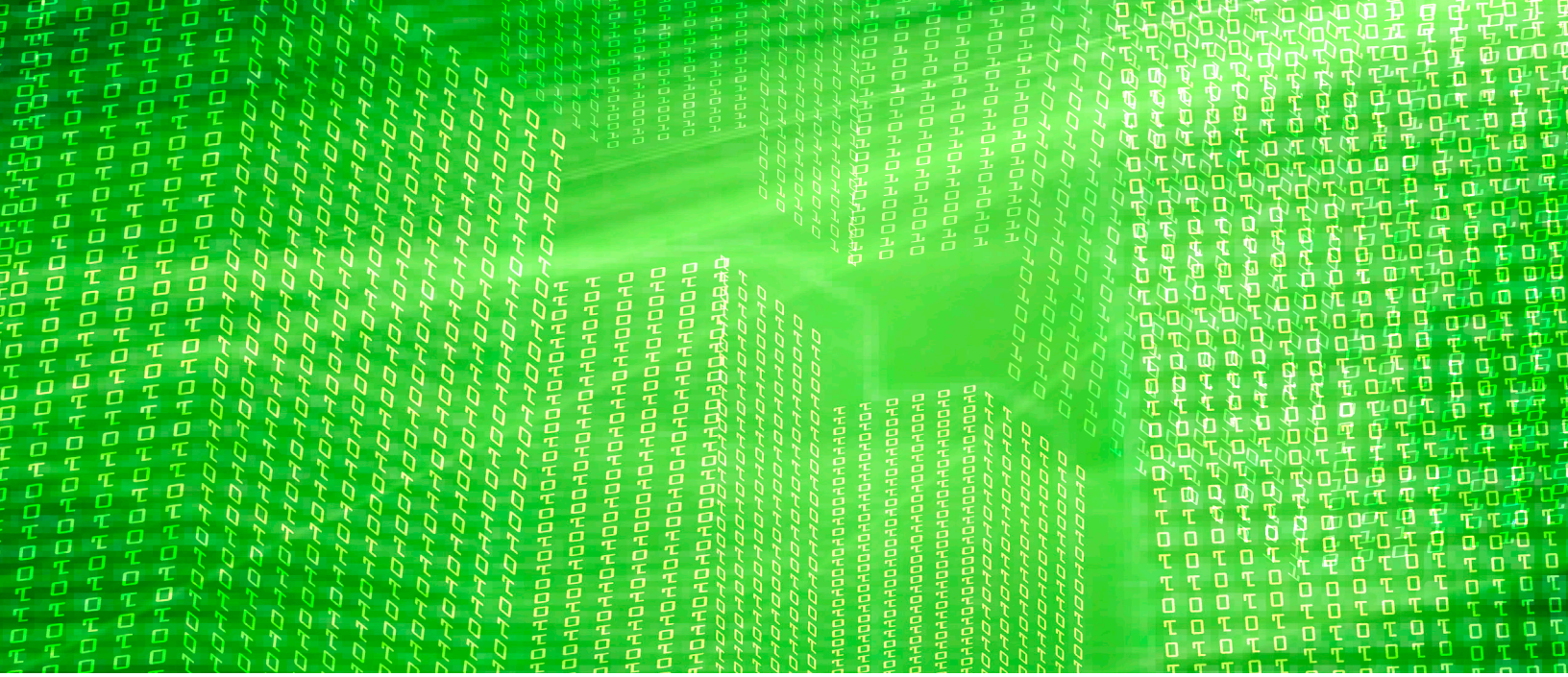

Institutional Readiness for Data Stewardship:

Findings and Recommendations from the Georgia Tech Research Data Assessment

Lizzy Rolando, Chris Doty, Wendy Hagenmaier,
Alison Valk, and Susan Wells Parham

JUNE 2013

Georgia Tech Library

1. EXECUTIVE SUMMARY

The potentials and possibilities afforded by managing, preserving, and sharing digital research data have been lauded by funding agencies, universities, and researchers alike. As federal funding agencies require data management plans and data sharing, questions around how to ensure that research data are managed and shared have come to the fore. Academic institutions and libraries are particularly interested in these issues, recognizing the need to support researchers in their work with research data. Accordingly, the Georgia Tech Library began investigating the research data practices and needs at Georgia Tech by conducting a campus-wide research data assessment. The assessment, which included a survey, interviews, analysis of data management plans submitted with NSF grants, and data archiving case studies, revealed a number of noteworthy trends, which are detailed more in the full findings of the report.

The major findings of the assessment were:

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 the final disposition of their research data.
8. Very few researchers deposit data into repositories.

Based on these findings, we make the following six recommendations:

1. Enhance institutional ability to support data archiving
2. Establish a campus Research Data Stewardship Group
3. Develop a formal data stewardship marketing plan
4. Create a repository of Georgia Tech data management plans
5. Provide data management training, especially for graduate students
6. Create and update the necessary and appropriate institutional policies

The challenges of caring for research data are many and constantly evolving, and Georgia Tech will need to adapt to the needs of our community. These recommendations are but a starting point for developing the institutional capacity to steward research data, but they provide important insight into the framework needed to properly care for institutional digital data.

TABLE OF CONTENTS

2. INTRODUCTION

In 2010, when the National Science Foundation (NSF) announced that all grant proposals were to include a data management plan (DMP), universities, libraries, academic publishers, policy makers, and researchers alike began in earnest to plan for data management, preservation, and sharing. Interest in sharing digital research data[1] had been growing prior to this announcement —the NSF required grant recipients to freely share the products of research funded by the NSF long before the DMP requirement[2] — but in the United States, the NSF announcement prompted levels of discussion and action previously unseen [1]. Accordingly, American universities have begun considering how to care for institutional research data; most major research universities are currently examining where researchers need additional support or services.

