Research Data Management Assessment

Tape library, CERN, Geneva by Cory Doctorow / CC BY-SA 2.0
Why Research Data Needs Assessment?

- Data Management Plan requirements (and high levels of federal grant funding)
- Interest in treating research data as important scholarly work
- Growth area for academic library support

General drivers behind academic library interest in data management and research data, and then we will focus on the specific assessments conducted at each institution.
Because these are the early days, we are not assessing our services, per se. What we were concerned with was assessing the state of affairs on campus, in order to uncover those aspects of research data management and archiving where the Library could provide institution-wide support.
The assessment was to investigate, evaluate, assess, and communicate Georgia Tech researchers’ data practices, processes, and outputs to enable the Library to understand and support their research data-related needs.

In order to do this, the Library assembled the Research Data Project Team, who planned, deployed, and analyzed the results of the assessment.

The data assessment was based on the Data Asset Framework (DAF), formally known as the Data Audit Framework, which is an assessment tool developed by the Humanities Advanced Technology & Information Institute (HATII) at the University of Glasgow in conjunction with the Digital Curation Centre. The objectives of DAF were to discover 1) what data assets researchers create; 2) how researchers manage, store, and share those assets; and 3) what researchers’ attitudes are toward managing, storing, and sharing their data. The DAF combines a set of methods with an online tool to enable data auditors to gether information about what research data assets exist at an institution and how researchers and organizations are managing these assets. The formal Framework includes four steps:

1. In the planning stage the purpose and scope of the audit is defined. Preliminary research is conducted and meetings scheduled to optimize time spent with the organisation’s staff.

2. The purpose of the second stage, identifying research data, is to establish what data assets exist and classify them according to their anticipated value
to the organisation. The classification step determines the scope of further audit activities, as only those data most important for your purposes will be assessed in greater detail in Stage 3, assessing management of data.

3. The information collected in Stage 3 will assist auditors to identify weaknesses in data policy and current data creation and curation procedures. This will provide the basis of recommendations in the final stage of the audit.

4. The knowledge gained from the audit will enable organizations to improve data management.

In preparation for implementing the DAF, the Research Data Project Team first determined internally what the scope of the assessment should be, and we identified available resources, such as funding, technical support, discipline expertise, and institutional partners. Based on these criteria, we modified the tool to match our local requirements. Rather than focusing on a comprehensive audit of a single school or research group, we developed a plan to canvas the entire campus, as we were hoping to develop a broad understanding of the research data environment across a university known for its de-centralized nature. While much attention in the professional literature is focused on the data-intensive disciplines within science and engineering, we also wanted to include other technology-rich disciplines that have a strong presence at Georgia Tech – including computing, architecture, music technology, and humanities-based digital media.
Based on our initial research, we ended up with a four-part assessment plan, each of which I will next talk about in more detail, but so that you get a quick overview of what’s ahead: we included a survey to gather some broad, general information about what research data exists on campus, interviews for a much more detailed and nuanced view of current research data needs, a dmp analysis to see what researchers were submitting to the NSF, and case studies, to more closely examine the data produced by researchers, as well the needed resources and expertise to archive those data.

The RDPT did receive IRB approval for the survey and interview.
The first method was the online survey. The questions were based on example surveys based on DAF that had been used at other universities, including the University of Edinburgh and Imperial College. Survey questions covered everything from the formats used for storing research data, the amount of data being generated, as well who has access to the respondent’s data, and what potential services related to research data the respondent might be interested in.

Before wide deployment of the survey, we conducted a pilot with the survey across all six Institute colleges, along with a number of major research centers and affiliated campus units. This was really important because we planned to survey research projects with a wide spectrum of methodologies, practices, budgets, and data management requirements, and we needed to insure that the assessment questions were not biased toward any one discipline or research scenario.

Based on the pilot, we did make modifications, including the removal and addition of some questions and edits to some existing questions; Examples of changes made included: originally wanted questions to be required, but changed after testing; questions about metadata were confusing to survey participants (moved to interviews because we could explain)
After testing, the survey went live on the Library’s website, where it lived from 2010 to 2013.

The survey was housed on a Drupal website on the Library’s site, and log-in through GT Central Authentication Service (CAS) was required

Ultimately, 77 members of the Georgia Tech campus took the survey, with members from all schools and all roles in the research process represented

There were several marketing campaigns conducted
to encourage campus partners to participate in the survey, including emails and active recruitment by subject liaisons, campus partners (HPC) as well as coverage in Institute newsletters and regularly communications (like email dailys, etc.).

It is important to note that this sample is self-selected and it is fairly small. Georgia Tech has about 5,000 faculty (both academic and research) and about 7,000 graduate students, so we really only reached a very small portion of campus. That being said, the information gathered was still incredibly useful to us, but we certainly cannot draw any hard fast conclusions from this sample.
Included this to show the type of distribution we had for the survey. We did assess the sample mid-way through and begin targeting particular areas, like Mechanical Engineering in the College of Engineering, in order to get a more representative sample.
In the survey, participants were given the option of volunteering for a follow up interview. 44 survey participants were willing to be interviewed, of which 26 were actually interviewed.

