My Discussion

- Data policy and the role of libraries
- Challenges: workforce and infrastructure
- Deep collaboration, trust, and proving value
Today’s question....

What is the Library’s role with “big data?”
What is the Library’s role with any information?
Core Library Roles

- Identify and **select** valuable resources
- **Organize** those materials
- **Describe** those materials
- **Preserve** those materials
- Provide **access** to those materials
  - Help others find them
“Big Research Data” Examples

- What researchers create, such as:
  - Video or other time dependent recordings of human response, physiological response, eye movement...
  - Predictive models of climate change impacts on decadal scale
  - Large-scale heterogeneous text corpora
  - The long tail of “small data”
The policy landscape has become “unstuck.”
- White House OSTP RFIs (2009-11)
- NSF Data Management Plan policy (2011)

Librarians are seen as partners for solutions.
- Consultation
- Infrastructure
Meet funder requirements for data management plans.

The DMP Tool allows you to: 1 2 3 4

DMPTOOL is a service of the University of California Curation Center of the California Digital Library
Copyright © 2010-2012 The Regents of the University of California
Privacy Policy | Terms of Use | Photo Credits

https://dmp.cdlib.org/
Some jobs posted since June 2011

- Data Management Consultant
- Data Mining Consultant
- Data Research Scientist
- Data Services Librarian
- Design Data Librarian
- Digital Archivist
- Digital Collections Strategist and Architecture Librarian
- Digital Humanities Design Consultant
- Digital Records Archivist
- Manager, Data Management Services
- Research Data Librarian
- Research Data Management Coordinator
- Scientific Data Curation Specialist / Metadata Librarian
- Scientific Data Curator
- Social Science Data Consultant
Inflecting our roles

- Identification & selection, but **according to what value?**
- Organizing collections, **but of what sort?**
- Describing **according to what principles?**
- Preservation, **but in what manner?**
- Access, **but also analytics?**
What does the researcher care about?
“Where can I put my stuff?”
The value proposition offered by libraries is longevity and durable access, not “storage.”
Towards Deep Collaboration

- Among libraries through consortial arrangements
- Commercial-public
- With professional academic societies
- Current examples:
  - DataNet (NSF) and NDIIPP (LoC) programs
  - HathiTrust
The Big Bet

“The bet is this: by curating and preserving this data we will allow for the study of ‘big questions,’ answers to which will benefit society. But how do you measure that? When will we know?”

Brian Schottlaender
University Librarian
University of California, San Diego
Thank you. Questions?

Mike Furlough | mfurlough@psu.edu
I would like to thank the organizers for giving me the opportunity to speak with you all today about Research Libraries and Big Data. Before I get started, I’d like to acknowledge the context from which I am speaking.

I am the associate dean for research and scholarly communications at Penn State University. We are a research intensive university; last year we spent over $800m related to research, and about half of that came from federal funds. We are not the largest research university but we are big.

Research is conducted at all colleges and universities, but most of what I say will be focused on libraries more like mine, e.g., affiliated with large research universities.

There is great variation among them, but we have some common characteristics:

We build collections and provide services to a **local** audience. Those services range from research assistance to instruction to operating production facilities in our spaces.

That audience is broad spectrum of students and researchers from dozens of fields. We try to serve them all equitably, but that’s not really possible.

Those researchers produce data big and small.

We have been experimenting with various scholarly communication efforts over the last decade, ranging from campus publishing to collections digitization to the development of e-research support for data management.
SLIDE 2 | MY DISCUSSION

So as I talk to you today, but I’ll be talking from that context. There are major research libraries that do not serve universities, such as Library of Congress and the National Libraries, but they have different roles. I don’t want to ignore them but for the sake of scope, I’ll focus mostly on the university context. Towards the end I’ll talk about large scale efforts that happen across institutions.

I was asked to give you some background on what has changed for us regarding big data over the past two years. I’ll answer that by focusing on some shifts in national policy and how that has galvanized libraries and librarians to provide new services.

Those services bring with them many challenges, but I will talk about two: workforce and infrastructure.

I was also asked to speculate about where things are heading in the next five years. I’ll respond to that by talking about the need for deep collaboration and the challenge of proving the value of what we propose to do with data.

SLIDE 3 | WHAT IS THE LIBRARY’S ROLE RE BIG DATA?

If this is the question you’re wondering about, I have a response that sounds kind of flippant, but truly isn’t.

SLIDE 4 | WHAT IS THE LIBRARY’S ROLE WITH ANY INFORMATION?

We address big data in the same way we address any information: for the long term. We are very good at thinking about the long term.
We are not necessarily good at thinking about cost-effectiveness, and for that matter, neither are researchers.

