

“New Frontiers in Access to Historical Primary Sources,” New England Library Association Annual Conference. Burlington, VT | October 22, 2006

Before taking us on a tour of the Open collections Program’s web sites and collections, and in complete contradiction to the rather bold statement about “new frontiers” in the title of this presentation, I’m am going to begin by taking us all back to library school for a little while.

So let’s start at the beginning.

By the time the great libraries like those at Alexandria and Ephesus were destroyed, humans had already been working on this business of accumulating cultural artifacts for quite a few centuries. The physical geographies in which these activities took place, were far more than just libraries. Today, they can best be compared with the variety of establishments (libraries, museums, archives) we collectively call ‘cultural heritage institutions.’

It is not without irony that in looking back at these libraries, we realize one of their central activities, was copying materials. This was their primary means of collection development. Those who once served as the librarians in these institutions were multidisciplinary scholars. Demetrius of Phaleron, who founded the library at Alexandria, had studied with Aristotle. Hypatia of Alexandria was a mathematician and a philosopher.

These librarians were scholars were scholars in their own right, often actively translating and editing the texts in their care. So these institutions were far more than libraries as connoted by the word today, they we very active, multi-disciplinary spaces for the transformation of information into knowledge. And the librarians weren’t just caretakers, but participants.

Almost 2 millennia later, in 1931 SR Ranganathan reminded us of these very qualities when he wrote:

“In a word, the librarian should be ‘friend, philosopher and guide’ to every one who comes to *use* the library. It is such sympathetic *personal service* and ‘such hospitality that makes a library big, not its size’”

But what Ranganathan is most renowned for are his five laws of library science.

The five laws were really all I learned about Ranganathan in library school, and I don’t think I ever read and of his books in entirety.

But about 8 months into my position with the Open Collections Program Ranganathan came up in conversation I was having with Tom Michalak (the founding director of OCP), and I decided to reread the whole Five Laws of Library Science.

I was astounded by the relevance of what he said 70 years ago to the task of creating digital collections for the 21st century.

I had seen a number of the recent reinterpretations of the five laws for the web, or for information architects, or open source software, but what I hadn't realized, was that Ranganathan doesn't really need to be reinterpreted. His message is really quite clear. He places modern librarianship on the same foundations that gave rise to the great Hellenistic libraries, the foundations that maintained the monastic libraries of the Middle Ages, and surprisingly those needed to anchor our own work at OCP: Librarian and philosopher, friend and guide.

In 2003, with the support of the William and Flora Hewlett Foundation, Harvard established the Open Collections Program. The goal of OCP is essentially, to open up parts of Harvard's libraries, museums, and archives to the world.

So, I began to think about this in the context of Ranganathan. If the world is our target audience – essentially our patrons for these projects– then suddenly I was realizing that it was our job to become “friend, philosopher, and guide” to the world. Kind of a daunting task. Tom and I started by developing a set of principles and standards by which the open collections program is designed to operate. These guidelines were extrapolated from Ranganathan's five laws and concerned

#### SELECTION, PRODUCTION, and ACCESS

So, let's start with selection.

Ranganathan's first law is that books are for USE. He describes this elsewhere in his books this way:

“If the library keeps books for use, the task of the librarian is not to dump down a mass of books and tell readers to help themselves. Nor is it to forcibly feed them on books of your choice. It is to help them; and, to help anyone is to co-operate with him in carrying out his own plans and wishes—to help him to help himself.”

Let's break this down and just look at the first part of this quotation.

“If the library keeps books for use, the task of the librarian is not to dump down a mass of books and tell readers to help themselves.”

This, to me, this is really the essence of why we select.

And in OCP, we built our selection standards around the goal of creating comprehensive, topic-based digital collections by carefully selecting both the topics for the collections, and the materials within them;

Harvard has a pretty big library system. There are 90 libraries and over 10 million records in the library catalog. Just in the Boston area alone, the libraries are spread over 4 or 5 different cities, so our first task was to figure out how to select a subset of the materials in all of these libraries, and to subset in a logical way that promotes use... is what it means to build a collection.

The first thing we had to selection was the topic for the collections; the second was the materials within those collections. We determined that each collection should have the following attributes:

- Topics of collections should be interesting and have broad appeal within and without Harvard
- Materials must be academically important for the study of the topic, yet interesting to teachers and students
- The faculty should be willing to shape the direction of the collection and provide guidance and advice
- There must be a critical mass of resources at Harvard to allow for significant exploration and study of the topic – it is surprising what Harvard DOES NOT have in its collection
- There should be a mix of formats available for each collection: books and pamphlets, manuscripts, photographs, etc.
- The resources should be drawn from a number of libraries across Harvard.

This last point was in contrast to what we saw going on with digital projects both at Harvard and elsewhere. LC's American Memory is a good example of the opposite approach. They have the African American Pamphlet Collection and then they have the Daniel Murray Collection of Pamphlets on African American Perspectives. These are separate analog collections, and from them they have created separate digital collections. You can use their cross-collection search to search across them, but you can't browse them together. The essentially remain separate.

