

Digital special collections: the big picture

A talk by Alice Prochaska for RBMS, June 25th 2008

As I started to think about the theme of my talk, and the overall theme of this conference, I found myself pitching into a morass of unknown quantities and speculation. The dominance of electronic journals and aggregations of databases, followed by the news-grabbing mass digitization program at Google, has diverted attention from the fact that some publishers, and some of our libraries have been digitizing special collections for a couple of decades or more. And many more have been evangelizing for the benefits of digital access to unique and rare materials for most of that time. There was a time when I thought I had a clear vision of the future for sharing digitized versions of special collections with the world of scholarship and researchers far and wide.

Making high-quality images of special collections available on the internet has opened up for archivists, curators and librarians some dizzying possibilities. We find our collections now far more often at the center of our libraries' attention than they were twenty years ago. We have a new set of choices in the realms of preservation, reformatting and security. We are able to pursue high ideals for sharing a common cultural and historical inheritance by digitizing rare and unique materials for a world-wide audience. Now we can return the documentation of traditional cultures in digitized versions to the communities from where those collections came. We play new roles, with new partners, in placing "our" collections at the heart of the learning and educational experience. New forms of research develop, facilitated by ourselves through the materials that we bring into the scholarly domain and the links and software developments that we create with our colleagues in information technology and the electronic industries.

But it was always a complicated picture. Complicated by problems of scarce resources, and expectations that continually outrun the available technology and expertise. Complicated by

politics, legal issues, and organizational boundaries. Complicated very often by the ways in which cultural and historical ownership interacts with the responsibilities and values involved in stewardship of the original materials in our care. Often, for example, the very communities we seek to serve, or ought to serve are also those from whom our special collections derive; and ownership in the digital environment is no simpler than in dealing with the physical artifacts. And now with the acceleration of mass digitization, awkward conflicts between publishers and libraries that show no sign of reaching resolution yet, and an exponential increase in the technical solutions that are on offer, visions of the future seem more elusive.

We deploy in our community words like Discovery, Exposure and Disclosure, (or at a more detailed level, “preliminary record” or “collection level record”) as though one of these, each with its nuanced differences of meaning, holds the key to the new policies that we must formulate and embrace. A speaker in a gathering of experts in this field ventures trepidatiously to define the “big picture”, knowing that every word carries with it potentially a host of different meanings. I take some comfort from the thought that this, being a general election year, is a season for flinging words around, willfully misunderstanding them, giving and taking offence, and generally massacring the meaning of everything. In just one day’s issue of the New York Times a few days ago, I found two suggestive examples. Gail Collins wrote after a news conference in the Rose Garden that President Bush, speaking of a Democratic spending bill, had almost certainly called it not “omnibus” but “ominous”. Perhaps that’s what he meant. And Michelle Obama appearing on The View, was heard to describe her husband as “sweet and pathetic”, when what she actually said was “sweet, empathetic”. I don’t wish to underestimate the potential for misunderstanding in the topic before us. But I do hope, to quote a source from the safe distance of over 200 years, Mrs Malaprop herself, that as I venture on to this difficult and sometimes contested field, I can do so without attracting too many “aspersions on my parts of speech”.

What are special collections today, and what will they be in the future?

The Association of Research Libraries set up in 2007 a new Special Collections Working Group composed of a mixture of ARL library directors, heads of special collections, and some ARL leadership fellows from special collections libraries. Conscious that we are building on large bodies of work on this subject, including that of RBMS, and the predecessor of our own group, a task force that did make an impact on the special collections agenda in ARL libraries, we set ourselves the following modest charge:

“The Working Group on Special Collections is charged with advising the Research, Teaching, and Learning Steering Committee on special collections issues on an ongoing basis. In this context, “special collections” is construed broadly to include distinctive material in all media and attendant library services.

Priority Issues

The two issues that the Steering Committee identifies as first priorities for the attention of the Working Group are

- 1. Identify opportunities and recommend actions for ARL and other organizations that will encourage concerted action and coordinated planning for collecting and exposing 19th- and 20th-century materials in all formats (rare books, archives and manuscripts, audio, and video, etc).*
- 2. Identify criteria and strategies for collecting digital and other new media material that currently lack a recognized and responsible structure for stewardship.*

These two issues are closely linked. An enormous amount of valuable material in all formats remains uncollected and risks being permanently lost. Coordinated strategies for identifying, collecting, preserving, and exposing these materials are greatly needed.