Trends, challenges, responses apply equally to “small data” as well as “big data.”

But its nature and policy governing its disposition require deeper collaboration with many parties.

**SLIDE 5 | CORE LIBRARY ROLES**

These are our core strengths. This is how we operate. And the most important of all of these is ACCESS. None of these other activities has value in the absence of the provision of access. So the real question we are always thinking about --- and assuming -- is that we have to find ways of helping our community of users gain access to all kinds of information.

What librarians are increasingly asked to do is to involve themselves at earlier stages of the information lifecycle. In other words, traditionally, we have focused on handling the products of research that have in some ways already been filtered through publication. That’s not entirely true in all cases, but that’s true for the most part.

NOW, we have begun to recognize that if we want to provide good, solid, valuable access, we need to work our way into the research process at earlier stages.

**SLIDE 6 | BIG DATA EXAMPLES**

Recently I had a conversation about “big data” with a researcher and he offered a definition of big data as “anything that breaks your common tools.”
By the definition offered by my colleague, libraries started dealing with big data when we automated our circulation systems and OCLC began the development of WorldCat, which now holds descriptive records for over 1 billion holdings in 72,000 libraries worldwide.

We see big data in nearly all fields. Even historians and english professors have data. The examples I have on the screen are all real. Some of them you are familiar with, at least in concept, others maybe not so much.

One point I would like to emphasize is that we often think of “big data” as the result of large scale, collaborative projects: the Large Hadron Collider produces big data.

But when you look at funded research at a major university you see that the distribution of awards follows the Long Tail: we have a few very high profile high $ projects. But the majority of bench scientists, social scientists, and humanists toil away with limited funding—or no funding—producing research data that in the aggregate might very valuable, but which is also, in the aggregate, very large and very challenging because of its diversity and heterogeneity.

For much of the past decade, some leading libraries had been working exactly in this area: the management and durable access to research data. Johns Hopkins, Purdue, UC San Diego and others were part of NSF or Library of Congress NDIIPP grants to focus on data preservation, to build infrastructure that could be used broadly.

**SLIDE 7 | LIBRARIES AND THE POLICY LANDSCAPE**

But the single biggest change around “big data” and libraries in the past two years has been brought about by shifts in the federal policy
landscape. These shifts have put a greater emphasis on PUBLIC ACCESS to the results of research funded by the federal government. These began with NIH’s policy that requires research written as a result of NIH funding to be deposited full-text in PubMed Central within 12 months of publication.

Next, the Obama administration showed an interested and their RFIs generated a significant response from libraries, publishers, and professional societies.

- OSTP Public Access Policy Forum (December 2009)
- Public Access to Digital Data: Public Comment (November 3 2011)
- Public Access to Scholarly Publications: Public Comment (November 3 2011)

The most important of these, however, was NSF’s announcement that it would require researchers to write and submit with proposals a “data management plan” that address the data to be created, how it would be safeguarded, how long it needed to be retained, and under what conditions would anyone have access to it?

Now we all had a reason to consult with researchers. And the faculty and administrators at our campuses wanted us to do that.

**SLIDE 8 | DMP TOOL SLIDE**

The skills that librarians have in organization, documentation, description, and preservation are all skills that relate to the effective management of data produced for the purposes of research. So we have been seen as natural allies.
While some libraries were very well prepared to expand their services in response to the NSF policy, many others were not, but knew that this was an opportunity to support their faculty.

So, many of us put together data management consulting teams. We worked diligently at educating ourselves. Our trade organization put together some tools and resources to help educate librarians about data management. Much of what the NSF was asking researchers to do had already been requested of researchers in Europe. So we have had best practices and resources that we could draw on, and which we could put in front of researchers.

Several libraries and organizations worked together to develop a new toolkit, call the DMP Tool. This resource collects the data management policies of funding agencies and private funders, including foundations.

It also provides some standardized questionnaires to guide the development of a data management plan that could be submitted. Over 75 institutions have set arranged for their users to be able to use the DMP Tool through their local authentication mechanisms.

This does not provide a push-button solution. But it provides a common set of questions, many of which researchers rarely, if ever, think about in advance of their research.

- The University of California Curation Center (UC3) at the California Digital Library
- DataONE
- Digital Curation Centre (UK)
- Smithsonian Institution
- University of California, Los Angeles Library
- University of California, San Diego Libraries
- University of Illinois, Urbana-Champaign
• University of Virginia Library

SLIDE 9 | SOME JOBS POSTED

This has also resulted in small explosion of hiring. This is a small sample of job postings that have gone out since the summer of 2011 (six months after the NSF policy went into effect).

