

We have been at this digitization thing for quite a while now. The library of congress began the pilot project that would later become 'American Memory' in 1990. By the mid-1990's there were hundreds of digitization initiatives in libraries around the world. Looking just at the digitization projects initiated by libraries, (so excluding for a moment the current projects by Google, the Open Content Alliance, and Microsoft) the Library of Congress now provides access to over 9 million objects, Harvard provides access to another 7 million. Gallica has hundreds of thousands. OIAster, the aggregation of OAI harvests from over 900 collections, provides access to over 16 million digital objects. Taking into account all of the digital materials at the small libraries, museums, and historical archives, I estimate that there may be as many as 50 million digitized rare objects freely available on the web today. And that number is only going to grow.

So we should have learned a lot. And we have. But we still have no collective strategy for access to these materials – even across the academic libraries. And that is why I think – with a couple of exceptions – these incredible materials that should be absolutely revolutionizing scholarship in the humanities are often somewhat ignored.

So, today I would like to propose a strategy, or at least the beginnings of one. The good news is that I think that we are already a long way down the path of this strategy, but we just aren't aware of it yet.

So, the structure of today's presentation will be as follows.

First, I am going to talk about the state of digitization projects by laying out a few of the essential differences between the large and small-scale projects in which we engage today. In doing so, I will over-emphasize the divide, but I think it is important to make my point.

Then, I will speak about why I think that libraries should focus their energy on digitizing and providing access to their unique holdings and just let the big projects—google, open content alliance, etc—do what they do. And then I will lay out the conditions under which I think this strategy can be successful and the current barriers to success.

So, let's look at the status quo in digitization projects, starting with the large-scale projects. (When I am speaking of the large-scale projects, I am thinking mainly of Google books and the Open content alliance, because those are the projects I know the best)

These projects are characterized by

- 1) their large size

2) their production speed, which is fast;

3) their low per page costs;

but also by their findability.

These projects, and the materials within them are for the most part, very easy to find.

So, how do they achieve this?

1. They have no intellectual input into selection, which is not to say they don't have a selection policy, just that it is not based on intellectual criteria. It is based on format, condition, size, availability, but not whether it fits a topic, or is part of a collection.

2. these projects – google in particular – are after the text. The page images are a by-product. These are projects to make text searchable, **not** to build collections. Building an enormous index of the written word is an important goal, but it is not building a library.

These are tools for discovery, for information retrieval, but not necessarily for use—at least not in the way that we are used to using library materials. They are only part of what it is to be a library and I strongly resist calling them digital library projects.

Looked at another way—and to borrow a concept from Paolo D'Iorio—these projects do not meet the conditions for the possibility of web scholarship. They do not easily allow for Quoting, Consensus or Preservation / Dissemination.

What are the characteristics of the **small** projects?

1. They are well crafted with attention paid to providing access to the object, not just the intellectual content. They more often include rare materials. They are well organized, and browsable.

2. They are also expensive.

3. But they are hard to find! They are hidden. What makes me say that?

I recently undertook some analysis of inlinks to digital project web sites.

Inlinks are defined simply as any link to a web site

If 4 different web sites link to Web site A... it is said to have 4 inlinks

Inlink analysis is similar in many ways to citation analysis... Except instead of measuring how many people cite a resource, you are measuring how many people link to it. And in that way you are taking a measure of how many people might be using a site.

The basic idea of the project was to count the number of inlinks to 289 library digitization web sites and to graph the resulting numbers.

In the end there were 2 sites – the Library of Congress' American Memory and the Project Gutenberg – that had a significant number of inlinks. And the dropoff after that is steep.

The fourth most linked-to site contains only 16% of the inlinks of the most linked to site...and the tenth most linked to site contains only 4%.

There are a number of reasons that could account for the steep dropoff in inlinks

- an obvious one is scope,

The top 5 sites are quite general and quite large, but I am not convinced that scope alone can account for it:

I think it is much more likely that so many of these sites aren't linked to because they aren't used, and they aren't being used because they aren't being found

So, returning to the comparisons between kinds of projects. If the characteristics of these small projects are that they are well crafted, and the objects are well described, but hidden, how are these characteristics achieved?

1. Through an understanding of users.
2. through an appreciation and understanding of books and materials
4. But how did they get hidden? Through a forgetting of what it is to be a library.

So, if we buy into the fundamental difference between these kinds of projects, what next? Well, when Google announced its library project, I think it is fair to say that there was a bit of a wave of panic across libraries, as if we were going to be put out of business. But what we should have heard was a collective sigh of relief... as if the heavy lifting is being done for us and now we can focus on the important work.

If the large digitization projects are changing research practices, then libraries have the potential to change scholarship by digitizing their unique and closely guarded materials AND making them freely available on the web.

Research, which is predicated on being able to retrieve information, is in many ways the domain of the products and services that libraries offer – journals, databases – the basic formats of configured and presented data.

BUT, creating a space to take all of that information and turn it into new knowledge – THAT is the domain of libraries.

Seneca once wrote: 'it does not matter how many books you have, but how good they are.' This idea, this commitment to quality over quantity, is one of the most significant factors binding these smaller projects together

Other than this commitment to quality, there is very little else binding these small disparate collections together.

Well, so what? Why should we care if there is no confederation of projects nor 1 single point of entry? Well, this goes back to the issue of access.... The problem of these collections being hidden.

If we look again at the graph of links to digitization projects, and we turn it on its side, we basically have a classic demand curve.