Institutional efforts to manage, preserve, and share research data continue with urgency today, particularly following the February 22, 2013 Office of Science and Technology Policy Memorandum [2], which called for greater access to the results of federally funded research. At the Georgia Institute of Technology, the Library began to explore ways to support campus members' work with research data in 2009, with the creation of the Research Data Project Team (RDPT). The full list of RDPT members can be found at the end of this document. In order to design the necessary research data services, members of the Library's RDPT designed and implemented a campus-wide research data assessment, evaluating current research data practices and research data needs. This report describes the work of this team, and then presents the important and noteworthy findings identified in the assessment. Finally, this report will make recommendations, for both the Library and the Institution, for how to manage, preserve, and share Georgia Tech research data.

---

[1] To facilitate conversations with researchers, the Research Data Project Team chose the following as the definition of "research data": *Research Data is digital information structured by formal methodology for the purpose of creating new research or scholarship. May be in a variety of formats suitable for communication, interpretation, or processing. Examples include: Observational data (e.g., sensor readings, survey instruments), Experimental data (e.g., lab equipment readings), Simulation data (e.g., climate models), Derived or compiled data (e.g., compiled databases, text or data mining). For our purposes, research data does not include published reports or papers based on analyzed data.* This definition was adapted from definitions by the MIT Libraries [3] and the U.S. Federal Government's Office of Management and Budget Circular A-110 [4].

[2] As early as 2004, the NSF expected "PI's to share with other researchers, at no more than incremental cost and within a reasonable time, the data, samples, physical collections, and other supporting materials created or generated in the course of their work." Retrieved May 9, 2013 from http://www.nsf.gov/pubs/gpg/nsf04_23/6.jsp.

2.1 BACKGROUND

Data management, preservation, and curation are all necessary to ensure that research data can be shared and reused. Data management typically refers to the actions taken during the course of research to describe, document, organize, and store active research data. While data preservation encompasses the processes that protect digital content for access and use at a later date (this involves procedures beyond traditional IT system administration, including media refreshment or bit-level integrity checks), data curation is "the active and ongoing management of data through its lifecycle of interest and usefulness to scholarship, science, and education," [5] and includes not only preservation but also those actions and services necessary to add value to the digital object over time. Examples of these services include metadata maintenance and creation, normalization of data formats, or migration of digital content.

Underlying efforts to better manage and curate research data is the assertion that digital research data should be shared or made openly accessible. Arguments in favor of sharing and reusing research data are many. When openly accessible, research data can facilitate scientific discovery and allow researchers to more effectively address the grand challenges of society [1, 6]. Data preservation and sharing are necessary to protect against data falsification or fabrication and to guarantee that research is truly reproducible [7-9]. Openly sharing data produced by publically-funded research funded allows for unanticipated re-uses [10], often by members of the public themselves [11, 12], and provides a mechanism by which researchers can remain accountable to the public that has invested in their research [13].

Inadequate data management and sharing can have severe consequences. In 2013, a graduate student looking to replicate a study by renowned economists Carmen Reinhart and Kenneth Rogoff requested access to the data underlying their report. The student found that the dataset contained errors, which once fixed produced significantly different results than those reported by Reinhart and Rogoff [14]. Harvard University recently investigated psychologist Marc Hauser for research misconduct, and auditors discovered that Hauser not only falsified data, in at least one study he had fabricated half of the data used in the study [15]. In 2011, Dutch psychology researcher Diderick A. Stapel confessed to fabricating data, some of which was used as the basis for dissertations completed by graduated PhD students he had supervised [16]. These are not isolated incidents, and they clearly

demonstrate the need for improved data management and data sharing to identify cases of accidental errors or intentional misconduct.

While discussions about how to care for research data have involved a variety of different stakeholders, including research scientists, funding agencies, and publishers to name a few, academic libraries have also been quick to engage with the open questions and issues around how to curate research data. In particular, academic libraries, given their common mission to acquire and disseminate information in support of research and education [17], as well as their expertise in areas of importance to data curation, such as digital preservation and metadata, are likely to play an important role in institutional data curation [18]. Further, as a recent feature in *Nature* on the future of publishing noted, many academic libraries have a long history of working with faculty at their institutions and they are trusted by their community [19].

Many universities and academic libraries have examined the current data practices at their institutions and have begun to develop the infrastructure, expertise, and services necessary to help researchers preserve and share their research data. Several of the findings from previous assessments are consistent across institutions. Researchers create and use a wide variety of types of data and formats [20-23]. They often consider themselves to be personally responsible for data management [22-24], and they regularly store research data on the hard drives of lab or office computers or on USB and external drives [20, 23, 25]. Unless a data management plan is required as part of a grant application or by a publisher, researchers will rarely create one, and many feel that they lack the resources and information to appropriately construct a plan [20, 21, 23].

At the Georgia Institute of Technology, the Library established the Research Data Project Team to investigate, evaluate, assess, and communicate Georgia Tech researchers' data practices, processes, and outputs. This has enabled the Library to understand and support the research data-related needs of the Georgia Tech community. Given the Library's history of effectively curating Georgia Tech scholarship, both in print and digital form, the curation of research data was a natural extension of Library curation services, and data curation was designated as an important strategic direction.

For the last 10 years, the Library has maintained all Georgia Tech theses and dissertations, as well as conference proceedings, technical reports from sponsored research, and faculty Open Access publications in SMARTech (smartech.gatech.edu), the Georgia Tech institutional repository. Institutional repositories have been widely adopted by many universities for curation of more

traditional scholarly outputs, and increasingly, they are now being considered as a possible solution for data curation [26, 27], as a supplement to well established disciplinary data repositories[3]. Similarly, the Georgia Tech Library, having successfully established SMARTech, has been pursuing ways that the GT Institutional repository can support data curation.

Curation of research data is critically important in ensuring that digital research data are available for re-use well into the future; therefore, in 2010, the Research Data Project Team designed, tested, and deployed the Research Data Assessment [28, 29]. Findings from the early stages of the Assessment highlighted the importance of Library involvement in data curation, but they also demonstrated the need for a broader collaboration between stakeholders all across campus. While data curation is a key aspect of research data stewardship[4], early results and discussions with stakeholders suggest that an institute-wide framework for research data stewardship is needed, a framework that includes policies, technical infrastructure, human expertise, and complementary data services [30]. The findings and recommendations reported in the document further reveal and detail the ways in which Georgia Tech and the Library can and should take action to support, preserve, and share the valuable research data generated by members of the community.