(I ought to make plain here in case anyone doubts it, that the focus of the ARL group is not intended to suggest that materials pre-dating the nineteenth century are unimportant or in many cases, not at risk: it is simply a pragmatic attempt to give to the group a manageable task.)

International efforts are underway – and more are needed – to support the digitization of 19th- and 20th-century newspapers and books. Even before such digitization is possible, strong efforts must be made to identify and acquire culturally significant materials from these periods. While individual libraries should ultimately take action to acquire and expose such materials, ARL can provide leadership for encouraging collective activities. These would include but not be limited to, collection analysis, identification of gaps, coordination, and use of a “preliminary record” for identifying and making accessible otherwise hidden collections.

General Issues

In addition to the priority issues listed above, the Working Group may also wish to advise the Steering Committee about the following:

- *Ways to illustrate examples of how special collections contribute to innovative research, teaching, and learning.*
- *Contributing to the work underway within ARL to develop qualitative and quantitative measures for the evaluation of special collections. These might include a target for surfacing hidden collections and mechanisms for tracking progress.*
- *Contributing to and/or validating the work being done by the ACRL/RBMS Core Competencies Task Force to define the skills needed for work in special collections.*

From time to time, the RTL Steering Committee (the umbrella committee for this working group) may ask the Working Group to provide advice on other issues. For example, the Working Group may be asked to address preservation strategies for special collections in both physical and electronic spheres following Steering Committee discussion of a report from the ARL Task Force on the Future of Preservation in ARL Libraries.”

If that charge serves as context, let me elaborate a little.

We are all, I believe, obliged to take stock of the way we define special collections in the modern world; I don't just mean what are the new special collections in a digital environment, although I'll come on to that in just a moment. I think also of the way we have construed our responsibilities in relation to materials from centuries past. And this is something of which the leaders of ARL libraries, where some of the great concentrations of special collections reside, are acutely conscious. It has been a common experience in the research community to rediscover the intrinsic value of different sorts of material, often in response to new scholarship and sometimes ourselves leading scholars toward neglected material that turns out to contain revelatory new insights.

To use one category of special collections as a metaphor for the rest, let's take the case of printed ephemera. At one time, few libraries collected printed ephemera or at least only the fugitives and strays that slipped between the pages of rare books or turned up among the papers of statesmen, literary figures, or local dignitaries. It was left to the hobbyists, maniacal amateur collectors of menu cards, bus tickets, beer mats, postcards, wine labels, playing cards, and on and on, to insist on the importance of such materials. The librarians and archivists still have to explain to the enthusiasts our need to be selective; and faced with the insistent donor, we give such items an occasional home, throwing up our hands at the challenges of preservation, storage, and still worse, the task of providing any kind of descriptive standards.

Among printed ephemera today there is still, I would say, a distinct hierarchy of materials. Anything that we might construe as relating to arts of the book or print history will generally find a place among the aristocracy of library ephemera. Playbills are theater history, of course, and posters, especially in such subject categories as politics or sport, constitute a class of their own. But I know, with some residual feelings of contrition that date from my days as director of special collections at the British Library, how difficult it is to persuade great research institutions

that philatelic collections possess research value, even some of the greatest and most comprehensive of their kind in the world. Many of us in this room will have our own personal favorite, or perhaps our own uncomfortable consciousness of neglected collections languishing in our most hidden corners. I suspect that no two of us would define the word ‘ephemera’ in quite the same way though, even if we share a certain rueful recognition of some of the adjectives that Roget brackets with “ephemeral”: impermanent, transient, insubstantial, fragile, fleeting, and fugitive.

In other words, the task of selecting collections for preservation and incurring all the costs associated with responsible stewardship was never an easy one in the pre-electronic world. That task does not disappear in the digital context, but it is sharpened and changed in important ways. How do we define a born-digital special collection? Is it defined by the way we handle it, and the special skills needed? Is it almost anything that is NOT an online journal or published e-book? What about second or third generation reproductions, for instance digitized slides. Are those special collections? Does email correspondence fall into a particular category, or do we treat it as a records management problem that only archivists will know how to handle? (And as soon as we start thinking about email we are drawn into issues affecting literary creativity in the cases of writers’ email, legal issues and the problem of “discoverability” which is quite different from the innocent term “discovery”, and more...) We need a concerted effort to define our terms in the electronic environment. Meanwhile, archivists and rare book and manuscripts librarians, including all those of us who are charged with the care of multi-media collections in numerous formats, face a perplexing array of issues and tough choices.