And on top of that you can layer Pareto's rule, or the 80/20 rule. In libraries, this is translated as, "20% of your collection will meet the needs of 80% of your users." But the Internet has done some very interesting things to the 80/20 rule.

Chris Anderson has made what I think are some pretty persuasive arguments about the economic viability of the 80% .. or the long tail .. of the demand curve. Amazon.com is his classic example. He argues that Amazon can afford to essentially give away the things on the best seller list, because where they make their money is by giving you access to the publishers back catalogs, or the titles from smaller presses. They give you equal access to the 80%

So, if we equate the more common holdings in libraries with the bestsellers, and the unique materials and archives with the long tail from which Amazon has become economically viable, the question becomes,

Will this model work for libraries? If people start getting access to that 20% from somewhere outside of libraries, if users are going to Google first, Can the potential aggregate size of the many small "markets" for these rare materials (the 80%) in fact rival the demand for the 20% of the collection that we know is popular?

I think they can, both in quantity of demand but perhaps more importantly in terms of quality—thinking back to what I said about the collections of rare materials being at the center of scholarship, not just research.

So, in my introduction, I said that we were already along this path that I have laid, because I think that libraries have largely ceased the digitization of their common holdings, or duplicating the things that Google and others are doing. And I think I am preaching to the converted here to say that we should focus on digitizing rare materials.

BUT, there is a significant caveat – a significant barrier to success – and that is access. In order to replicate the Amazon model, the long tail portion of a collection needs to be just as accessible as the high-demand materials.

On Amazon, It is no more difficult to purchase a copy of a book from an obscure publisher with a small print run than it is to purchase a best seller. But libraries are still using the old bookstore model.

It's pretty easy to get a copy of Hamlet, pretty much any bookstore will have one. But if you want to find a copy of 'The Inanimate Tragedy,' you probably need to find a bookstore that specialized in the works of Edward Gorey. And so someone who is dedicated to finding a copy of 'The Inanimate Tragedy' will either go persistently from store to store, or will already know where to go because the last thing they bought was equally hard to find.

We need to level out the extremes of access – to make the rare materials as easy to get to as the common ones. And once they are digitized there is absolutely no reason that this can't happen. But in large part, I don't think it is.

Lynne mentioned yesterday a report that they commissioned about the information seeking habits of the Google Generation. One of the conclusions of this report was that although very technically literate, this group is really not very good at finding information. And there is certainly a role here for libraries to teach information literacy, but I think that we have to take equal responsibility for not always making these things very findable.

I would really like to see an amazing revolution in access to primary resources. When the dead sea scrolls are digitized, I would like to think that not only the amount, but the quality of the scholarship produced about them will increase sharply from the times when there were just a small handful of people who had access. And I think that libraries should be playing the central role in reconfiguring access to these materials. They should be leading this revolution.

So, what are the solutions? How do we get there? How do we facilitate a revolutionary change in access, because I don't think the 'build it and they will come' model is working.

I think that the answers are already built into the principles of librarianship – we just need to remember what it is to be a library.

To really understand the essential qualities of libraries—academic libraries in particular—I like to turn to the work of SR Ranganathan.

Ranganathan was a mathematician, a librarian, and perhaps most importantly, a philosopher.

He worked in a number of libraries in India in the first half of the twentieth century and championed what he called the 'open access' movement in which he argued for open stacks in all libraries.

He is most famous for writing the 5 laws of library science.

His five laws are:

- Books are for use
- every reader his or her book
- every book its reader
- save the time of the reader
- a library is a growing organism

I like to look beyond just the laws, though, and at his very broad interpretation of them. Because Ranganathan understood what it means to provide true access to library materials. and this is just as applicable to opening up access to digitized rare materials as it is to opening a public library in India.

So, Ranganathan had something to say about every aspect of libraries.

On the appropriate height of library shelves he said:

“no rack shall be higher than what can be reached by a person of average height, while standing on the bare floor”

What on earth does that have to do with building online digital collections?

Well, to me it means:

- don't make it difficult for people to get to the materials
- build easy to use interfaces
- understand who your users are

About the opening hours of libraries he said:

“In no country where the concept, BOOKS ARE FOR USE, has taken root in the Public Mind, will any library be allowed to close until the majority of humanity go to bed and cannot use it”

which I take to mean:

build a stable infrastructure..

if users come knocking at your door too many times and you are closed, they will stop coming back.

This does not mean you should not take risks, or that you should trade usability for stability. Just perhaps that this is not the place for the ‘fail early and often approach’ that can be quite common in the technical development environment.

Users need to develop a level of trust in these collections.

On providing appropriate reference services Ranganathan wrote:

“The majority of readers do not know their requirements, and their interests take a definite shape only after seeing and handling a well-arranged collection of books”

Online, this means build collections that can be easily navigated and allow for serendipitous discovery of materials. This is one of the primary differences between providing information retrieval services and building a library. So much of what it is to build a space to facilitate the creation of knowledge comes from being exposed to that which you were not originally seeking.

And on the location for a library, Ranganathan said:

“a library that is keen about its books being fully used will plant itself in the midst of its clientele”

which leads us to marketing...

I think we all engage in digitizing rare materials because we are keen to see them used...

So this means Go where your patrons are

And That entails not only understanding your users... and not just on the surface -- their fads and trends-- but understanding what the essential qualities of scholarship are...

We need to understand users, and we need to be both transparent and outspoken about what it is we are trying to do.

Because, to close with another quote from Ranganathan, (and remember that he wrote this in 1931, but I think it is equally true today)

“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.”