3. METHODS

The work and research underlying this report was conducted by the Research Data Project Team (RDPT), a committee of library employees committed to better understanding and caring for institutional research data. Beginning in early 2009, this group met monthly to discuss current issues in data curation, to assess the current data landscape at Georgia Tech, and to develop a plan for how Georgia Tech could develop and provide the necessary infrastructure, services, and support needed for proper data curation. The Research Data Assessment that was developed and deployed by the

[3] For example, in Social Science research, depositing research data into the Inter-university Consortium for Political and Social Research (http://www.icpsr.umich.edu/icpsrweb/landing.jsp) is common practice, as is depositing data into GenBank (http://www.ncbi.nlm.nih.gov/genbank/) or Dryad (http://datadryad.org/) for Biological or Bioscience research. Funding agencies that require data archiving recommend that grant recipients archive their research data with the appropriate disciplinary repository, if one is available.

[4] The Committee on Science, Engineering, and Public defined data stewardship as "the long-term preservation of data so as to ensure their continued value, sometimes for unanticipated uses" and noted that stewardship "embodies a conception of research in which data are both an end product of research and a vital component of the research infrastructure" [6].

RDPT consisted of four components (survey, interviews, DMP analysis, and case studies), each of which will be described in more detail below.

3.1 SURVEY

The first phase of the assessment was an online survey based upon the Data Asset Framework (DAF), an assessment tool developed by HATII at the University of Glasgow in conjunction with the Digital Curation Centre [31]. The survey was built in Drupal and was live on the Library website from 2010-2013. A complete list of questions from the survey can be found in Supporting Documentation.[5] Several marketing campaigns were conducted to encourage campus partners to participate in the survey.

In all, 77 members of the Georgia Tech campus took the survey, with members from all schools and all roles in the research process represented (See Table 1). One respondent participated in the survey twice and was not aware that he had done so (he agreed to be interviewed only the second time around.) Both responses from this one participant are included in the total results.

	Graduate Assistant	PI or Co-PI	Research Faculty	Unspecified Role	Postdoctoral Researcher	Author	TOTAL
College of Architecture	1	4					5
College of Business			1				1
College of Computing	2	4	1	1			8
College of Engineering	5	8	2		1		16
Ivan Allen College of Liberal Arts	2	14	1			1	18
College of Sciences	6	15	1	1	1		24
CETL						1	1
GTRI		2					2
EI²		1	1				2
TOTAL	16	48	7	2	2	2	77

Table 1: Breakdown of survey participants by College or Research Center affiliation and by role in research.

[5] The supporting documentation for this report can be found at http://hdl.handle.net/1853/48188.

3.2 INTERVIEWS

Survey participants were given the opportunity to volunteer for a follow up interview. 44 of the 76 unique respondents indicated in their survey that they were willing to be interviewed and ultimately 26 survey respondents were interviewed (See Table 2 for a breakdown of interview participants). A complete list of the questions and prompts used during the interviews can be found in the Supporting Documentation. Interviews ranged from 30 minutes to 75 minutes, with an average of 45 minutes per interview, and the interviews were conducted by either one or two members of the Research Data Project Team. All interviews were audio recorded, transcribed, and complemented by the interviewers' memos on noteworthy topics and themes. Based on the interview questions and the interviewees' responses, we developed themes and codes which were then incorporated into a codebook and full coding process using the Dedoose software[6]. Themes were tested and refined through further coding. When interview passages are used in the results, a note indicates the Georgia Tech College to which the researcher belongs, as well as their role in the research project, such as "College of Sciences - PI."

	Graduate Assistant	PI or Co-PI	Research Faculty	Postdoctoral Researcher	TOTAL
College of Architecture		1			1
College of Computing	1	3			4
College of Engineering		2	1	1	4
Ivan Allen College of Liberal Arts	2	6			8
College of Sciences		4	1	1	6
GTRI		1			1
EI²		1	1		2
TOTAL	3	18	3	2	26

Table 2. Breakdown of interview participants by College or Research Center affiliation and by role in research.

[6] http://www.dedoose.com/

3.3 DMP ANALYSIS

In cooperation with the Georgia Tech Office of Sponsored Programs, we examined National Science Foundation (NSF) DMPs submitted by Georgia Tech researchers during the first eight months of the NSF DMP mandate (January 18 through September 6, 2011). Of the 335 submitted proposals, we reviewed the content of 181 plans. Proposals that were grant supplements or transfers were excluded. Using plagiarism software, we searched DMP content for information related to repository services, inter- and intradepartmental sharing of DMPs and the prevalence of cloud-based tools.[7] Additionally, plans were individually reviewed by members of the Research Data Project Team to determine whether submitted DMPs identified particular standards or repositories, as well as whether the DMPs discussed the delegation of data management responsibilities among graduate assistants.

3.4 CASE STUDIES

The final phase of the assessment, which is still underway, is a series of data archiving case studies. Each case study involved the transfer of unique research data from a campus researcher into a data repository for long-term preservation and public access. The step-by-step process for the case studies varied from researcher to researcher, but most followed a similar pattern. First, members of the RDPT team identified potential campus partners and arranged for a meeting with the researcher. This initial meeting was followed by research by the RDPT team members into possible data archiving solutions. After potential solutions had been identified, these options were explained to the researcher, who was ultimately able to make a decision about where their data would be deposited. Once a plan was agreed upon, the researcher transferred their data to the Library, where a librarian prepared the dataset and accompanying metadata for deposit into the appropriate repository. At the time of this writing (05/13/13), one case study has been completed and four are underway. The completed case study was with a member of the Interactive Computing School, two of the ongoing projects are with members of the School of Physics, one is with a faculty member from the School of Aerospace Engineering, and the other is with a member of the School of Literature, Media, and Communication.