There is a mix of roles here:

• Some are set up as “translator” roles—able to help connect users with IT or other staff, and to provide some training and education.

• Some are set up as real “archivist” roles—these may or may not work with research data. They may be working with electronic business records of their universities. But the principles of archiving, and the forensic skills that go into archiving of electronic records, are applicable to research data as well.

• Some of the newer titles here are research consulting roles. The Data Mining Specialist is likely to be someone who helps someone conduct their research, to use tools that can facilitate the analysis. The Digital Humanities Design Consultant will help humanists frame research questions in relation to computational methods, e.g., help a historian perform geographic analysis.

• Note also that some of these are focused on some specific knowledge domains, like architecture, or broad ones, like social sciences.

SLIDE 10 | INFLECTING OUR ROLES

These new roles raise some significant questions about our workforce and about how we carry out our core activities.
First of all, the preceding list suggests that Libraries are only trying to hire for new sills and roles. This is not true.

We are also working to re-focus our current librarians—especially those in our cataloging departments, and our subject librarians—on these issues. Many of these librarians are used to working with researchers to help them acquire and find information resources.

We are not historically trained to understand, in a deep way, how information and data are created and what mechanisms should be put in place FROM THE INCEPTION OF THE PROJECT to make it easier to collect, organize, preserve the data being produced.

However, librarians who have a background in cataloging and metadata functions are exceptionally good at conceptualizing relationships among data and are often just the people you want at the table when someone is trying to develop a schema from scratch.

We are seeing trends towards more specialization, research backgrounds in specific disciplines, and a de-emphasis on traditional library training as the Union Card. However, we are also seeing LIS programs with strong archival programs re-defining the focus of those programs to emphasis data curation and preservation. UNC-Chapel Hill and UIUC are two that come to mind, but there are others.

We are not yet sure how we understand POTENTIAL VALUE in data. We have always bought books, journals, databases from organizations that pre-filter through peer review or economic factors. Data moves us to a world in which value is not dependent upon scarcity of access, or pre-filtering. But we still have to come to terms with new ways of determining future impact.
SLIDE 11 | WHAT DOES THE RESEARCHER CARE ABOUT?

ULTIMATELY this is the key question. How it is answered depends upon which side of the process are you on at the time it is asked. Are you the creator, or the potential users?

I think we hear more from the producers than the users. It may be that the policy shift has pushed us in that direction. It may be that a lot of use is still theoretical.

Researchers tend to be narrowly focused on the domains, their research, their questions. It is often hard for them to imagine how anyone else might use the observational data they have collected.

But what looks like noise to one researcher might be the critical data to solve another’s problem. But many researchers don’t believe their noise is valuable.

Do we save everything? And if so, in what way? Or do we only save the inputs? Or some subsets? Or just the bits, and good luck to you in the future if you want to read this obscure file format. How can we effectively curate this data so that it becomes more valuable, and doesn’t hinder other potential uses? It is very difficult to predict future use.

One thing we have learned from consulting with researchers on data management plan is that Researchers who produce data are often going to focus on one key question.....

SLIDE 12 | WHERE DO I PUT MY DATA?

That is to say: for many researchers, data management is all about storage. Storage may be called “archiving” by some technologists, in
the sense that you move some data from a site of immediate access to a near-line access method, such as remote tape, or whatever. But that’s not what Libraries mean by archiving. Archiving means arranging the data, documenting, describing, and making accessible the data.

But if you frame data management only as a matter of storage you miss the true intent of the NSF policy, which was to enable new research, public access to research funded by federal dollars, and enabling reproducibility for research.

Outside of this frame of reference, if you have big data, and you need storage, then you might think the answer looks like this....

**SLIDE 13 | CLOUD**

......ah, yes, THE CLOUD. Which is a great metaphor because that single word tends to obscure so many different issues. We’ve overused that phrase, but we can easily point to what it might mean by naming vendors and particular services.

It is definitely the case that Amazon Web Services and Google Cloud Services can provide, in some cases, cheaper access to computing cycles or to storage than a university’s existing IT services and data centers. Some universities are exploring how cloud services fit into their local infrastructure. You could imagine a role for University IT services—if not libraries—in brokering access to some of these cloud services, and in helping researcher use those services if they need to do so. And, as I’ll note in a minute, it’s important for universities to start thinking about developing their own multi-institutional clouds.

And the reason is that we are not Amazon....
SLIDE 14 | THE VALUE PROPOSITION

“...there aren't other vendors for preservation--just for storage services, like Amazon. Competition would be non-sensical in the digital preservation space. Digital preservation only makes sense in the context of other services that support access, analysis, and re-use.”