This talk focuses primarily on the digitization of special collections, which I take to mean rendering analog materials into electronic format. But as archives increasingly arrive in hybrid form, and we are challenged also by special collections that have been partially digitized already, but not necessarily in complete or ideal versions (I think of published aggregations of rare books, for example, or editions of newspapers that leave out the advertisements), the activity of

digitizing frequently coincides with choices about how to treat materials that are already available in some kind of digital form. So the question “what ARE special collections?” lurks behind much of what I will be saying.

The environment of digitization

The ARL Special Collections Working Group will present recommendations including the following headings: Collecting Carefully (with close attention to the total costs of caring for and making available a collection, be it digital or in other formats); Advocacy against restrictions on Access; Transparency over provenance, and the source of acquisition; Good practices in records management; Ensuring discoverability and access; Address the hidden collections problem; and the Digital Challenge. I will focus now on the environment in which the digital challenge presents itself, some glimpses of the current scene, and finally a kind of peering, hopeful glance into the future. I hope I have said enough already to make clear my belief that the collective thinking about digital special collections inevitably grows out of our experience in the physical, analog world, and must be informed by it.

But before I move on, let me issue an invitation. If there are two national organizations that between them embrace most of the professional wisdom on special collections in the research libraries, archives and historical organizations of North America, they are RBMS, and the Society of American Archivists. The Association of Research Libraries is essentially a meeting place for the directors of research libraries, advised by experts in the many fields that our organizations encompass. Although we are wonderfully supported in our working groups and committees by program officers, specialist fellows and some co-opted professionals from the particular areas of work affected by each committee (and several members of the Special Collections Working Group are here today), we can only be effective if we know we have drawn deeply on the professional experience and insights of the community, and we can only be influential if we have their support. So I am hoping to get some feedback from all of you, and from the SAA, as soon

as our group can get our draft report into presentable shape to circulate. I hope to work out while I am here, how best to do that.

The environments in which we work differ, of course, enormously. There are huge disparities in resources, a wide range of different missions, and any number of different collecting traditions. Here are some hugely oversimplified generalizations.

On university campuses and in the broader scholarly community, the essential elements to support digitization of special collections include an infrastructure to support preservation, access and the re-purposing of digitized versions. Digital repositories of anything approaching adequate size or sustainability are still the exception rather than the rule. On the other hand, increasing numbers of universities are putting this infrastructure in place, at a cost of millions, sometimes tens of millions of dollars. And faculty members increasingly expect it. The presence on a research campus of a trusted repository that can hold and sustain both the digitized collections of the university's libraries and museums and the vast array of databases and image-based research output of faculty in all disciplines will come to be a defining quality of the leading research university. Where the technological know-how and the capital are not available in a single institution, partnerships and collaborations are bound to grow up. State institutions can provide this sort of facilitation sometimes more readily than private universities, the classic example being the California Digital Library.

Smaller-scale independent scanning facilities exist now in most libraries and archives, and there are few that lack a web-site on which they can display images of some of their treasures. Few are totally without the capacity to provide scanning on demand, and indeed it is through ad-hoc digital services of this sort that many special collections departments in university and college libraries first began their own digital libraries. Preservation issues form part of the context (and I will return to this question when I come on to the topic of mass digitization). The capabilities of the special collections staff, and their attendant metadata skills and standards inevitably lie at the

basis of many decisions about digitization as well. Most ARL libraries and the big archives, especially national organizations, possess at least some of the infrastructure they need in terms of both equipment and capital investment, and staff.

For a huge number of smaller organizations throughout this country and around the world, that is not the case, and it seems to me that there are compelling arguments for collaborations and consortial arrangements. Among the compelling arguments for higher education to set up robust, sustainable infrastructure for digital special collections, whether within single institutions or in consortia, is the huge growth in teaching with primary sources and independent undergraduate research, creating an inexhaustible appetite for digital versions of materials that can supplement the original or, even when original rare books, manuscripts, photographs and so on dealing with the chosen topic are available, can preserve them from too much handling. Equally compelling is the scholar's need to search large quantities of rare material without moving from the spot: a facility already available in the form of electronic serials and large scholarly databases, and for which there is now an exponentially growing appetite.

I do not want to neglect, in this brief survey of the environment and market for digital special collections, the broader societal benefits of making our resources available throughout the networked world. Digitization of collections that relate to particular communities be they local or defined by ethnic or cultural origin can bring great social and educational benefits to the users and political benefits to the originators. There is a continuing explosion of interest in local and family historical research. Political pressures to extend access to unique assets include the recent proposal by Senator Charles Grassley, to tax the wealthiest not-for-profits, including universities such as Harvard and Yale by 5% if they do not spend a specified proportion of their endowment to increase access. Access to collections is part of that scenario.

