with. Moreover, we have a few special collections where just performing an inventory would be a significant undertaking. We therefore deferred inventory to the next stage and decided to start with something more general.

III. THE TYPOLOGY

A typology is a general classification of items that provides a structure for understanding items by highlighting the properties either shared or not shared between them. A typology can be contrasted with an inventory where for any sizable number of items an inventory becomes unwieldy and is not useful for understanding the items as a whole. We see a typology as being in the middle ground of detail between the conceptual Collections Spectrum and an inventory.

Our focus was preservation, so the typology was organized to assist with the preservation (and non-preservation) of digital content. The attributes were chosen to be those that are significant to deciding on preservation actions.

The types are each defined by unique attributes. Some attributes relate to the control and access of the types, such as whether an item is managed by the Library or vendors; whether an item is a physical object, e.g. a VHS tape; or whether an item is digitized, i.e. a digital surrogate of a physical item. In case a digital asset is the result of digitization then we would need to maintain the knowledge of the link between the physical item and the digital surrogate.

Our typology identified three broad types for the Library's digital collections, listed below. These are high-level and likely to be applicable to other institutions. More detailed types are expected to vary between institutions and lead to different preservation approaches, depending on resources and constraints.

Vended collections are resources that the Library is given permission to use for a limited purpose or timeframe. Access to vended collections by patrons generally takes place online at platforms provided by copyright holders or licensors. Examples are electronic journals, books and databases. Unlike physical purchases, vended content may "disappear" from the Library's collections once the license is contractually terminated.

Library-managed collections are resources the Library chooses to or is obliged to exercise stewardship over. This content is mostly kept on university operated equipment, but occasionally external services are used to host content and provide access. This category contains two subtypes: digital surrogates (or digitized) and born-digital. Digital surrogates are the electronic captures of physical items in our collections. "Born-digital" are items that do not have a corresponding physical item in the collection.

"Physical media" is a "related type", which includes physical items that have digital information or that we expect to be digitized, e.g. computer disks, CDs, VHS tapes, and audio recordings.

In-house digital creation are resources produced by the Library for various purposes that are not collection items, including marketing and instructional material, materials related to contests and student awards, source code, websites, blogs, access copies, LibGuides, etc.

IV. METHODOLOGY

To gain an understanding on the digital material the Library currently deals with, the project team interviewed 37 faculty and staff members. The majority were subject librarians and staff who work

directly with digital content. The interviews were free-form, but there was a written prompt to help guide it.

The questions were over the kinds of digital content interviewees either handled, purchased, or had in their collections. If interviewees knew any item counts or storage sizes, we recorded that as well. The answers pointed not only to large collections hidden-in-plain-sight, such as nearly 8000 VHS tapes, and more than 15k DVDs, but also to unique one-offs, and comments on processes and policies related to digital content. The interviewees were often unable to provide exact counts or indicate where to obtain the numbers. This confirmed our assumption that more detailed inventorization is required, allowing us to plan and prioritize the collection items that are in scope for preservation.

From the interview notes, the team then compiled a list of every content or file type mentioned. Since the content names were transcribed directly from the interview notes, there were many variants of the same name, so the content names were normalized and then grouped into similar kinds. The team reviewed the resulting list, extracted and organized attributes based on how they aligned with preservation decisions, and developed recommended preservation actions for each type.

V. Using the Typology

The main application of the typology is the definition of the scope of digital preservation for the Hesburgh Library, followed by a broad approach for each type, that either builds on existing workflows or points to new work. By having a comprehensive typology comprehensive, we could make intentional preservation decisions for all of our content.

For example, it became clear that not every type or all instances of a type need to or can be preserved. The Library's digital preservation program should focus on the long term stewardship of collections managed by the library, as well as digital surrogates produced from physical media. Digital surrogates produced for access, for example images from a book scanned for use on a poster, are outside the scope of long-term preservation.

Similarly, instances of In-house digital creation are generally not intended for long term preservation. They may have significant short term value or be required to support the Library's operations, so need to be kept safely, as with any other operational data. Some instances within this type may become a part of the **Library-managed collections**, for example Doctoral dissertation or Master's thesis. Formal process should be followed to accept the material which conceptually moves the content from one type to another. Some instances may be considered University Records, and should be dealt with in accordance with the University's Records Management and Archives Policy.