But our theory at OCP was that the user probably cares more about being able to get to and efficiently navigate through these collections than they do about where they are housed. Why should collections separate, now that we are dealing with a format in which physical location makes no difference?

Let's have a look at Women Working as an example of these selection principles put to use. Women Working is about women in us the US economy from 1800 to the great depression (or, coincidentally, just around the time works published in the US are no longer in the public domain). The topic was selected for its contemporary relevance, but also for its fit within our selection guidelines.

Material have been selected from 9 libraries and 1 museum.

The collection contains:

3000 books, pamphlets, serials

And about 10,000 pages from 23 different manuscript collections

These same principles of selection can be also be seen in our second collection, just released last week:

Immigration to the United States from 1789-1930.

Immigration, like Women Working, was selected for its contemporary resonance. And with the recent debates over immigration legislation, border fences, and working visas, the materials in this collection are even more salient than we imagined when we began.

The collection contains about 2000 published titles, and over 6000 photographs  
It was selected from 10 libraries and 1 museum.

So let's look back at that quotation from Ranganathan. This time the second half.

“...Nor is it to forcibly feed them on books of your choice. It is to help them; and, to help anyone is to co-operate with him in carrying out his own plans and wishes—to help him to help himself.”

We can see that it is important to create a logical subset so that we don't dump materials on our users, however there needs to be a diversity, within that subset, not just of formats, but of the character of the materials. We don't want to define the outcomes of our patrons research through our own bias. Yet, there is bias inherent to any library collection, therefore, we try to cover the gaps in Harvard's collections by linking from OCP to other digital collections outside of Harvard.

In Women Working we also partnered with Cornell and made extensive use of links into their collection of Home economics materials. In immigration, we rely on NYPL's Schomburg Center to cover the African American Migration Experience because they represent it far better than Harvard's holdings could. So, accounting for bias, we still need to select materials that accommodate the needs of diverse users: economists, political scientists, religious scholars. Not force feeding readers with books of our choice, as Ranganathan puts it, means not just

selecting the expected. Women Working, for example, has much more than materials on women, it has a lot of primary sources from the progressive era in the United States. Child labor, labor legislation, housing and living conditions are all well covered. We also gathered a number of important labor history resources for this collection, including a complete set of the NY Factory Investigating Commission reports, which were published after the Triangle Shirtwaist Factory fire.

As an aside, only fragments of these reports existed in any single library at Harvard. But, because we had the luxury of selecting materials from across the Harvard libraries, we were able to digitally assemble a complete set.

We also selected materials, like trade catalogs, for women working that not only have some visual appeal, but illuminate the lives of working women.

- What were they buying, for example
- And what was being marketed to women
- In many cases the trade catalogs that we have digitized also illustrate the goods that women were employed in creating
- We have lots of materials on the boot and shoe industry, for example, so we thought it would be helpful to show students what the shoes they were making looked like, and what they cost

For Immigration, we focused a lot on the social sciences. We digitized the entire Social Ethics Collection from the Fogg Art Museum, which was put together in the early 20th century to document social conditions and social services around the world.

It contains photographs, like this, of schools, hospitals, housing conditions around the world. Many of these photographs are by famous photographers like Lewis Hine and Jessie Tarbox Beals, but many are anonymous. The social museum collection also contains a lot of data, like this image documenting the activities and expenditures of the Children's Aid Society of Boston, or this one discussing the nutritional composition of dairy products. In selecting for immigration it was also important to tell ALL the sides of the story.

Apparently, Harvard in the 19th century was not a big advocate of immigration to the United States. It was, in fact, home to the newly founded Immigration Restriction League that included the likes of Charles Warren, who later founded the Charles Warren Center for History at Harvard. So, in the spirit of honest disclosure, we have digitized the complete minutes of the League as well as a series of surveys they conducted asking respondents from each state to list the "types" of immigrants they would like to see from each ethnic group.

Once materials are selected for inclusion in a collection, they are digitized. This involves various combinations of imaging and OCR, and also the creation of structural and descriptive metadata. We are not just creating digital copies, but digital preservation masters that are described and registered in OCLC's Registry of Digital Masters. Our goal is to digitize these materials in such a way that they don't need to be digitized again. In the context of Ranganathan, our principle about production standards incorporates both his first and 4th laws.