[7] More information about this project and early results can be found in [30].

4. RESULTS

The raw results from the survey assessment have been archived in Georgia Tech's Institutional Repository and can be found at http://hdl.handle.net/1853/48198. This study produced a substantial amount of results, all of which should not be reported here. In the sections that follow, results from all four studies - the survey, the interviews, the DMP analysis, and the case studies - have been synthesized into key findings that are detailed below. Because our sample for all four different studies is small, these findings may not be applicable to all researchers at Georgia Tech. However, despite the small sample, we observed some very clear trends that elicit notice and comment.

4.1 DATA MANAGEMENT PLANS ARE STILL A FRUSTRATING BURDEN FOR MOST RESEARCHERS

Despite some interviewee's assertion that the construction of a DMP is trivial, many researchers found data management planning to be a complicated and confusing process. 44% (34 of 77) of the survey respondents indicated that they would like information about developing a formal data management plan or other data management policies (fourth most requested service in the survey) and 36% (28 of 77) of the survey participants indicated that they would like assistance meeting data sharing and/or data management requirements of funding agencies (sixth most requested service in the survey). 44% (26 of 58) of the survey respondents who specified that they did not have a data management plan stated that a lack of information about data management plans was one reason why they did not have a plan. This general sentiment was also echoed by those interview participants who indicated in their interview that they did not have a data management plan.

With the exception of a handful of researchers, interviewees felt a general uncertainty about what was expected in DMPs, how the plans were to be reviewed, and what weight the plans would have in the overall proposal evaluation. Three interviewees specifically discussed ways in which they had struggled with the creation of a DMP. One of these interviewees raised a number of questions about the data management planning process, saying,

> I felt like in our data management plan, we overpromised. I had never done one of those before, so I made these ridiculous promises in the plan to archive everything…in ridiculous ways…So one thing, I think as a PI, I could use help with is scoping my data management plan

> appropriately. What do they really expect? And how much time is it going to take me to do this? And is there some way that the data management part of the project needs to be in the budget so that someone has been compensated for this new work that is being added on? And I don't know the answer to any of those things (College of Computing - PI).

Another interviewee explained that a recent grant application to the National Science Foundation had been declined, and the reviewers made a point of noting that the data management plan was inadequate. When asked whether he had been given any other feedback about the plan, he explained, "Presumably, one or more people on a review panel, who they're not with NSF, they're just academics, who use this type of equipment, has a different view of how this should be done. And if I had a different panel, I might have got a completely different response (College of Science and College of Engineering - PI)."

The review of DMPs submitted alongside NSF grants very clearly revealed the need for additional guidance in how to properly develop a data management plan. Despite the requirement that all NSF proposals include a DMP, some proposals did not contain one, and those plans that were submitted contained minimal information. Additionally, several plans indicated that they would conform to institutional policies, without indicating what those policies were or what provisions or requirements those policies contained. Much of this minimal text had been copied from one plan to another. The DMP analysis revealed that researchers were regularly sharing text between one another, as one third of the plans contained large sections that were identical to at least one other researcher's DMP. Two thirds of the text was shared between just a pair of faculty members, while the other third consisted of groups of four, five and six different faculty members. Unfortunately, in some cases, the shared text was outdated or incorrect.

4.2 RESEARCHERS LACK THE GUIDELINES, RESOURCES, STANDARDS, AND POLICIES TO PROPERLY CARE FOR THEIR DATA

Almost all researchers, regardless of their role in research or their discipline, lack the guidelines, resources, standards, and policies to develop thorough data management plans or to decide how to appropriately care for their research data. Further, in the few cases where these resources do exist, most researchers are unaware of them.

A handful of interviewees, when asked if they felt their data management or data archiving procedures were adequate, responded that they did not know how to evaluate their practices, and that they suspected superior methods or practices existed, as the following participant notes: "There's probably some improvements out there. I'm not sure exactly what they are (Ivan Allen College of Liberal Arts - PI)." A few went so far as to say that the Institute provides no guidance about how researchers are expected to manage or store their data. The desire for more information about best practices was also seen in the survey, as 52% (40 out of 77) of the respondents indicated that a potential service they desired was "Information regarding data management best practices." This was the third most highly requested service in the survey.

Similarly, interview participants pointed to the lack of centralized resources and services on campus. As expressed above, many felt that campus or their discipline likely had developed tools or services to meet their needs, but they were unaware of them. One interviewee, when asked what services he would like to see, responded "It would really be some collaborative tools, at least to know what is available. Maybe, where are collaboration tools, because there are a lot of things around, but is difficult to know what is good, what is bad (College of Engineering - PI)." As will be discussed later in the results, often the resources do not currently exist, or Georgia Tech has not developed or adopted the tools necessary to meet their needs.

None of the interview participants could name a disciplinary specific metadata standard that could be used for their data, and only one interviewee indicated that he was using a disciplinary repository. Further, none of the interviewees were aware of our subscription to the DMPTool[8], and only a few knew about our institutional repository, SMARTech; those who knew of SMARTech did not know the repository is able to accept data deposits.

The DMP analysis revealed a similar lack of awareness of available resources. Only 15% (27 of 181) of the evaluated DMPs stated that the PI planned to archive their data in SMARTech. In the review of the DMPs, as in the interviews, the vast majority of researchers could not name disciplinary standards or repositories. Further, for each of the case studies undertaken, the participants were

[8] The DMPTool is a free web application that takes federal funding agency requirements for data management and sharing plans, and chunks these requirements into digestible sections. Each section has prompts, questions to consider, and links to relevant resources. Georgia Tech has subscribed to the DMPTool since October of 2011. Since subscribing, Georgia Tech has had 57 unique users, and 60 plans have been created. The DMPTool can be found at dmp.cdlib.org.

generally unaware of standards or community best practices; only one researcher indicated that within her sub-discipline, the type of metadata expected to accompany the data is fairly well understood and standardized.