Countervailing pressures include the thrust to protect intellectual property. There is intense but unpredictable scrutiny of material that is available on the web and making special collections

available digitally carries with it the potential penalties of higher visibility. As just one example from among the problems that universities have encountered recently, Cornell faced a law suit from an alumnus who objected to seeing a decades-old unfavorable story about himself on the web in the digitized student newspaper. Digital special collections need to be packaged. They carry with them the danger of misinterpretation, which cannot be mitigated by personal intervention on the web in the way that can happen in a reading room or classroom. And the huge opportunities to extend the accessibility and benefits of special collections that digitization presents undoubtedly carry with them potential drawbacks in the shape of more claims to the ownership of content, and more contested interpretations.

The current range and scope of digitization in research libraries

If that is the environment, what sorts of programs are flourishing within it? Each of us will have favorites. There are so many local or subject-based enterprises that it is impossible even to characterize the landscape at this stage. Which are the projects that will turn out to be the foothills of mature mountain ranges, or the streams that become tributaries of some great Mississippi of digital content? (The Amazon metaphor, of course, is already spoken for.) Among my own favorites are some digital initiatives at Yale designed to support teaching and learning, with more than thirty courses now supported in an interactive way through collaborations between faculty, librarians, teaching and information technology specialists. And I am also impressed by the proliferation of collaborative digitization programs that build on collections in museums, galleries and archives as well as libraries. My personal all-time favorite is Documenting the American South, that wonderful project based at the University of North Carolina which was set up by Joe Hewitt, who chaired the ARL special collections task force, predecessor of my working group. I wish I had time to dwell on this project as a paradigm for the digitization of special collections within a broad theme. If you are not familiar with DocSouth, do take a look at its web site, and don't neglect the letters and emails from users ranging from the descendants of slaves and slave owners, to an overseas MA student, and

numerous scholars and community organizers. It brings home as vividly as anything can, the reasons why digitizing special collections, in an intelligent and programmed fashion, can be so worthwhile.

Mass digitization programs meanwhile are numerous, and not confined to Google, the Open Content Alliance and the recent, relatively short-lived Microsoft Live Book Search. Specialist mass digitization, of newspapers for instance, is also becoming an attractive option, building on earlier projects such as EEBO that are basically aggregations from numerous sources. (Are newspapers special collections, by the way?) Mass digitization programs are generally projects involving multiple libraries; and from the perspective of each library, the task is too great to be taken on without the involvement of significant corporate involvement and economies of scale that go beyond the capacity of even the largest research library on its own.

The impact of Google Print (now Google Book Search) has been momentous. Set aside the culture wars that threatened to break out at one time between the Anglophone and Francophone digital empires, and the fierce debates in both the scholarly and general press, about the merits of digitizing large quantities of books. It is indisputable that Google's extraordinary coup, incorporating five of the largest research libraries in the English-speaking world into their initial program, which has now extended to a number in the forties and across the world, has had a huge impact on the way libraries do their business. For a brief and digestible conspectus of many aspects of the Google impact, I commend the current issue of the *Journal of Library Administration*, entitled "Googlization of Libraries".

Among the issues that particularly affect digital special collections are those to do with quality of the output. Famously, Google Books are not always well presented, with folded-over pages and the occasional intrusion of an operative's finger, not to mention problems with inconsistent and sometimes inadequate metadata. The quality of handling is a source of anxiety. Google's machines are not widely demonstrated, and there are concerns arising out of ignorance, as much

as empirical results, about the way in which books are treated at the point of scanning. On the other hand, it must be said, the libraries working with Google currently are not reporting significant numbers of damaged books returning from the scanning operation; and Google have been working hard to improve the quality of their product.

More intractable is the problem of copyright. When the legal dust has settled, and that surely will not be soon, there still remain many questions about presenting just ten per cent of a book that is in copyright, for viewing on screen. One of my own great concerns is about the impact of this way of presenting books on the whole educational enterprise. It seems to me to do violence to the principle of scholarly argument, to facilitate the use of small snippets of a book, out of its overall context. Several leaders in the world of research libraries (Chuck Henry of CLIR is one) have expressed serious concern that Google's business methods, which include an extremely high degree of confidentiality, locked in by non-disclosure agreements with each participating library, have made it impossible for the library community to discuss adequately, still less to influence, the process whereby millions of their volumes appear on the internet. Given the tight contractual restrictions on re-using the digital files, it is a serious concern that certainly limits the universality which surely was one of the great attractions of this enterprise in the first place.

