Long Term Financial Costs of Maintaining a Digital Collection

Wilhelmina Randtke
Florida Virtual Campus

Introduction

The point of this article is to give an idea of what costs go with long-term maintenance of digital collections. Libraries have moved most access to materials from print to electronic. Understanding maintenance costs, as opposed to initial costs to build a digital library, helps to understand long-term implications of the shift. This article is heavily academic oriented, because academic and state libraries/archives have traditionally kept long-term historic materials.

With print materials, storage costs are visceral. We can see and feel linear feet on the shelf. Rent and air conditioning are things we pay for ourselves in daily life and which must be paid on a storage facility. However, digital materials conceal costs. Home photos taken by phone may be the closest we come in daily life to understanding issues involved in long-term maintenance of digital objects.

Shift from Print to Electronic

The past 20 years—20 years!—has been the same story repeatedly in library archives. Electronic material becomes available. Libraries rent access to it but cannot own paid electronic packages. Libraries catalog, but they do not harvest open access electronic materials. Now the library has it online, so purge the books and have a nice physical space, even perhaps a nice physical space for a very different department than the library.

Acquisitions Decisions are a World Apart from Digital Preservation

Meanwhile, just-in-time purchasing has been the replacement for just-in-case hoarding of materials. The library profession has touted the logic of this with the catch phrase "patron driven acquisitions," or PDA. The Internet has made it easier to find and buy out of print materials. You can buy an electronic resource at a faculty member's request and have nearly immediate delivery. There is an idea hanging in the background that electronic materials never go out of print and that future price fluctuations will occur regardless of what a library does this year.

Proof that long-term preservation is divorced from the acquisitions department is right there in the buzzwords. PDA is also short for "personal digital archiving," a huge catch phrase in the world of digital preservation. There are many conference presentations on "PDA," with two very different meanings in two very different library worlds. A reused acronym is perfectly clear and unambiguous only as long as the two fields remain separate. Whenever you see the acronym "PDA" in use, this means that acquisitions and digital preservation have not yet met.

Here, I am using PDA to mean patron driven acquisitions, the idea of buying materials on demand rather than buying and archiving "in case" someone needs it in the unknown future.

Just-in-Time Purchasing: The Assumption that Materials Can Be Purchased

Just-in-time purchasing builds on the assumption that materials, once created, may always be acquired and purchased.

For current info, this probably works. The market can decide each year what exists. The publisher's choice is always: Create it or not? The library's choice is always: Buy it or not? The desire to buy happens in the same year that the availability or unavailability of the resource is apparent, so if something important disappears, there is a chance to voice concerns, know why, and even make a change.

For older info, the market cannot decide each year what exists. The choices are: Keep it available or not? and How much does it cost this year? However, once the decision is to discard it, there is not a way to go back and undo that. You need a copy to make a copy. If no one keeps a copy, then it is gone. Once an individual library decides not to retain **ownership** that decision is out of the library's control. (Remember, there is a large amount of content that libraries rent rather than own.) Once everyone decides not to retain, that decision is permanent. Awareness of that decision may come long afterwards and long after there is time to voice concern.

Historically, libraries used to keep the backlog of older material. "Publishing," at heart, means making the resource available. It used to be that a commercial publisher/printer would make something available short term. Libraries and consumers

would purchase. Then, once material was out of print, libraries made the resource available long term. Libraries for a long time had a very concrete role in acting as the long-term archive of older materials, making those materials available for centuries to come.

Outdated material matters for law practice. Professors studying legal history are one thing, but the Kafkaesque nature of law throws many very relevant curves. If you have someone with a prior conviction, but court records do not state the degree of the conviction, then you need the older statute to get the degree of the conviction. That matters for priors on a subsequent conviction and for background checks of employment. Older material matters when a discovery rule starts the statute of limitations running. Think of pollution—once there was no regulation of pollution, then more and more materials have become regulated. Statutes of limitations run on the discovery rule. If someone this year discovers pollution in their neighborhood and the dumping was long ago, then the statute of limitations starts now, but the legality of the pollution depends on older laws in effect on the date of the pollution.

The premise of PDA is the idea that nothing is really ever unavailable.

However, the PDA model of purchasing puts long-term preservation of resources with vendors. Historically, the publisher/vendor never could make the decision not to retain, but now can make that decision.

Maintenance Costs and Vendors

Online delivery of material is relatively new. The release of the first web browser was in 1993, less than 25 years ago. There was an initial phase of adoption connecting enough people creating an audience for Internet material. Most early digital library buzz was about digitization or producing newly available digital collections. It is far more recent that focus has shifted from launch to long-term preservation.

Understanding costs in keeping digital material available may help to understand the issues here.

Vendors exist for profit. Profit is generally part of corporate by-laws. (Yes, many vendors are nonprofit, but copyright concerns mean each resource stays with the owner until the sale of those rights.)