Another area of ambiguity and confusion was around what institutional policies would affect a researcher's work with their data, as well as the researcher's rights and responsibilities with respect to their data. One researcher bemoaned the difficulty his graduate student had locating and understanding the relevant regulations for maintaining or destroying consent forms for human subjects:

> I just graduated a PhD student in December, and the question came up…all of these consent forms, what do we do with them? How long do we have to hold them anyway? As tenured faculty, I'll hold onto things. Well how long do I have to hold on? Well it doesn't really matter. As long as there's storage space, we will hang on. But for a PhD student who's going away, he wants to make sure all of the things are tied up correctly…I thought it was very wise of him to go and look into that, but I had no idea (College of Computing - PI).

The DMP Analysis revealed that a surprising number of plans referenced Georgia Tech policies, but the policies either do not exist or they were merely named without discussion about what parts of the policy would have bearing on the particular project. The line "The preservation and sharing of these data will be governed by Georgia Tech's policies pertaining to intellectual property, record retention, and data management," was contained in many DMPs, without giving any detail about what these plans specify or where a reviewer could find these policies. This may not be entirely the fault of researchers - in many cases Georgia Tech has not developed institutional policies to which researchers can refer for this information.

The data archiving case studies have revealed a similar pattern. Researchers trying to submit data collected on human subjects were unsure about how to obtain permission to archive their data. Investigation into Georgia Tech policies regarding research data revealed that all research data at

Georgia Tech are classified, by default, as Category III data[9], and the procedures for determining whether a researcher is an exception to this policy are unclear.

4.3 A DISCONNECT EXISTS BETWEEN THE EXPECTATIONS OF PRINCIPAL INVESTIGATORS AND GRADUATE ASSISTANTS

A disconnect exists between the PIs and those creating data policies and data management plans for a research project and the members of the research team who would implement the policies and plans, typically graduate students. Students assume data management is the responsibility of the PI, and the PI assumes that students are responsible. Based on the interviews, graduate students were rarely given clear guidelines for how they were expected to manage or care for their research data. In many cases, the PI's let graduate students work with their data as they please. This type of arrangement was present in the following interview participant's research group. When asked about how adequate she feels her methods for organizing her data are, she answered,

> Ah, completely ad hoc. But I mean each student is in charge of their data, and for them if it works, it works. But I don't get into that level of micro-management. I mean… it would almost be disrespectful to go to the student and say, 'So, how are you organizing your files?' The projects are very graduate student oriented, so the management of the specific data files is a layer of detail lower than what I get involved in (College of Computing - PI).

Similar to the above example, a few of the interviewees noted that graduate students are allowed to operate independently because the PI doesn't want to micro-manage or because they recognize the need for someone who focuses solely on procedures and policies around data management, suggesting that PI's do not want to devote graduate student time to data management activities.

Even in cases where a grant or project has a data management plan, the staff and graduate students rarely know about it, despite the fact that a few of the analyzed DMPs specified that students and staff would play a critical role in data collection, analysis, and management. Of the survey respondents who answered the question, "Designate if you have a data management plan," 13% (10

[9] The full Data Security Classification Handbook can be found at http://www.oit.gatech.edu/sites/default/files/DSC_handbook.pdf, and the Data Access Policy can be found at http://policies.gatech.edu/data-access.

of the 77) were unsure whether or not they had a data management plan, and this confusion among post docs, staff researchers, and graduate students was echoed through the interviews as well. PI's typically knew definitively whether or not they had a plan, but they were unsure about whether the plan was being followed; conversely, graduate students and staff were often unsure about whether a plan had been written or they did not know what was contained within the plan.

In a few cases where the PI or the lead on the project had developed some standards, policies, or a data management plan, they acknowledged that their students rarely comply with these policies and standards or that they are not used consistently by members of the collaboration across the entire project. This was the case with the following participant, who had worked to develop guidelines for describing and documenting changes to data but was unwilling to confirm that his students were following them: "You know, some research projects, the graduate students end up taking much more of a leadership role on. And then at a certain point it becomes micromanaging, if I would show up and say, 'Hey, you! You don't name your files correctly' (Ivan Allen College of Liberal Arts - PI)."

Often, this disconnect leads to a situation where students act as gatekeepers to the data, either the entire data collection or the subsets with which they were working. In these cases, other members of the collaboration or research group are dependent upon the student for access to the data, because practices are not standardized and are therefore not understandable to anyone other than the original student. This problem is exacerbated by the high turnover rate for graduate students, who often leave without first documenting their work or directing others to where their data now reside. The following interview participant, when asked whether she had lost data when students left, replied, "Oh, I've had that problem. Yeah, that's not good. When they are doing little side projects that they've done themselves, like on MATLAB, that's when I lose them (College of Sciences - PI)." Also of note is one interviewee's point about the common practice of allowing students to use their personal laptops when working with research data, and how this is not only an irresponsible practice, but also opens the PI to potential liabilities that could be avoided if students were never allowed to store research data on their personal computers.

The DMP Analysis revealed a similar trend to what was seen in the interviewees – only 14 of the 181 plans discussed the role of graduate students in data management, and what information was provided typically highlighted the need for students to maintain a lab notebook or carry out regular

backups. Based on the DMPs submitted to the NSF, PI's do not plan in advance to have graduate students or post docs involved in data management, despite the reality that PI's leave most data management decisions to them.

What is perhaps most disquieting about this disconnect is that the survey results show that PI's are overwhelmingly believed to be responsible for data management, with 73% (56 of 77) of the respondents choosing "PI or co-PI" in response to the question, "Identify who manages the data associated with this project." Only 8% (6 of 77) of the respondents indicated that graduate students manage the data. This may be what survey respondents believe should happen, but the interviews and well as the case studies reveal that this is rarely the case.