One undeniable benefit of the Google mass digitization program is its influence. The Open Content Alliance based on the Internet Archive, although not enjoying the wealthy resources of a corporate behemoth, is making available a large quantity of digitized material and has great potential. Its open source principle makes it possible for libraries and archives to share their digitized content whenever it is free of copyright restrictions, and the potential benefits to research from this collective approach are breathtaking.

Meanwhile, for a short time Microsoft's rival approach, with the Live Book Search project into which several libraries entered, including the British Library, Yale, Columbia and Cornell may have been short-lived but will leave a legacy of expectations. As we at Yale wind down our

Microsoft operation, we value the careful handling, and closer engagement with the library that characterized Microsoft's approach; and we are keenly interested in continuing our program with the technical partner, Kirtas Technologies, whose robotic page-turning machine offers a promising solution to the handling of fragile bound materials. We are also hoping that with Kirtas, we can move quickly into the realm of mass digitization of special collections: pamphlets and rare books on the robotic machines; maps and archives, perhaps on additional flat scanners.

I will not list all of the now numerous separate mass digitization enterprises that have sprung up alongside Google, Microsoft and the Open Content Alliance. But there are a few more consequences that should be mentioned as an important part of the environment. Scholars now have an appetite for increased quantities of digitized material, and this in turn has an impact on individual universities' policies, helping to build the support we need, and make the case for capital investment in a robust infrastructure. We are now able to pay close attention to collection development issues. It was a revelation to my own library to learn from a preliminary analysis of OCLC's Worldcat, that we have something like one million volumes that are not held by any of the original Google Five libraries (and for them too, presumably, there must be huge quantities for unique holdings). Mass digitization has undoubtedly spurred OCLC and RLG programs to develop their collection analysis tool, a wonderful step forward that will inform libraries' local decisions about collection development, as well as laying the ground for future digital partnerships. Here perhaps, is one possible route to the holy grail of collaborative collection building; if not in the physical world, then in the digital one. And here for sure is a tool in the hands of all of us who wish to increase the availability of our special collections in the digital environment.

The future

What priorities derive from our responsibility as the stewards of multiple inheritances?

First of all, one critically important recommendation of the ARL working group will be that there should be no digitization without metadata. Discovery, that primary function of research in special collections, is not possible without description and guidance. The Hidden Collections agenda, as defined at a conference organized by ARL at the Library of Congress in 2003, has moved into the center of the stage, and the importance of surfacing special collections is essentially what that agenda is about. Most of you will be aware of the new program of grants from CLIR, supported by the Andrew W. Mellon Foundation, which is sending out a call for proposals in mid July. This program explicitly exists to promote cataloging and description, not digitization for its own sake. With a budget of \$20 million to be spent over five years, it will make a significant contribution to the fundamental task of exposing and disclosing material that our organizations lack the resources, on our own, to place in the public domain. No digitization without metadata. The future, in this respect, looks like the past but, we may hope, with somewhat more resource to support the work we need to do. And that resource will necessarily be spread thin. Although the specifics of the call for proposals have yet to be published, CLIR is looking for projects that will provide models of good practice for the lean, effective, description of collections that until now have been hidden altogether from scholarly enquiry. Access begets access: let's get the stuff out there, and then work on it some more when it's clear that it's needed.

That still leaves us with any number of issues as we shape the future of digital special collections. For example, what is the future for born digital collections, and how do we shape that? What ARE born digital collections?

We really only know for sure, at this stage, that these are some of the skills we have to develop and the issues we need to address:

- Digital curation

- New relationships to our users: how do we connect with them?

- Working with our users to describe collections

Mirroring traditional functions in OAIS models

Coping with volume

Identify partners

Be prepared to wait for new technology to help achieve more satisfactory curation

Finally, as we contemplate the future and think about the contest over standards and metadata, platforms, decision about commercial versus in-house development, the battles that are in progress and yet to come over adequate funding for infrastructure: as we contemplate the emerging outlines of this still clouded future, we can at least tell ourselves that special collections librarians and archivists are not unused to such battles and uncertainties. Our work will ALWAYS be a work in progress. We are not going to achieve perfection. But, as we work together to invent and reinvent our future and the future of our collections, we can at least aspire as Mrs Malaprop might have hoped, to be models of the “very pineapple of politeness”.