We created a digital curation program, although with no direction; I laid out the alternatives for hosting or not hosting other departments’ data; and then started exploring all the avenues that opened up.

This is a big departure from what many of us are used to, and there are a lot of things that have to be covered at the start; so I am going to ease us in with a story first (it will seem to go down a rabbit hole, but it will give us insight), then stop for a definition of data, and then point out the connection to humanities and other disciplines not always thought of as concerned with data. With that foundation set, we’ll look at the various elements of data service in a library, and functions librarians perform that are relevant to data, and contemplate the administrative perspective of planning a library service. This is a progress report on our beginnings at UT Arlington; the title of this talk is “designing a library data service” because that’s as far as we’ve gotten, and we’re hoping for a discussion today that will share components useful to us all.

At the ASIST 2006 annual meeting, I heard a remark that changed my scientific life: that scientific research has moved from laboratory experimentation to gathering and archiving huge quantities of data—also, research data are now available for secondary analysis. Researchers have a general idea of how to organize and store files, as they have always done; but now, maintaining data through version changes, changes of storage location, description and association of metadata, and organization are becoming too important, and too complex, to be worth the time of researchers, when information professionals have exactly the knowledge necessary to ensure robustness of this change.

In 2011 I created an action plan, which was a grassroots mini strategic plan, for setting up a library data files service, comprising accessioning, reference, and information literacy acquainting people with the new importance of data and getting them to think strategically about it.
I had been preparing to approach our colleges and departments with offers of my services; and since the final project of my capstone course at UNT’s iCAMP digital curation program, last year--
--required me to go out and interview a data power user about that department’s data curation needs, I went there, suspecting and finding that their needs were extensive, beyond curation to a complete data program,* which I *could* shape according to digital curation.

The ten-second course in curation for librarians is (and we’re going to get back to this, so don’t worry about scribbling it down right now): selection, recording of metadata including details of people and organizations associated with the materials collected, placement into stable long-term formats, and designing access to ensure security and benefit the largest audience possible.

Archivists take a long view, planning for users being able to make sense of archived materials decades or centuries after the original staff are gone. How much that applies to research data is a good question; but the techniques of metadata collection, file format selection, and access interface design are clear and present needs.
When I talked to the director of research for our college of nursing (now the College of Nursing and Health Innovation), I could see that before we got to curation, they needed some professional advice on how to manage their information resources. They are academics, caregivers, and major collaborators, and their data household was getting chaotic and messy. They need at the very least some kind of data audit and a way of creating and maintaining their own records, and they probably could use an actual systems analysis (appropriate to their changing activity and the availability of personnel to do it) so they could be able to find things and have some procedures in place. Only after that could I think about curation.

But of course their primary expression of need was, hardly surprisingly, that they wanted me (or the Libraries) to provide them with some data sets, both for their graduate students to practice on for secondary analysis, and for real research and discovery.

Aside from this, of course, is the public service of data reference, navigation, assistance with file manipulation, and instruction on the whole culture of data.

That was the rabbit hole. See, we glimpsed a whole new world beyond immediate public services.

All this might sound vague, especially to those whose background has not included experimental research. If you *have* done research, you know what data is (or are—let's settle this, even if temporarily, right now. “Data” is a plural, even if common English usage is to treat it as a singular; you
do see it correctly as a plural sometimes; I compromise by using it as a singular when referring to the
general concept or the practices of management of it; “big data” originated in business where it refers
to the power and pervasiveness of data to predict or dictate conditions and decisions, hence the more
you have, the better you’re able to choke or run circles around your competitors): data are the records
you generate by observations, either of the experiment you’re doing, or of one part of the world as it
exists—you dig things up, or catch them, or see them in your scope, or observe people doing or saying
them, and you write down (or have an instrument print out) the count, or color, or words, or electrical
charges, or test answers. Usually you organize them into some kind of set or file. You might need a key
or legend to remember what they all stand for (that’s called a code book).

Thus we have the definition, as given by my colleague Helen Hough, “Data: elements that are discovered
or manipulated during the course of a research project.” They take various physical forms (and I was
too lazy to search the net for file photos of all of them), but they can be paper printouts, or mechanical
tracings on some sort of medium, but in this modern age they’re mostly Excel files or notes in Word or
ASCII, or images.

Data: elements that are discovered or manipulated during the course of a research project

Now this may seem more science-focused than your subject area; but computational technology has
gotten so fast and complete that it is able to explore and answer some of the questions we thought
were too soft or complicated for cold silicon: we now have digital humanities. We are using data to
analyze connections, influences, imitation and plagiarism, and just big counting projects in the study of literature and the arts. I am seeing how some of the information automation that started in the sciences is able to serve the humanities as more than just an electronic printing press.