Think about cancellation decisions by libraries. (Electronic "acquisitions" decisions often instead tend to be electronic "cancellation" decisions.) Analytics play in. It makes sense for a library to cancel less-used content. Similarly, for a publisher, it makes sense to cancel and no longer offer for sale less-used (i.e. less profitable) content.

Maintenance Costs in Running a Digital Collection: Computer Programming, Systems, and Testing

This section is about costs to operate the platform, as opposed to maintenance costs of the content. In general, a software platform has an economy of scale, where maintenance costs spread over content and more content does not have a big impact on costs.

Costs of Not Maintaining Code

All digital library platforms are written in computer code. That code lives on a server. Over time, deprecated versions of software run on servers, and, along with that, servers must be upgraded. The server configuration allows the code to run, and a newer version may not be backwards compatible with regard to what the code allows. In fact, PHP, which is a very popular language for web interfaces, is not backwards compatible. Code that can run on a PHP 5.4 server probably needs some lines rewritten to be able to run on a PHP 5.6 server. You can theoretically keep running a server with an outdated PHP or MySQL, and so keep on going with no changes to the platform, but running older server configurations opens up maintenance issues in terms of security and eventual difficulties in sourcing hardware. For example, if you purchase a physical server, it may have some limitations. If you purchased a physical server long ago and need a part now, that may be harder to source and more expensive than a new server. If you rent a managed server, vendors will have rolling cut offs for what versions of PHP, Structured Query Language (SQL), Java, etc. the vendor will support for you. Eventually, running a very old server architecture may involve significant surcharges. Troubleshooting and tech support may become unavailable. Going out of date costs money. Long term, not rewriting code is expensive.

Going out of date may also not be feasible as the community around you moves on. Users have expectations about how a site should look, and those expectations change as the entire web changes. Having a site that looks a bit outdated can be off-putting. Having a site that works in an outdated way can be a deal breaker and may prevent users from getting to material. For example, consider sharing out metadata records. Fifteen years ago, Z39.50 was looking to be a leading standard for sharing metadata records. Today, Z39.50 is still in use but out of favor. A fifteen-year-old platform designed to share records exclusively in Z39.50 likely could have participated in federated search projects more so in the past than it could today.

That is because harvesters built around Z39.50 are not the norm now, and even for harvesters that support them, the skill and experience to troubleshoot or test is scarce. If you are running an archive, and the world has moved on, it is out of your control to fix the problem without changing your archive.

Costs of Maintaining Code

Keeping up-to-date costs because you have to rewrite code—there is no getting around that.

In open source, the wider community is keeping the code up, but once you use it, you are part of that wider community. You can contribute code. Adding code to a project requires a huge time commitment in order to understand the wider project fully and to make edits in an appropriate or cohesive way. It cannot be an afterthought and likely would require a dedicated position within the organization. That is a significant cost. You can contribute software testing, which is actually a huge part of debugging and huge time commitment. (Hint, hint, law libraries: this is a skill that might be easier to staff, and software testing is a skill that dovetails well with legal research. It is still a deep time commitment in order to notice the significant details that matter, but that detail oriented work to really understand how the software works for a researcher overlaps with reference and teaching activities.) Alternatively, you can contribute funds to the board managing the open source software, by a membership fee or by pledging to specific projects where multiple institutions each pledge a small amount and once achieving the goal, there is the development funding of a new feature.

In proprietary, the vendor is keeping the code up. The vendor uses salaried staff and contractors to keep software up-to-date.

There is quite a bit of overhead in keeping a whole platform running over time. Libraries can somewhat see this when getting prices for institutional repositories or digital publishing platforms. Some of the pricing is training and support, but not all of it.

Maintenance Costs in Running a Digital Collection: Metadata

Metadata maintenance comes up in two areas: migration to a new platform and keeping up with changes in the wider world. Often, this cost scales up with the amount of content. More content leads to higher costs.

Migration and Metadata

Migration is a necessary part of long-term management of digital collections. Digital content lives in software. Eventually, the software will change. Then all that content has to move over without getting all messed up. Library catalogs, which technical services is familiar with, have been in Machine-Readable-Cataloging standards (MARC) for decades, and it is glaringly obvious that MARC is the dominant standard.

However, in digital material, it is not like this. There are several widely used metadata standards for digital collections: Dublin Core, Metadata Object Description Schema (MODS), qualified Dublin Core, and others. All those standards are common, and all are very different, and it is possible to build a digital library platform around any of them. On top of that, it is also common for a platform to keep indexing information in something that is totally and completely specific to the software. For example, most metadata schemas do not have a place to represent statistical information on downloads, but some harvesters require statistics, so statistics tend to be pulled from elsewhere in the software. Also, think about a collection that an object is in. Collections can often be represented in metadata, but almost all digital library software stores, updates, and pulls collection information from something other than the metadata record. In addition, it is difficult to represent serials-specific information, such as volume, issue, etc., in any library metadata schema, so it is common for platforms oriented around journals to handle that information in ways specific to the software.