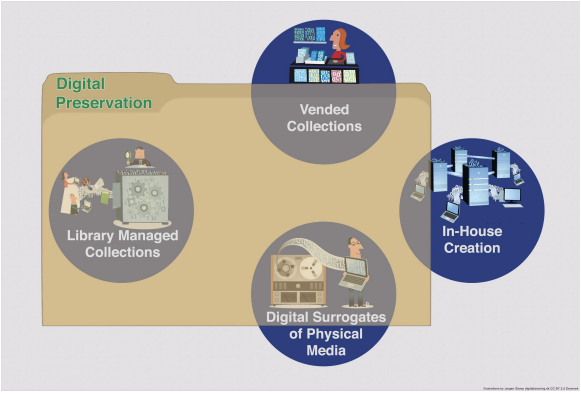

Figure 2 Typology and Scope of Digital Preservation.

For Vended collections, there is relatively little that the Library can do to preserve the content pro-actively. One area where the Library can exercise some influence is in the license negotiation process. We therefore recommended to continue the current practice of asking explicit questions about continued access, the possibility for archiving, and the content providers' preservation commitment. We also expressed a word of caution with regard to local archival copies, which do not guarantee perpetual access. They become the Library's responsibility and should be treated as part of Library-managed collections.

Library-managed collections are the focus of digital preservation. We recommended short term projects, as well as longer term, more strategic work:

1. Develop guidelines and tools to help curators determine where to store and expose digital content - this is essentially an communications effort, intended to make sure that the available storage and access options are well understood.

2. Organize immediate effort to mitigate the preservation risks related to obsolete physical media items in Rare Books and Special Collections, digitizing and moving these to stable and managed storage.

3. Utilize existing infrastructure including the institutional repository and campus archival storage service as (interim) preservation storage for Library-managed digital collections. The goal is to accommodate as many of our collection items as possible, especially those that for various reasons are not benefiting the current digital preservation care in place.

4. Devise a process (supported by the necessary tools and systems) that systematically identifies Library-managed collection, and maintains an ongoing overview. Such overview of our holdings is essential to digital preservation: you simply cannot protect your data if you don't know your data.

V. Discussion and Conclusion

Overall, our typology project has been very useful. Constructing the typology gave us an opportunity to consider our collections holistically and record the various places content had been stashed over time.

The typology has created a common terminology between different groups in the Library facilitating discussion and planning. The typology also raised awareness of the specific characteristics leading to the different approaches required to preserve our digital collections.

A caveat is that our typology provides a content-centric view. The broad approaches recommended based on the typology are action-oriented and should not be seen as a digital preservation strategy. Our work will certainly inform the development of such a strategy, but it does not take into account the organizational component, addressing issues such as staffing and training needs, and financial requirements. In this sense, the typology is only a starting point, eliminating what is irrelevant at a high-level and pointing us to a way forward.

While our focus was the big picture, interviewees did mention tools and systems that they used to handle digital collections. In hindsight we should have been more systematic with collecting this data and aggregating it so that we would have a more detailed list of place to examine for follow-up inventories and analyses.

The typology is a high-level framework that helped us describe and categories the Library's collections with enough details to understand the broad patterns yet prevented us from not seeing the forest for trees, too early in the process. These patterns were used to define the scope of broad approaches for digital preservation, and allowed us to arrive at the priority areas rapidly.

Acknowledgement

The authors are grateful to Patrick Rader for creating Figure 2 and to all other project members and interviewees for their contributions.

References

[1] "CurateND". https://curate.nd.edu/.

[2] L. Dempsey. "Library collections in the life of the user: two directions." https://www.liberquarterly.eu/articles/10.18352/lq.10170/.

[3] Ibid.

[4] "DCC Curation Lifecycle". http://www.dcc.ac.uk/resources/curation-lifecycle-model.

[5] J. Liu. "Appraisal and Selection for Digital Curation." *International Journal of Digital Curation*. VOL 9 NO 2 (2014). https://doi.org/10.2218/ijdc.v9i2.272.

[6] S. Vermaaten, B. Lavoie and P. Caplan. "Identifying Threats to Successful Digital Preservation: the SPOT Model for Risk Assessment." *D-Lib Magazine*. Volume 18, Number 9/10 (2012). http://mirror.dlib.org/dlib/september12/vermaaten/09vermaaten.html.

[7] Digital Preservation Management. "Digital Content Review: Process and Results." https://dpworkshop.org/workshops/management-tools/process-results.