I wrote a post in my Open Science blog (which is where I put items relating to data) a while ago about the way the humanities were picking up digital techniques.

So whether you’ve come over the path of STEM or arts and language, you have Excel files or notes in Word or ASCII, or images, if you’re dealing with original, locally produced, data. There are plenty of sets of data of all kinds available over the Web, especially in this age of Open Data; they may or may not be organized or cleaned of errors, but at least they’re electronic and you don’t have to worry about digitization—which we’re not going to get into here. One of your first responsibilities is selection, and often you can set out certain standards or requirements for what you’ll deal with.

So either way, you’ve got some spreadsheet files or images or PDFs. As a librarian, you can see the issues right off: are these from a reliable source, are they in an accessible, transmissible format, how should they be organized and accessed, are there any restrictions on permissions, are they in versions which have to be tracked for being up to date, and of course how do we provide reference or navigation services.
Traditionally, experimental data have been under the purview of experimenters, and libraries didn’t touch them. Actually there have been some sources of data that libraries kept, such as Census or government documents, or some scientific references. And we ensured completeness, comprehensiveness, currency, and accurate retrieval for our patrons, because we are information professionals. This is why we are the best people to manage data.

How much we will get to do is a question to which I think we can still influence the answer. Let's look at what we bring to the table and how we would look at working with data.

Traditionally we sum libraries up into collections and services (or maybe collections, facilities, and services, as well as administration and management). Thus we have the handout which starts out as a checklist, so organized.

Probably our first or most common question would be, where can we find, or help patrons to find, data resources. We'll get to pages of links in a second, but we might notice first the underlying question of, since a lot of data are available free on the Web, what does it mean to acquire things that don't cost anything or take up any space. As our free electronic resources policy implies, we do pay a cost for adding links to our server pages: we have to catalog or arrange for them to be added to our Web site; we have to re-check them periodically to ensure they still work, as mentioned in our free electronic resources policy. If they're not free, or are but within a government documents structure, we have to deal with purchasing, licensing, and data use agreements.
Who will find these resources? They're not packaged into acquisition catalogs that we can just select into our approval plans. They're likely to be found by disciplinary faculty and students, and my experience is that they'll turn up when you are reading and corresponding within a subject discipline. There are a few data search engine type resources. I keep two folders, which I label “Find” and “Finding.”

http://libguides.uta.edu/researchdata/datasites
“Find” is sites I have found, that are specific containers of content. “Finding” is search engines, wikis, blogs posting recommendations. I post some selections from both “find” and “finding” folders in my research data LibGuide.

Backing up a bit, I'm now going to move to curation, which I also group under collections. Actually the first item is not subordinate entirely to curation, but it definitely comes up there. As I talked to the nursing people, one of their sources of daily uncertainty was hesitancy about approaching other people for their data, just on the basis of whether it’s “their” data or it’s all in some kind of communal pool, especially after somebody leaves—who can still access the old files? They just aren’t sure. Who has copyright? Does the university have that, and does the university actually own and control the files?

There are data ownership policies, such as Johns Hopkins’, and there are also intellectual property policies such as Valparaiso’s, which cover copyright, patents, trademarks, and royalties.
But since many of these get into the grey area of whether faculty or university have the rights, and apparently grey areas can be firmed up by whoever asserts the rights first, which could lead to inconsistent individual instances within the same institution, my thinking is that we should be aiming for not a cut-and-dried policy but a philosophy of ownership, with principles of who can grant access, who can take stuff with them, who has to give up possession, who can claim copyright (there has been a tradition that non-faculty university employees cannot get copyright because everything they do falls under the legal rubric called “work for hire,” but that’s not a hard and fast law.) From such statements of principles, individual cases could be formulated or interpreted, and everyone would work from the same general understanding. Of course in the end we’d have to meet with intellectual property lawyers to make sure we were on solid ground.

Once that foundation is set, the first step of curation is selection—what to enshrine in your archives, or deal with in a temporary project Web site—and along with that, decisions about retention: is it a permanent accession or does it have a finite life span. Choice of what records are to be retained and curated is usually left to the curators, although regulations and funder requirements may impose restrictions or access requirements.
Selection (and retention)

- What serves the researchers
- What serves the University
- What is required for compliance
- Preservation planning
  - *(just anticipation—whole section on preservation policies & procedures comes later)*
- Policy when PI leaves UT Arlington

Considerations include:

- What serves the researchers
- What serves the University
- What is required for compliance
- Preservation planning (just anticipation—whole section on preservation policies & procedures comes later)
- Policy when PI leaves UT Arlington

Selection criteria can parallel library collection development policies—works collected and not collected, characteristics of the community—and are based on the “designated community” statement, which is an archivist term for a paragraph describing the audience, if you write such a thing as part of your
collection development plans.