Interviews ranged from 30-60 minutes, with about a 45 minute average and they were conducted by members of the RDPT. The interviews were all recorded, and then transcribed, so that we could code them, which I'll talk about more in a few minutes.

There were ten main questions, with additional probes to either follow up or clarify. The interview questions were adapted from multiple other interview templates used for data assessments at other institutions (like Purdue and MIT). Based on many available examples, the RDPT created an interview instrument specific to Georgia Tech.

Questions were designed to gain more in-depth knowledge about the survey responses. It was also an opportunity to ask about things that did not readily lend themselves to survey questions – so questions about metadata and documentation, or how the research data were organized. Generally speaking, the interviews allowed us to gather much more in-depth information about researchers’ work with research data, as we could ask for clarification or ask follow up questions, which isn’t really possible with an online survey. We also included a question about SMARTech, our Institutional Repository, and this often served as a chance to inform the participant about the existence of the repository (and its ability to accept research data) as well as the Open
Access policy that had just passed.

This required IRB approval, so interviewees were asked to sign a consent form before we began the interview.

As with the survey, the sample is heavily self-selected. This was probably even more apparent when we started the interviews, because many of the people who were willing to be interviewed were usually not working with Big Data projects but smaller, more specialized research, and they had lots of questions because they didn’t have the support much larger, “big data” projects had
Again, we had at least some representation from each of the six colleges at Georgia Tech.

We had a particularly strong showing from the College of Liberal Arts, which is a comparatively small college (they have only 202 faculty and 283 graduate students), so we knew going into the analysis, that our sample (which it included someone from every college and major research center) was not representative of all of the institution.
In addition to the more traditional assessment techniques described before, the group wanted to also look at what researchers were actually saying in Data Management plans, and the easiest way to do that is to go to the source itself.

In cooperation with the GT Office of Sponsored Programs, we examined NSF DMPs submitted by Georgia Tech researchers during the first eight months of the NSF DMP mandate (Jan – Sept. 2011).

The original sample contained 335 submitted proposals. We excluded grant proposals that were grant supplements or transfers, or proposals that did not contain a plan or the plan only included one line, which left us with 181 data management plans.

Using plagiarism software (a free software tool called SPIaT: http://splat.cs.arizona.edu/), we searched DMP content for information related to repository services, inter- and intradepartmental sharing of DMPs and the prevalence of cloud-based tools.
This method was really important because we were able to observe things that would otherwise have been very difficult to determine, such as our finding that many plans contained incorrect or improper information about our repository, SMARTech, and that the incorrect language was wholesale lifted from one plan to the other.

Also, unlike the survey and interviews, this was not self-selected. We received the proposals whether or not the researcher volunteered, so there is less of a bias in that sense. However, because we only examined NSF plans (which is really all we could do because no one else regularly requires DMPs), research areas that more regularly seek NSF funding are more highly represented.

A handful of librarians, from Purdue, University of Oregon, Oregon State University, University of Pennsylvania, and Georgia are beginning work on an IMLS grant to continue this work further, but it is really early on, so stay tuned for more information.
The last method employed was a number of case studies of data archiving, so we recruited researchers from the Institution who had data they wanted to archive, and we walked through that entire process, slowly, so we could better understand what their data look like at the point they think they are ready to archive data, what issues arise when you try to archive data, what workflows facilitate or hinder this process, and so on.

In addition to recruiting some people through the interviews, a call was sent out through subject liaisons.

In total, we met with 8 different researchers (in Aerospace Engineering, Astrophysics, Computing, and Digital Media), and three resulted in deposit of the data, and one is still in process (and data are actually going to Archives).

In addition to the preliminary interview, we recorded information about the process of preparing and ingesting the data (workflow steps, time spent on these steps, etc).

In addition to helping us assess the needs and readiness of researchers to deal with data archiving, the case studies were essential for our own internal assessment of Library infrastructure. The case studies raised issues we had not previously anticipated, like IRB issues and consent forms, researchers trying to transfer too much data to Library servers through a Secure FTP, and the fact that our Dspace repository cannot support more than 2GB of digital data per record.
After collecting all of the data through the four previously discussed methods, we were left with data analysis.

The largest chunk of data analysis time was spent analyzing interviews and trying to make meaning out of our qualitative data (which was largely the interviews).

I would call our method of analysis “grounded theory” lite, because while we were definitely guided by the grounded theory method, we did not adhere strictly to it. For those not familiar, grounded theory is a research method (developed by sociologists Barney Glaser and Anselm Strauss) that values discovery of theory through the analysis of data rather than other, more traditional forms or research that begin with a hypothesis, which will guide the collection of data. So in our case, we gathered whatever data we could, and after looking at what we had collected, we began to identify some emerging themes, which we use to further examine and analyze our data.