4.4 RESEARCHERS RECOGNIZE THE IMPORTANCE OF DOCUMENTATION AND METADATA, BUT FEW CAPTURE THIS INFORMATION ADEQUATELY

While the survey did not ask about the use of metadata, almost all interviewees indicated the need for thorough and accurate metadata, whether for locating data or for understanding what data they have. When asked about what tools or services would be helpful, one interviewee replied saying,

> I think having really good metadata, that's one of the things that comes to mind. That makes a huge difference. It's just night and day difference between guessing about various aspects of the dataset and being able to look at a header file or equivalent of that and being able to see the dimensions and array sizes and things like that... What's made the NETcdf format so successful is that they have really good metadata so you can see that information really clearly, so that's not just for me, that's just something I would like to see develop for all scientific datasets. Something that tells you what is this thing you're looking at. What are some of its relevant features (College of Engineering - Post doc)?

Only one interviewee employed a community standard for recording and sharing metadata. Most interviewees were not aware of standards available to them, nor had they considered looking for one to use in their own research. Eleven of the interviewees noted that they do use some form of naming convention for files or directories, but even in those cases, researchers were using locally developed standards that are not necessarily interoperable with those being used by other labs or research groups. Further, even among those who have made an effort to collect and record metadata

(including the participant who uses a community standard) metadata creation among the entire research group is inconsistent and ad hoc.

Only three interviewees indicated that they create and maintain human readable text files, either embedded within the data file or in the directory that contains the data files. These text files are developed and maintained so that the original researcher or their collaborators will understand the data files when used at a later date, which suggests that researchers are aware that documentation is important. A handful of interviewees indicated that they maintain lab notebooks with information about the experiments and procedures; however, as one participant noted, keeping the information in the physical lab notebooks associated with the digital research data is often very difficult. Although two interviewees used collaborative data management software tools that they felt helped their work and supported the capture of metadata, five felt their field lacked the tools to support proper metadata collection.

Interviewees discussed two different types of metadata in their interviews, without necessarily ever noting the difference – usability metadata (documentation and metadata that would be necessary for future use, either by the original researcher, their collaborators, or unknown future users) and discoverability metadata (metadata needed to locate the data). Neither type of metadata is created in a consistent, intentional, or standardized manner. In working with researchers for the case studies, the allocation of responsibility for metadata creation has come to the fore. While librarians and repository staff can create metadata to support the discoverability of the datasets and can help normalize the metadata to conform to community standards, researchers themselves must create much of the usability metadata necessary for future reuse – they alone have the knowledge and information that needs to be documented.

The review of DMPs revealed that researchers were seldom discussing metadata in their plans. Only 48% (86 of 181) of the reviewed plans included the term "metadata," and only 20% (36 of 181) provided any details about the metadata they intended to collect. Examples of the types of metadata discussed in the DMPs include file naming conventions or a separate file that graduate students working on the project will create that will define column headings or experimental design procedures. Some plans indicated that the PI would review these metadata to verify that they are accurate and complete. However, far more investigators either did not discuss metadata in their plan, or they merely included the word in the plan (usually in the heading of the section "Format and

Metadata Standards") without explaining what metadata they were going to create, how they were going to collect it, who would be responsible for collecting and verifying the metadata, or how they were going to share the documentation and metadata. Further, very few of the DMPs indicated that the applicant was planning to follow community metadata standards (only 6 of the 181, or about 3%, of the plans named a community metadata standard).

4.5 SHARING DATA WITH COLLABORATORS OUTSIDE GEORGIA TECH IS CHALLENGING

In the interviews, participants overwhelmingly indicated that sharing research data with external collaborators is very difficult. Researchers who were transferring or sharing smaller files were affected less than their colleagues who need to transfer very large files, but small files are often shared through email, Google Documents, or Dropbox, which is discouraged by the institution because of the increased possibility for security breaches. Researchers sharing medium to large files that could not be shared across cloud tools because of their size found data sharing very challenging because of firewalls and other security measures in place at Georgia Tech.

One interviewee explained how security measures on campus computers affected her work, saying,

> So right now, we are really exchanging data in a very ad hoc way. So, this is really more about sharing than archiving, but we would like to have it someplace where we can archive and share, I guess. Maybe those are different things, but at this point, we really need to share just to get started, and the files are too big to email. So you know, we used to just FTP, but now with firewalls, and people don't use UNIX so much, and it's not as convenient for everyone to FTP, so we've been using Dropbox (College of Engineering - PI).

The use of cloud services to share data with those outside Georgia Tech was evident in the survey as well, as 38% (29 of 77) of the participants indicated that they share data through a collaborative web space and 58% (45 of 77) responded that they used email to share data. Only three indicated that they use Dropbox specifically, but Dropbox was not a choice on the survey - three people wrote this answer in the free text section of the survey. The DMP analysis also revealed that Dropbox was used by researchers to share and store research data, as five of the DMPs discussed using Dropbox to store or share data.

One interviewee explained that she struggled with the security on the local machines in her lab, noting that the high performance computing machines were much easier to access. For her, the disparate levels of security made her work with her data and her collaborators unnecessarily challenging: "Their computers are easier for me to get into then ours. Like I can't go from their computer into our computer. I have to go from our computer into their computer, so it's more those kind of issues. Like computer security issues than the actual data sharing issues (College of Sciences - Post doc)."

A handful of the interviewees noted that they would prefer to give their collaborators access to the servers here at Georgia Tech, but they cannot do so because of security concerns and measures taken to protect institute computing resources. One interviewee described her need, saying,

> All I really want is a good housekeeping seal of approval. What I want is to know that I'm sharing this with [School X] and [School X]'s IT folks meet the green checkmark good housekeeping seal of computing. But if I'm sharing it with Nowhere State, their network has not been approved, so therefore I need to handle it differently. And then I need to know how to handle it differently if I'm trying to share sensitive data with Nowhere State. One way to handle that is to somehow give access to our system to the partner, and say 'Look you can use this data but it has to stay on our servers.' But then our IT people get nervous, and they don't like that (College of Computing - PI).