Curation
Philosophy of ownership
Selection and retention
Ingestion workflow
Identifiability and privacy for follow-up and future studies (metadata)
Preservation
formats
restoration
preservation planning and technology updating Storage and
hosting
Choice of platform

Also (mentioned below a well), occasionally a need to ingest data after the original investigator has left, or without a specific request for curation from the investigator.

You may have to chart workflows, listing all the steps such as data integrity checks of your files (like checksum, to make sure you’re not archiving a corrupted copy), and create a file manifest.

One tradition archivists have is to gather or create contextual metadata at the time of ingestion. Gather all the details of who wrote things, what department it was from, what permissions are specified, and anything else that might be pertinent, now, while the parties concerned are alive and present, because if the archive is accessed 75 years later, you can’t count on anybody remembering what stuff was, what it was intended for, or who’s never allowed to see it.

Metadata falls into several categories, one of which is privacy and permissions, and that may be critical now and in the future. There are metadata schemata with fields to remind you what to fill in. Whatever curation platform you have will set the mechanics up for you.

For preservation, the heart of curation, you have to choose formats that you expect to last a long time, as well as plan for migrations when the technology goes out of date. There are some people, especially old computer hands, who say, “just upgrade every time,” but the idea of preservation formats is that they’re stable enough that while you do have to migrate or upgrade once in a while, you don’t have to mess with them every few months or years, especially if you have a lot of them. On the other hand, some formats supposedly designed for stability actually require other compromises, such as PDF/A which is called archival but actually doesn’t work with certain document sizes.

But what preservation truly is, is saving materials in such a way that they can be brought back and recreate the experience of using them when they were new. In more specific technical terms, it entails
storing in a format (that may be a compression to save space or expense), and also having a protocol for restoration. This is why backup is not the same thing as preservation, although it is an essential component of it.

Your choice of platform has preservation considerations, in that some systems may be around longer than others. Preservationists take into account the organizational environment of their hosting company or university or department, estimating longevity, planning updates, assessing likelihood of maintenance, and possibly some choices in metadata.

Curation plans contain short-term and long-term sections. Short term plans tend to be operational in nature, whereas long-term plans are hedges against the unknown, and describe technology migration issues.

Curation also involves plans and criteria for granting access and ensuring security, as well as access for all those entitled to it, so a welcoming, usable public interface would be desired unless there are considerations against it (such as cost, limited audience, or security).

That gets into whether you’re housing active data or not. A friendly, technologically sophisticated interface might be desirable in many cases, but if your audience is limited enough you may consider “cold storage,” which has noticeably limited or infrequent access. Some curation platforms, such as Amazon Glacier, are purposely designed that way.

A do-it-yourself form of cold storage is compressing archives into ZIP files, and copying and distributing those on request. This is actually preferred by users who want to manipulate the data to their heart’s content, and aren’t submitting frequent individual queries or updates, the way they would with active data.
Another aspect of curation decisions is when you have to curate to meet requirements. This is most well known in the data management plans required of grant recipients of the National Institutes of Health and the National Science Foundation. They require grantees to have a plan for storage, management, and access (often Open Access) of data generated in funded research. The actual federal requirements are pretty mild and generic, but they can provide for a grantee to develop rigorous or extensive plans, and curation is going to have to accommodate those plans (or the plans are going to incorporate curation—mostly they’re concerned with making the data available). Institutional Review Boards are there to prevent harm to research subjects, and that includes consequences of breaches of privacy. Thus they might specify how the data may be used, or to whom the data can be released. This can prevent secondary analysis later if somebody wants to look at the data in a new way. Curation can be affected by this in metadata and security configurations.

Now as a collections and technical services person who looked at data first as a resource noun, I am moving to data as a service verb, which is what many of you probably wished I had done first. So TTYL your cat. I’m going to move relatively quickly through some broader generalities from more of a management level, since these are going to be more familiar concepts.

