A Recipe for a Successful Digital Archive: Collection Development for Digital Archives

John McDonald
California Institute of Technology, john@library.caltech.edu

Follow this and additional works at: https://docs.lib.purdue.edu/atg
Part of the Library and Information Science Commons

Recommended Citation
DOI: https://doi.org/10.7771/2380-176X.4242

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.
A Recipe for a Successful Digital Archive: Collection Development for Digital Archives

by John McDonald (Acquisitions Librarian, California Institute of Technology, Pasadena, CA 91125; Phone: 626-395-6427; Fax: 626-792-7540) <john@library.caltech.edu>

At the 2002 Charleston Conference this past November, I was fortunate to sit on a panel addressing issues in Digital Archives. Along with my presentation about Caltech’s Digital Archive initiative, CODA (http://library.caltech.edu/coda), the panel included presentations by librarians at MIT about their DSpace project (https://hpdsl.mit.edu/index.jsp), from Ohio State about their Knowledge Bank (http://www.lib.ohio-state.edu/KBinfo/), and from director of production at JSTOR (http://www.jstor.org). This session highlighted the varying approaches that academic libraries and non-profit institutions are taking towards digital archiving of materials.

These project descriptions have led me to believe that we are at the right point to shift the focus of digital archive development from the technical to the methodological. We now need to apply collection development techniques to digital archives to make them useful, utilized, and important. I originally wrote that Caltech’s recipe for building our digital archive project included six ingredients: an entrepreneurial attitude, iterative process, learning to communicate, collaboration, defining and redefining roles, and patience. The new recipe will include a seventh ingredient: content.

Basic issues for digital archive development have in the past focused on technology — how to get an archive up and running, how to maintain it, how to fund it, how to staff it, etc. Most of these technical issues have now been solved or are being tackled on a grand scale (Eprints, DSpace, etc.) and there are now multiple technological approaches to building a digital archive. Content has not been at the forefront of digital archive projects in the recent past, but now should be. Some digital archive projects have been scattered out of necessity — items placed in the archive were readily available, easy to put there, either since they were already digital documents or were the easiest to convert, or were unique items that received special funding to convert (maps, images, etc.). Focusing on content, just like libraries in general, is what will drive digital archive projects in the future. Digital archives will be needed and used only if the content is relevant, accessible, and properly promoted. Digital archives are not only archival projects in the traditional sense but also libraries and need to apply principles of each to develop a common theory of digital archive collection development.

Collection development is built around the identification and evaluation of materials based on demand, quality, cost, and other local factors including storage and access points, both physical and bibliographic. The number one consideration for selection of materials for inclusion in a collection is demand from the primary user group, for current or future use. Selectors define their primary (i.e., faculty and students), secondary (i.e., community members), and tertiary (i.e., other libraries) user groups and select items that they feel meet the needs of those users. This demand is balanced with quality of the material and its cost — both initial and ongoing costs of the item and its processing and storage. In addition, local factors, such as space, language, and the ability to access the item (physical or bibliographic) can be taken into consideration.

continued on page 23

<http://www.against-the-grain.com>
Archival theory is built around many of the same principles. Archivists must select records based on their perceived future demand, arrange and describe the records to allow them to be found and used at a later date, ensure long term preservation of the records, and publicize and promote them. All of these must be balanced against the potential costs associated with each activity. Archivists are likely to be very selective in accepting documents and collections due to high costs associated with processing and storage and low or unknown prospects for use.

The stakes are high for digital archives since the infrastructure costs are significant and potential long-term costs are not yet known. This makes it especially important for developers, librarians, and archivists to balance the principles normally applied to collection development of print materials with those of archival documents to the digital archive environment. These libraries are collections of documents that should be chosen because they are important, and needed by their user group, and the materials' preservation and dissemination will enhance the scientific record.

Few libraries have the flexibility and resources to continue with the "build it and they will come" style of collection development for standard library materials. But as new types of materials are developed and come to the forefront of the attention of library users, libraries often revert to this technique of collection development. Now where is this more apparent to me than in the recent collection development models for electronic journals — consortial deals, package purchases, and the "Big Deal," all developed out of the idea that if we provide a multitude of material then our users will utilize it. While this technique works for some user populations and some institutions, it does not work well as a long-term collection development strategy for most libraries. Archives have never functioned in this manner, archivists are selective about their content and build in a particular area of strength or based on format. Archivists have known for a long time that the technique of "build it and they will come" does not work for the specialized collections that they need to preserve.

Digital archives have, in the past and by necessity, focused on materials that are easy to acquire, easy to "make digital," or present few difficulties for the archive developers. Since most developers had little knowledge of exactly what would work when building the digital archive, they used these types of materials as test resources. At Caltech, the computer science technical reports collection was chosen as our first collection for digital archiving due to its ease in conversion and visibility. The collection was already a part of a national project (NCSTRL) but needed to be brought up to date. Additionally, the materials were already electronic, making conversion to archival standards relatively simple. Another example was the start of our Theses & Dissertations collection that was already being planned for new dissertations but that was populated with converted print dissertations due to flood damage to the backup archival copies — making their conversion time-sensitive and necessary anyway. All of these efforts were in the early stages of digital archive development at Caltech and were useful in the procedural and technical development of our archive. Now we must turn the focus towards building an archive that our user population, however defined, will use while still accomplishing the goals of protecting local research and promoting institutional resources.

Building a digital archive is a long-term commitment to the content, making it especially important to select content that has long-term interest or prospective demand. We must define our user community, whether it is local users, regional users, subject specific users, or national and international users. This can be done on a case by case basis, as in the print world. For example, Caltech has the Cavitation 2001 digital archive that is built around the proceedings for a specific conference published at a specific point of time. The user community at that time was the conference attendees and the future users included those individuals plus anyone researching that subject at a later date. Archive developers should have a clearly defined user community and should select content that is most relevant, either immediately or in the future, to that user group.

In addition, we must define our scope of collection — retrospective, current, or both? A
provides a collection that spans a number of years, making it more likely to be used, but also presents some difficulties with populating the archive with disparate document types. Most archives at Caltech are mixed retrospective and current archives. I expect that most digital archives in the future will include items retrospectively scanned or converted plus newly produced items.

Additional considerations should be given to other issues in collection development as they relate to digital archives. These issues include continuing economic commitment to the collections, the costs associated with processing and adding material to the archive, the comprehensiveness of the collection, the uniqueness of the material, and physical and bibliographic access to the material. All of these issues impact collection development in the print libraries and archives and will affect digital archives in the future.

Most importantly, to maximize the effect of digital archives, we must not only identify the material that is most valuable to the defined user community and seek that material for our collections, but also promote and publicize that content. Participating in federated searching as an OAI data provider is the first step in getting the archive indexed and accessible by end-users. In addition, archives should be spied by major search engines, publicized in library and subject specific publications and listservs, and otherwise promoted in any way.

As libraries build and promote digital collections, we must follow and build upon the principles of collection development that have been established for the print world and for the archival world. If we build high demand, high quality collections at a reasonable cost that can be maintained for the long term, we will take the first steps to becoming a major part of the scholarly research dissemination chain.