Against the Grain

Volume 15 | Issue 1 Article 8

February 2003

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

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

Recommended Citation
McDonald, John (2003) "A Recipe for a Successful Digital Archive: Collection Development for Digital Archives," *Against the Grain*: Vol. 15: Iss. 1, Article 8.
DOI: https://doi.org/10.7771/2380-176X.4242

Acquiring Minds Want to Know
from page 20

ship can achieve the goals of integration and multi-purposing of material.

Another initiative that promotes many innovative projects is the **Electronic Cultural Atlas Initiative (ECAI)**. **ECAI** focuses on global mapping, imagery, and texts to provide access to "research based on digital technology which presents complex combinations of data from multiple disciplines visually and immediately." (*http://www.ecai.org/*) Based at **UC-Berkeley** and under the Dean of International and Area Studies, **ECAI** often deals with materials from or about other countries or regions of the world, but not exclusively. In fact, the *Valley of the Shadow* has a project component with **ECAI**. **ECAI** supports research into new technologies and infrastructures, develops standards, and sponsors conferences. It also acts as a collective where participants can list information about and provide links to their projects. Some are **ECAI** developed projects and are hosted at the **ECAI** site. Additionally **ECAI** collaborates with the **California Digital Library's eScholarship** program, another locus of new scholarship (*http://www.escholarship.cdlib.org/*).

The **Andrew W. Mellon Foundation** has been a strong supporter of digital scholarship, particularly that which is created in the humanities and social sciences. **Don Waters**, Program Officer for **Scholarly Communication**, has often spoken about the projects that they have sponsored. Their scholarly communication program supports digital scholarship because of the impact it may have on the current system of conducting and communicating research.

Unfortunately, scholarship of this nature is still so new that often it goes unrecognized and unrewarded in the promotion and tenure process. The younger scholars who have grown up with computers, and are most likely to experiment with and engage in this new form of scholarship, are discouraged from exploring and developing this path. **ECAI** is tackling this dilemma head on. On their Website they state, "We are entering into a dialogue with university administrators to ensure that **ECAI** publications are just as acceptable as books and articles to the people who make hiring and tenure decisions. In this way, we hope to create a climate that allows young scholars to publish in the medium that is most meaningful and intellectually exciting to them." Let us hope that the recognition afforded by awards such as those made by the **Mellon Foundation** or by being part of **ECAI** will begin to change the system and will allow digital scholarship to flourish.

Libraries are becoming more supportive of and engaged with scholars in the creation of digital research and dissemination of its results for several reasons. For one, librarians often have useful skills to apply, such as the creation of metadata schemes. For another, librarians have become increasingly involved in the research of faculty and the teaching of information literacy to their students. Librarians also have a strong and vested interest in fostering movements that have the potential to change the current system of scholarly communication. Libraries also have some of the equipment and space necessary to set up these projects, having carved out these resources for digital reformatting of library collections. Last but not least, librarians are in the business of preserving scholarly materials in many formats, and the best way to preserve digital scholarship may be to actively participate in its entire life cycle.

The current drive by libraries to create institutional repositories has not yet grappled with the complexity of true digital scholarship. Institutional repositories more often than not are starting to try to capture and preserve more text based and simple forms of digital works. Digital scholarship will require an exceptional level of commitment and risk for libraries over time. Will our efforts be worth it? The scholars will surely let us know.

*(**Note**: See also the November 2002 **ATG** article by **Tony Ferguson** (p. 94). He touched on a few aspects of digital scholarship and mentions the **ECAI Silk Road** project in his discussion on **GIS**.)*

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://hpds1.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 scattershot 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

A Recipe for a Successful . . .
from page 22

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. No 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

continued on page 24

A Recipe for a Successful . . .
from page 23

retrospective archive is finite, developed to house a set of documents that have been published. Once populated, the archive is considered finished. An example is the **Cavitation 2001** archive at **Caltech**, where documents were produced for the conference proceedings making the archive complete at that time. A current archive is developed to house documents that are currently being produced. These archives include digitally born documents or print converted documents, but the archive is built as documents are being produced. This has the combined effect of producing archives that are initially small but are relatively current. A mixed retrospective and current digital archive includes items that were previously published but also those items that are being produced currently. This 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.