## Libraries can offer

- Curation
- Procurement, identification of sources
- Metadata creation
- Consultation:
  - Data management
  - Data manipulation
  - Advice on data management plans
  - Data literacy (individual instruction, consultation on inclusion into syllabi)
- Data literacy programming

The top level categories of service I’ve pulled from what I’ve seen are what I’ve mentioned already. This is just a summary, and finally I’m talking about public services. We can offer general data reference, helping people find data sets for particular subjects or evaluate sources, and we can also work with them, as one of our librarians who started out in GIS did, helping them with file formats, embedding, printing, and other such technical hand-holding. We can act as consultants (once we learn how) for writing data management plans (although faculty in those disciplines may want to do that themselves—I’d bet they’d welcome the assistance, though) or even systems analysis, possibly. We can deploy ourselves as curators. I mentioned hosting briefly here—there are libraries which have installed the hardware to house faculty and students’ data physically; they can provide the dedicated personnel, and are knowledgeable about versioning, backup, restoration, and organization. It is, of course, a major
undertaking but I did a white paper on what it would entail at UT Arlington; you can still get university IT to handle the actual iron and do routines at 2:00 A.M., etc. And of course we can devise and present instruction and create ongoing data literacy programs, since our patrons are as new at this as we are.

We could set up certain individuals as data resource people, to whom public services librarians could refer patrons or projects, after having been trained on how to recognize data opportunities.

Reference service levels

(Kellam, *Numeric data services and sources for the general reference librarian*, Chandos, 2011, quoting Geraci et al.):

- Data identification
- Data reference interviewing
- Data content recommendation
- Data file advisory
- Data extraction
- Data analysis advisory

A valuable pioneer of library data services, Lynda Kellam, who wrote probably the first book on the subject, is reference-based, and she drew up these levels of data reference service, which include many of the things I’ve been mentioning.

As an analyst, I like to devise services by looking first at how people use information, or in this case, interact with data, and I came up with a short list from my experience as a researcher as well as an information practitioner. We have been talking about services we could offer, and we should tie them to our patrons’ and institutions’ actual needs.
How people interact with data

- Find and access data as readers
  - Open notebook science
- Make their own data available
  - Data Management Plans
- Show students data to teach analysis (secondary analysis)
- Show students how or where to find data

As we consider our service offerings managerially, we might consider a continuum, which I developed a few years ago for an ALCTS e-forum exploring the library’s role with data, and think about where we are along this continuum, where we’d like to be, and how to characterize our peers.

Continuum of service:

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Do nothing—leave it to researchers</td>
<td>Provide info. about data’s new importance, hosting, curation, copyright, data mgt. plan, but not do it</td>
<td>Provide info. about finding data sources</td>
<td>Provide advice on classification, metadata, structuring</td>
<td>Library institutional repository</td>
<td>Provide curation, copyright, data mgt. plan advice; provide forms &amp; institutionally approved templates</td>
<td>Partner with institutional IT for hosting</td>
<td>Provide hosting, curation, copyright, and data mgt. plan advice</td>
<td></td>
</tr>
</tbody>
</table>

Another, less linear and more territorial visualization would be aspects. I think of this like the Rumi fable of the blind people describing the elephant, although I kept coming up with way more than three panels.
Panels of data service

<table>
<thead>
<tr>
<th>Open Science</th>
<th>NSF/NIH/other grantors</th>
<th>Linking to content (especially journal articles, citations, c.v.’s)</th>
<th>Preservation as local personnel come and go, as well as standard preservation (backup and restoration) concerns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retention policies</td>
<td></td>
<td></td>
<td>Access for specific people in future</td>
</tr>
<tr>
<td>Data science, systematics</td>
<td>Visualization</td>
<td>Security, access control</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>- access credentials</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>- vulnerability (data classification schema needed)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>- intellectual property</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>- data use policies/agreements</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Reference and individual technical support</td>
<td></td>
</tr>
</tbody>
</table>

And finally, these are the professional societies I’ve been developing or maintaining relationships with, to keep current, and find other people to talk to. The first, which is the International Association for Social Science Information Services & Technology, best known as IASSIST.
It’s small and personable, and can’t put forth reams of material and services, but it’s the most targeted. There is the Research Data Alliance, but it’s pretty weak and laid back. It’s free, as far as I know.

Several years ago, ASIS&T started a thing that isn’t formally a SIG, which is ASIS&T’s divisions or caucuses, on Research Data Access and Preservation, or RDAP. They hold annual, or almost-annual, summits, and they’re also small and everybody can meet in one room and talk with each other.
SLA has organized formally a Data Caucus, and we’re doing strategic planning and organizing convention sessions.

None of these are particularly large or active, so you may need to put together more than one to get enough to work with.
Questions for further discussion:

Does curation provide us useful insights into service to our patrons with data, or can we just look stuff up to answer questions? Do the issues of metadata and versioning come up if we are doing instruction?

What would instruction for data look like? What would we need? Should there be a staff member serving as a resource for data?

How can we approach the gulf between disciplinary faculty and information workers as stewards of data?

What do you see that's missing on the handout? What services would you add, or configure differently? Any reactions to the checklist or any remarks?