What this means is that if you change software, then you have to look closely at the metadata for each platform. Digital library platforms do not necessarily store metadata in a schema. They almost certainly will output metadata in a schema, mapped from something internal. Therefore, the material may behave as desired in the software, and you can export records in Dublin Core and in MARC, but that export may have concatenated two similar fields. The fields are stored separately, because one of them is not kept in a standard schema; the distinction is important, but to be able to export in a standard, they must be concatenated, and the distinction is lost. When you move to a new software, you will have to do a deep dive into each software and what metadata each is using to make sure you can keep certain distinctions. Even if both platforms are using a standard, there is probably some important metadata used for the search and interface not stored in that metadata schema. This metadata search and interface runs in a customized way specific to the platform. Staff time to make that deep dive is a significant cost, along with a programmer's time to implement recommendations.

Focus and thought to migrate is often collection-specific, because the use of each collection of materials is slightly different. Think about researching statutory history in all 50 states. There is a general process of how you do it and what matters, but the details vary from state to state, and the details vary within a state over time.

There are more layers than that. Think about looking at pages of a book online. Each page is a picture. While you experience a book, the software must organize and present hundreds of individual pages. There is a standard for keeping pages in order, Metadata Encoding and Transmission Standard (METS) using the structMap elements in METS. METS is an Extensible Markup Language (XML) based standard. There is no quick query of XML, so if information is stored in XML, then it has to be pulled into something else in order for a site to operate and show pages quickly enough not to bore users. All software is really running on an index made from the XML. Software is using that index, not the XML, and the XML can seem superfluous in a self-contained system, such as your current library platform. Within a contained system, the XML is baggage and adds an extra step. Inertia cuts against writing software around XML. For page order, it is the norm to store information in something custom, so a migration to new software is extremely likely to require mapping from one standard to another. Moreover, if you are thinking about Portable Document Format (PDF) as a solution, PDF is a file wrapper and a PDF of a book is holding pictures together in a specific order while at the same time not necessarily being an open standard. The International Organization for Standardization (ISO) standard, PDFA, is tremendously limited in terms of what can be stored in PDFs in general. Happily, page order is usually a system-wide puzzle and tends to require a flat amount of time, regardless of how many items or collections exist.

Metadata Enhancements

Just as software updates can keep a site from looking long in the tooth, there are times when metadata needs overhauled to keep a once-current site from looking and behaving stale. For example, ten years ago, lacking faceted search was not a huge deal, but at this point, it is strange if a site does not have facets. Metadata that perfectly supports search does not necessarily support facets. Often, in adding nuance to facets, some systematic clean up is needed across a set of records.

There is no getting around touch ups, not even for perfect metadata. Think about the Virtual International Authority File (VIAF), which is Library of Congress' controlled vocabulary and identifier system for disambiguating author names. Users expect to be able to distinguish by name—they expect to be able to separate "Tom Wolfe" versus "Tom Wolfe" by facets, by browsing from a single record and elsewhere in the interface. Twenty years ago, for a common name, you would have looked to the Library of Congress Authority record and then pulled in the demographic info to the level of detail that would allow you to differentiate. For example, you might add birth and death dates, middle name, etc., to the author in the record for the item. Then you would periodically check over all the records for an author and make sure that author was always indexed the same (i.e. if one record has a middle name, then all records have a middle name, never just a middle initial and never no middle name). In addition, of course, no misspellings. VIAF uses an author identifier to disambiguate similar names and then looks to the Library of Congress Authority record to pull the demographic info. Let us assume that in ten years, it has become standard to facet on name, to use VIAF to disambiguate similar names, and that VIAF integration has become an essential building block to a clean interface. That will require a metadata clean up and enhancement of existing records. Even a digital collection with perfectly consistent internal records would have to go over records and implement VIAF in order to integrate well into a federated search project. (After all, when you purchase electronic resources and get MARC records to boot, your catalog is the federated search project.) Internal consistency definitely saves time, but it does not eliminate the need to make changes over time and the need for a skillset to manage the change.

Metadata is as much a specialized skillset as any other tech field, and it is just as inscrutable to outsiders. Complexity in a schema is harder to maintain, and there is a balance between enabling search interfaces and still having a manageable set of metadata. There is a whole world of standards to navigate and understand, and a quick Google search is better than nothing, but not as good as deep background knowledge and understanding. It costs money to get access to that skill through staffing or through outsourcing.

Conclusion

As much as this article has been long and detail oriented, the key point here is that maintaining digital collections over time costs money. Understanding the nature of these costs and why they are unavoidable helps to know what you are buying beyond content when you maintain your own digital collections and when you rent platforms from vendors. It also helps in understanding the reasons why less popular and hence less profitable digital content might disappear from the marketplace.