Codes and codebook were developed by paying attention to emergent themes observed during data collection, but many codes were developed out the interview or survey questions themselves. For example, the “data security” code was a fairly direct match to the question, “Do you need security measures due to sensitive or proprietary information?” but the code “Motivations for Sharing Data” was developed in response to the fact that so many respondents touched upon this particular issue in their interviews.

The interviews were coded by three different people. We did make an effort to ensure
inter-coder reliability, each coding the same interview and then meeting to Once all the interviews were coded, we looked at the datasets from each code to try and find the overarching themes or concepts for that code, which proved to be very difficult, in part because our codes were too general. We wound up with some pretty long code reports. What was useful, once we had started to recognize some patterns, was to create a persona for the respondents, where we could focus on one particular element and look quickly across the different respondents to see if there was consensus.

Emergent themes and findings were mapped against findings from DMP’s and case studies. In some cases, this required going back to a dataset. For example, one finding from the interviews was about the disconnect between the expectations of the PI’s and the graduate students. Given this, we went back over the DMP’s to look for cases where this was discussed in the plans.
We used Dedoose cloud software for the coding process.

Dedoose is designed specifically for social scientists who are undoubtedly going to be working with human subjects, so they take data security very seriously, including measures to secure the servers where your data are stored, as well as encryption of data. Of course there are limits to what they can do, so you want to consider your needs and the sensitivity of your data before using a cloud tool.
Good things – using multiple methods provides a richness and dimensionality that would never be possible with only one method. In addition, many of the methods (mostly the interviews and case studies) allowed us to promote our services and it actually served as a means for greater campus engagement, generally. Finally, I am a big proponent of learning through practice, and so having the case studies to complement the survey, interviews, and dmp analysis was incredibly helpful in seeing how the themes observed in that data would play out on the ground.

There were some not so good things too – the codebook developed got the job done, but many of the codes were probably a little too broad and not discrete enough, so it became difficult to narrow in on a specific issue. For example, we had a code for “motivations to share data” and one for “conditions on sharing data” and those two particular themes tend to overlap quite a bit, so both code reports ended up looking pretty much the same, and it was really hard to separate out researcher feelings on say, sharing because it’s good for the public.

Another issue, that probably plagues most assessment efforts, is that our sample was not representative of campus (or very big) so we can’t draw concrete conclusions. And actually, for a few of our findings, I suspect that if we had a more representative sample, we would see different results. For example, we found that in general, folks were very willing to share their data. But the type of person who volunteers to take a survey or be interviewed about their data is probably much more inclined to share than someone who wouldn’t, and given the emphasis Georgia Tech puts on patents and the amount of
corporate sponsorship we receive, I have my suspicions that campus-wide, the willingness to share is actually much lower.

Finally, there are some important limitations for most of our methods, which is that most only allowed us to understand what researchers say they need or do, and if those practices are not actually what they do, we run the risk of developing services that don’t meet their needs. I think this is really important to note here because of the nature of the topic, things like data management responsibility or willingness to share. There are legitimate reasons why a researcher would intentionally or inadvertently skew their answers to sound more culturally or socially appropriate. Becoming embedded in a lab may not totally alleviate that issue, but I do think being in the research environment, or actually trying to locate data from researcher are ways to help augment findings like these, or at least ground them in a bit more reality.
So what did we learn from our assessment? We have a very lengthy report that details our findings, the link is below, but there were 8 main findings, which I won’t go into too much detail about. These findings are pretty consonant with other institutional assessments I’ve seen, but despite some generality, the findings have clear local applications as well.

1. Data management plans are still a frustrating burden for most researchers.
2. Georgia Tech researchers lack the guidelines, resources, standards, and policies to properly care for their research data.
3. A disconnect exists between the expectations of Principal Investigators and Graduate Assistants.
4. Researchers recognize the importance of documentation and metadata, but few capture this information adequately.
5. Sharing data with collaborators outside Georgia Tech is challenging.
6. Researchers are willing to share their data, but the conditions under which they are willing to do so vary widely.
7. Researchers rarely plan for the final disposition of their research data.
8. Very few researchers deposit data into repositories.


So what did we learn from our assessment? We have a very lengthy report that details our findings, the link is below, but there were 8 main findings, which I won’t go into too much detail about. These findings are pretty consonant with other institutional assessments I’ve seen, but despite some generality, the findings have clear local applications as well.
The next steps for us are to apply some of the findings to our local services, which is already underway.

We have also submitted a proposal to campus about creating an institutional framework to oversee data management activities, which would include, in addition to things like coordinated policies, the development of a data repository for all of campus.

We are part of an IMLS grant that is going to examine submitted data management plans to develop a rubric that librarians or other information professionals can use for evaluating plans.

And finally, we do have plans to assess the services we are developing, as well as researcher awareness and attitudes.


Contact Information

Lizzy Rolando
Research Data Librarian
Georgia Institute of Technology
lizzy.rolando@library.gatech.edu
library.gatech.edu/research-data