Not only do IT security measures at Georgia Tech make sharing and distributing data with collaborators on a project difficult, the measures also complicate sharing data with researchers outside the original group.

4.6 RESEARCHERS ARE WILLING TO SHARE THEIR DATA, BUT THE CONDITIONS UNDER WHICH THEY ARE WILLING TO DO SO VARY WIDELY

Almost all interviewees were willing to share some of their data with researchers outside the original research group (only 3 did not identify motivations or situations for sharing). 41.5% (32 of 77) of the survey participants expressed interest in sharing data with researchers at other institutions, 21% (16 of 77) would like to share data with project sponsors, and 12% (9 of 77) wanted to share data with the general public. However, most researchers are concerned about making all of their research data available openly. 44% (34 out of 77) of the survey respondents indicated that one reason why

they would not share data is because the data are confidential, proprietary, or classified, and 30% (23 of 77) responded that intellectual property concerns were an impediment to sharing their data.

This was noted in the interviews as well, as 11 of 26 interviewees indicated that some of their data were sensitive in some way (either because they were collected on human subjects, the data were proprietary, or they were to be used for commercial development), as was the case with the following interviewee: "Well, so one thing I should say, further down the road, I would be totally happy to share my data. Right now it's part of a company I'm trying to start and there's intellectual property involved. Five years from now, whatever, if you can find a use for it, you can have it (College of Computing – PI)." The apprehension about sharing sensitive data is likely much more prominent across campus, as participation in the survey and interview were voluntary, and those interested in sharing data were likely more motivated to take this survey than someone who is unwilling or unable to share their data.

Some participants were concerned about the time required to prepare their data for someone outside of their research group, noting that they either had not been asked for their data or they were very rarely asked, as was the case with the following interviewee:

> It's so rare that people like me would be asked for our data, that to do any work up front, feels like a waste of time. For every study you do, and I know you guys will try to keep it relatively minimal, the [discipline X] database tried to keep it minimal, but it still was work. And, it seems like time saving to wait on the low probability even that someone asks you for data, then you do the work on those data to say, well here's the…but it does not allow for unrelated uses, like these databases allow, which I think could be a good thing. It'd be good to have those available for those (College of Sciences – PI).

Other worries about sharing data included the lack of appropriate tools, the desire for attribution or involvement in the resulting project, or worries about possible misinterpretation of the data. The full list of concerns or conditions about sharing data can be found in the Appendix D. However, despite some participants' trepidations, like the interviewees above, many recognized the value in preserving and sharing their data, and were their conditions met, they would be willing to share their data with others.

4.7 RESEARCHERS RARELY PLAN FOR THE FINAL DISPOSITION OF THEIR RESEARCH DATA

Overwhelmingly, researchers do not plan for the final disposition of their data. Most researchers have no strategy for identifying their valuable data that should be preserved, they have no plans for ensuring that valuable data are not inadvertently lost, and in most cases, they have no specific timeline for when they expect to dispose of data. While most participants understand the value in permanently archiving their data, researchers do little to secure or preserve their data once a project is completed. This pattern of behavior was observed in the survey results, the interviews, the DMP analysis, and in the case studies.

Many interview participants want to store their data forever, despite their inability to anticipate when or why they would use those data or who else would be interested in them. During the case studies, participants were generally interested in preserving their data, but only through discussions with the librarians were they able to understand which data they would like to preserve and what actions would need to be taken to preserve the data. Many interviewees and survey respondents felt that because of the decreasing costs for storage, they had no reason not to keep all of their data:

> I can't imagine a situation where I would intentionally delete data, ever. Because why do that? I think it would be worth buying another hard drive then deleting data. So, I think that I would…that's how I would think of that. But I don't…I wouldn't do that because I thought these data were important, that it would really ever matter that I had them. But just on the unknown, unforeseeable chance that, well I might want to look back (College of Sciences - PI).

Despite the fact that researchers continue to store data from past projects, these data suffer from benign neglect, and typically, researchers name only the price of the storage medium when discussing the cost of preservation and perpetual storage. Almost all participants overlooked the human expertise needed to maintain the storage and the data files. Researchers rarely take action to ensure that their data would be understandable, readable or usable in the future. Participants never indicated that they utilize integrity checks, format migration, or media refreshment. Further, most researchers had no plan for where their data were to be permanently stored, and because rarely was anything done to archive the data, wherever the data lived during the course of the research was where those data were to live forever.

Perhaps unsurprisingly, participants' responses varied when asked how long they plan to retain their data. While many interview participants expressed interest in saving their data forever, survey results showed discrepancy between individuals, both in terms of how long they plan to save their data and the reasons for that retention period. 35% (27 of 77) of the survey respondents indicated that they plan to retain their data for 1-5 years, followed by 23% (18 of 77) who responded that they plan to retain their data indefinitely, and then 19% (15 off 77), who plan to keep their data 5-10 years. Respondents who plan to keep their data less than five years overwhelming felt that this time period was consistent with the goals or length of the project. Those respondents who planned to keep their data longer than five years overwhelming responded that they had chosen that time period to allow for future re-use, either by the original investigator or someone from outside the original project. This suggests that researchers who plan to re-use or share their data expect to retain their data for longer periods; however, beyond planning for a longer retention period, most researchers did not take additional steps to safeguard their data so they could be re-used in the future.

4.8 VERY FEW RESEARCHERS DEPOSIT DATA INTO REPOSITORIES

Researchers very rarely deposit their data into repositories. The assessment revealed very little use of the Georgia Tech institutional repository SMARTech or disciplinary repositories for the curation of research data. Only four interviewees had deposited data into a repository. One interviewee intends to deposit his data into a relevant repository because of the NSF's data management plan requirement. One interviewee submitted data to a database because the journal required data deposit into a repository as a condition of publication. Another interviewee keeps his raw, paper surveys in the Archives at Georgia Tech, although others are not granted access to them. One interviewee submitted data to SMARTech during the course of the Research Data Assessment. Aside from these cases, no other interviewees deposit data into a repository, and many were unaware of repositories that would allow them to deposit their data.