Books are for use  
And  
Save the time of the reader

By making digital reproductions available, we are facilitating simultaneous use by large numbers of users. We save readers time by accurately describing materials and making them full text searchable. By making certain that our digital reproductions are faithful, we insure that the original rarely needs to be consulted. The actual production process is very detailed and I would be happy to answer any questions at the end if anyone would like to know specifics, but, one of the things I would like to highlight, is that a number of the steps in the production process deal with the description and cataloging of the materials being digitized. Books cannot be used, unless they can be found. Everything that we digitize is described, and linked to, from one of Harvard's library catalogs (either HOLLIS, for textual materials, or VIA for images)

We are creating library materials and they should be described in library catalogs, so for each item in a collection, we either create a new catalog record or improve the old one, so that we accurately describe that which we are digitizing. But while cataloging and description are part of the production process, they are really about helping the user efficiently find what they are looking for. In other words, it is about access. Library catalogs are one way of connecting the book and its user, but how do we connect users and books in a web-based environment. Sending people into a catalog with 10 million records to find a few thousand digitized things is once again sort of like dumping down a mass of books and tell readers to help themselves.

So, how do we solve this, we design web sites, right? But if we want to save the time of the reader, abiding by Ranganathan's 4th law, we need to design web sites to anticipate the needs of the user (and this is what we, as librarians, have been doing all along). At OCP, we have tried to do this by providing contextual materials on our sites to quickly and easily access subsets of the collections. As an example, our women working home page we present featured materials. Both collections contain pages that highlight people important to the topic in some way, or organizations, often whose papers we have digitized, or themes important to the collection. These pages are great ways to browse the collection. They are for the user who doesn't necessarily know what they are looking for.

They provide ways of breaking the collection into small, usable selections. Each web site also provides a way for users to access materials through a list of pre-determined topical keywords. These pages also serve as a sort of index of what is contained within each the collection. We also provide methods for searching the collections: We have a fielded metadata search, which searches across the both the image and text collections, but also allows you to limit to either format; or we have a full text search, that searches across the complete texts of all of the printed materials.

So, all of this web site business can seem really intimidating sometimes. And there are moments when I think that it would be a heck of a lot easier just to leave all of this at the point of describing things in the library catalogs and letting people limit their searching just to digitized materials. But creating these web sites are really no different than the kinds of paper pathfinders that libraries have been making to help users. First they were paper, then they were web-based, but it is still essentially the same activity--helping users navigate collections.

So, what is next for OCP. We are currently in production for our 3rd collection, Contagion. Which will be available in 2008. It will cover significant episodes in the history of contagious disease and medicine.--events like the creation of the smallpox vaccine and the debate that ensued over vaccination, or the discovery that malaria was spread by mosquitoes. After contagion, we are doing a small Islamic Heritage Project that will include Koranic commentaries and a number of maps. But beyond OCP, my interests also lie in looking at what is next in this whole field.

Ranganathan writes:

“It is no wonder that, when the library has been extending its scope, changing its outlook and altering its very character and functions, there should not be adequate understanding among the public as to what has been going on.”

So, what has been going on, and where is it going next? Currently, we find an establishment in our midst—Cyberspace—that through its primary information portal, the World Wide Web, provides opportunity for cultural heritage institutions to relocate from confining physical place to a digital analogue of libraries’ original inclusive identity. For this aim it is important to remember two definitions of the word library: “(a) a great mass of learning; (b) the objects of a person's study” (OED) because it is precisely this definition—of library as an interactive space for the creation of knowledge—that the Web so well enables. We can again render libraries a space not just for disseminating information, but technological laboratories rendering information into knowledge.

So, if the movement of cultural heritage onto the Web signifies in some respect a return to its roots, what has become of the librarian? Will those long associated with the organization, storage, and retrieval of information retrieve their identity

as philosophers and scholars? If librarians specialize in ontology and philosophers epistemology, will the Web facilitate a cross-disciplinary reunification? Current library projects reference such a unity only semantically, yet it is essential these disciplines truly collaborate to best harness the tremendous potential the Internet embodies.

Today, cultural heritage professionals mostly react to innovations in Internet technology. We respond to it as a mechanical technology, using the Web as a glorified photocopier to simply disseminate information the way Walter Benjamin understood the function of photography in his *Work of Art in the Age of Mechanical Reproduction*. Why should a profession that has long played a leadership role in the design and creation of ontological systems and the crafting of epistemological doctrine suddenly take a back seat to others just because the format of delivery has changed? How might we engage Cyberspace as a postmodern intellectual domain where we can again inquire about the transformation of information into knowledge and knowledge into political power.

History reveals a theme of private individuals creating public access to knowledge: Muhammed Ibn Al Ahmar, Gutenberg, King George, and Jefferson. Tim Berners-Lee is in many respects the torch-bearer of this tradition, but the outcome of his work will be far more powerful if we re-unify the librarian with the scholar and philosopher, and 'public' cultural heritage institutions with 'private' information technology companies. If this does not occur soon, we are in danger of information privatization on an unprecedented scale.

So, in conclusion, I would like to leave you with a question. Along the way, a lot of things have stayed the same, but on the surface, a lot of things have changed. The library at Alexandria has certainly changed a lot on the outside...But where will we librarians take this profession for our future?