David Minor, Program Director, Chronopolis, University of California San Diego Supercomputing Center and University Libraries.

David’s point echoes one I made earlier: preservation makes absolutely no sense in the absence of access. It’s why we have tended to emphasize words like “curation” instead of “preservation.” Curation is what we do—organize, enhance, normalize, preserve. And vendors are not necessarily in it for the long term. Libraries are.

So, as I’ve said, we have the skills sets, the backgrounds, the values, the mission to engage with you and others on problems in “big data.”

But that doesn’t mean we can solve the problem independently.

SLIDE 15 | TOWARDS DEEP COLLABORATION

So when I am asked: what are we likely to see in the next five years? The best answer I can give you is: More cooperative/collaborative efforts.

The infrastructure challenges are too great for single institutions, let alone libraries to solve independently. But while all of us are not-
for-profit organization, we all compete: for research funding, for faculty, for students and their tuition funding. Collaboration does not always come naturally.

But we have some groundwork in this area.

The DataNet awards from NSF have been made to several institutions, and in all of the ones I am aware of the PIs or co-Pis are Librarians or library science researchers. The goal of these efforts has been to build large scale networks to provide support for effective, distributed data management and access. These activities have included building repositories and developing services that hook to publishers, but it also means investigating researcher needs and habits, developing powerful analytic tools, or even developing enhancements to simple, and commonly used tools, like Microsoft Excel (the number one data analysis tool used by researchers).

HathiTrust is a cooperative investment made by over 30 university libraries to provide a preservation and research access service for the materials produced in mass digitization efforts like those of Google. Based at Michigan and Indiana University, it holds a huge corpus of textual data that researchers can access for “non-consumptive use,” e.g., Text Mining

10,588,112 total volumes
5,570,069 book titles
275,879 serial titles
3,705,839,200 pages
475 terabytes
125 miles
8,603 tons
3,268,765 volumes (~31% of total) in the public domain
On this screen you see several examples of the kind of collaboration I’m talking about.

DataOne, Data Concervancy and Data to Insight Center have all been funded by NSF through DataNet.

MetaArchive, a dark preservation archive making use of peer-to-peer technology, and GeoMAPP, focused on the preservation of local/state government spatial data were both funded by Library of Congress’s NDIIPP program.

There are new ones emerging all the time.

The Library of Congress’s NDIIPP program has given rise to the National Digital Stewardship Alliance, which includes members from academia, industry, and government, convened to work on Content | Standards and Practices | Infrastructure | Innovation | Outreach.

DataCite is an organization founded by several European national libraries and including some North American libraries to work with the publishing industry to develop the mechanisms to assign persistent, unique identifiers to datasets so that they can be cited.

The Digital Preservation Network is the newest and perhaps the most ambitious. Over 75 universities have contributed at least $20k to fund an investigation of a NATIONAL preservation network. Two strategies of DPN worth noting are:

Preservation in enhanced not just through replication/storage, but by replicating/storing in a diversity of repository infrastructures. Chronopolis at San Diego, Hathi Trust, the Data Conservancy all have slightly different approaches to data storage and transfer. Relying
on that diversity reduces the likelihood of loss/failure based on a single repository type.

Second: Preservation of cultural and scientific artifacts is a library issue, but it’s also a UNIVERSITY issue. It has been spearheaded by both Librarians and CIOs, but they have engaged our presidents and provosts, and attended their meetings to make the case for their effort.

In other words, they have raised the stakes and made it clear that preservation of big data and our cultural heritage is a joint responsibility. And that responsibility extends to the national libraries and agencies that many of you may represent.

Ultimately, these projects make it clear that big data is not a technology issue, but a social, economic, and political issue. Addressing it involves establishing new, deeper, more complex trust relationships across domains unlike what we have seen before.

**SLIDE 17 | BRIAN SCHOTTLAENDER QUOTE**

But to what end?

In order to engage those players, gain the investments needed, we have to keep focused on what we are trying to create for the greater good. But note that the bet placed here can’t be won in a short timeframe. In fact, the metrics for success are not clear. And they require us to invest in ways that do not have immediate payoffs. That’s fine—libraries in particular are not always about the immediate payoff. But if we truly must collaborate in ways that go deeper than before, it means that we can no longer afford to focus solely on our local context. It will mean that research libraries lose some of their identity based on distinct collections and instead begin
to focus it more on how we enable and support the creation of big data that can answer big questions for generations to come.

**SLIDE 18 | CONCLUDE**

Thanks very much. I look forward to your questions.