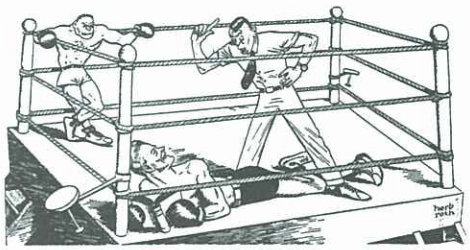
Additional considerations should be given to other issues in collection development as they relate to digital archives. These issues include continuing economic commitment to the collections, the costs associated with processing and adding material to the archive, the comprehensiveness of the collection, 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 spidered 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.

The Market Power of Publishers

by **William M. Hannay** (Attorney, Schiff, Hardin & Waite, 7200 Sears Tower, 233 S. Wacker Drive, Chicago, IL 60606-6437) <whannay@schiffhardin.com>



A Presentation to the Charleston Conference

A good deal of concern has been expressed in library circles in the past few years about the growing market power of publishers, particularly commercial publishers of scientific, technical, and medicals serials and journals. The existence of this market power is seen in significant subscription price increases and in the bundling of journal subscriptions.

As an example of real world market power in academic publishing, here's a quote from the May 12, 2001 issue of *The Economist*: "if a company owns a must-read title in say, vibrational spectroscopy, it has a nice little captive market."

A fair question to ask is, "what is market power?" and in turn, "do publishers have market power?" From the perspective of antitrust law, "Market power . . . is the ability profitably to maintain prices above competitive levels for a significant period of time. [T]he result of the exercise of market power is a transfer of wealth from buyers to sellers or a misallocation of resources." This widely-accepted definition appears in the *Horizontal Merger Guidelines*, jointly issued in 1992 by the **U.S. Department of Justice** and the **Federal Trade Commission**. The Guidelines go on to note that market power can lead to effects beyond price. "Sellers with market power also may lessen competition on dimensions other than price, such as product quality, service, or innovation."

In order to understand the application of this definition, let us ask whether hypothetical Publisher XYZ has "market power." Let us assume that Publisher XYZ has only one journal focusing on the subject of, say, brain lesions in rats. Further assume that no one else publishes such a journal. Query whether, at this point, we know enough to answer the market power question. The answer is actually no. Answering the market power question just isn't that easy.

To assess the presence or absence of market power, there is a variety of other data that one needs to know. For example, one needs to ask what, if any, leverage XYZ's customers have in the market. How "elastic" is the demand? To judge the elasticity of demand, we need to ask whether libraries just have to have it. Will they pay largely any price to get it? Put differently, how special is this rat brain lesion journal? Could a competing journal enter the market? Would new entry be easy or hard?

In addition to these fundamental sorts of questions, what else affects market power?

What if Publisher XYZ has 100 titles or 1,000? Is its market power over the hypothetical rat brain lesion journal any greater? And what if Publisher XYZ merges with one, two, or ten other publishers? Is its market power greater? Has the possibility of market entry by a competitor been reduced? Eliminated? These are complex and thought-provoking questions, but perhaps we can learn something from a recent case which analyzed market power issues in the context of the 2001 acquisition of **Harcourt General Inc.** by **Reed Elsevier PLC** under the antitrust laws of Great Britain.

As required by those laws, the proposed acquisition was notified to the British authorities and, in turn, referred to the **U.K. Competition Commission** for investigation on February 21, 2001. As part of that investigation, a "Statement of Issues" was sent to the parties by the **Commission** on March 19, 2001. In U.K. practice, a "Statement of Issues" is similar to a subpoena or civil investigative demand in the United States, requiring the parties to submit detailed answers and/or related documents to a government agency. The contents of that "Statement of Issues" bear some consideration in detail as an illustration of how an antitrust regulator examines the market power issue in the context of publishing.

The investigations focus on the parties' business was narrow, because the regulator concluded early on that "the only parts of their businesses with potential to give rise to competition concerns were sales of STM journals, in both printed and electronic formats, in the U.K." (Press Release P/2001/351, 5 July 2001, "UK Competition Commission's Investigation Clears Reed Elsevier/Harcourt General Merger").

Under U.K. antitrust law, a merger is illegal if it will operate, or may be expected to operate, against the public interest, taking into account the following factors:

— Will the merger maintain or promote effective competition?

— Will the merger promote consumers interests re: price, quality, and variety of the goods or services supplied?

— Will the merger promote cost reduction and product innovation?

continued on page 26