The interview results are consistent with the findings from the DMP Analysis, which found that only about a fifth of the plans indicated that SMARTech would be used either for research data or for theses and dissertations stemming from work on the grant. With the exception of a few plans (6 named community or disciplinary repositories and 4 named repositories at other universities), most DMPs did not specify that they planned to deposit their data into a repository.

Despite low levels of repository deposits, overall, interviewees were interested in submitting some of their research data into a repository or archive so that others could access and use them. This interviewee, when asked about services to improve his work with research data stated:

> I think a data repository for the [government agency] type data or someone unselfishly putting their [government agency] data, like just the raw files, because it takes a lot of time to download the right variables and all that good jazz. Finding that in a data repository, like the ICPSR, has… you know, the Census data in a more manageable format than checking the variables you need and things like that (Ivan Allen College of Liberal Arts – PI).

Some participants expressed interest in archiving their data, but not for public consumption. Rather, they would like a safe, private preservation environment for their data. The ability to deposit their data into a repository, so that those data could be retrieved at a later date to be re-used by members of the original research team was an attractive prospect to many participants.

5. CONCLUSIONS

The research data landscape at Georgia Tech, like that of many other American universities, is varied, complex, and constantly evolving. However, as detailed above, certain aspects of managing, preserving, and sharing research data are especially challenging or frustrating for researchers. In order to meet the needs and expectations of researchers, institutional stakeholders, publishers, and policy makers, Georgia Tech should continue to assess the research data practices and needs of its researchers. While flexibility will be critical to the success of data stewardship at Georgia Tech, our assessment has identified key areas where researchers need assistance and where the institution is poised to provide the resources and guidance necessary to support campus members in their work with research data. In the section that follows, we lay out recommendations for how to address the findings detailed above. These recommendations are not exhaustive, but they provide specific examples of actions that can be taken to begin to develop a comprehensive institutional framework for data stewardship.

6. RECOMMENDATIONS

Based on the findings detailed above and the data collected for the full assessment, we propose the following six recommendations:

1. Enhance institutional ability to support data archiving
2. Establish a Campus Data Stewardship Group
3. Develop a formal data stewardship marketing plan
4. Create a repository of Georgia Tech data management plans
5. Increase data management training, especially for graduate students
6. Create and update the necessary and appropriate institutional policies

These recommendations are described below, and then mapped to the key findings, to demonstrate how the recommendations will address the issues identified in the results, in Table 3.

6.1 ENHANCE INSTITUTIONAL ABILITY TO SUPPORT DATA ARCHIVING

Given researchers' interest in data archiving, federal agency's requirements for data management and data sharing, as well as the need for the institution to capture scholarship and intellectual property and guarantee compliance with federal policy, Georgia Tech should invest in the development of a repository for research data, both to facilitate wider access to the data and for preservation. Investment in the repository is necessary to develop the curation infrastructure required to protect valuable assets created by researchers at Georgia Tech for future use. For some researchers, disciplinary repositories are available for data archiving[10], and in these cases, the institution should support the use of the pre-existing resources. However, many of the researchers at Georgia Tech produce research data that lacks this type of infrastructure. Therefore an institutional solution is necessary. The Library should lead this particular effort, given their expertise in digital curation and because they have extensive experience working with and developing digital repositories. The list of formats of data used by researchers, obtained through the survey and located in the supplemental file, will be instrumental in helping to highlight the formats the repository should be able to support and preserve.

Potential functionalities that could be built into the repository include the ability to keep a collection private and to require mediated download of data in the repository. Many researchers were interested in being able to deposit data for safe keeping, but they were uncomfortable with making

[10] For example, researchers in the Biosciences can deposit data underlying peer-reviewed publications into Dryad (http://datadryad.org/), and seismologists can deposit research data into the Incorporated Research Institutions for Seismology repository (http://www.iris.edu/software/).

their data open. Although allowing researchers to keep collections private would not immediately support the sharing and reuse of data, Georgia Tech would be better able to preserve and manage valuable institutional assets. Further, researchers who initially are uninterested in sharing data may feel differently later. By providing these researchers with the tools and services to preserve their data into the future, the Institute would facilitate researchers' ability to share their data at a point later in their career.

Researchers were also interested in having some control over what happens with their research data after they have submitted their data to a repository. Many were concerned that their data would be misunderstood or misused, they wanted to be involved in the resulting project, or they wanted assurances that they would be given credit for creating and publishing their data if they were used by another researcher. In some cases, researchers wanted to share their data, but because of the sensitive nature of the data, the data could only be shared under specific circumstances. One way to accommodate and address some of these issues is to require end users to agree to a license or terms of use before they can download data from the repository, or to require approval from the depositor before the data are released.

6.2 ESTABLISH A CAMPUS RESEARCH DATA STEWARDSHIP GROUP

Because research data stewardship is an issue that interests and affects groups beyond the Library, the Institution should develop a Research Data Stewardship group, drawing upon the expertise of current Institute Data Stewards, as well as additional stakeholders, such as the Library and researchers themselves. The group would coordinate data management services and related infrastructure development and propose institutional policy to promote research integrity as it relates to the management and preservation of research data. While data curation falls well within the purview of the Library, proper care of research data requires a much broader perspective, one that focuses on the stewardship of research data. Stewardship includes data management planning, secure retention and disposal, sharing and publishing of research data, and compliance with institute policies, legal requirements, and ethical standards. A proposed structure for this group is as follows:

- Joint sponsorship between Executive Vice President for Research, the Dean of the Library and Learning Excellence, and the Chief Information Officer
- Faculty advisory board to ensure that the needs of researchers are met across the disciplines

- Management council to coordinate and develop services (composed of representatives from each of the major service providers)
- Coordinator to facilitate management council and day-to-day operations
- Implementation teams to carry out specific tasks

6.3 DEVELOP A FORMAL DATA STEWARDSHIP MARKETING PLAN

The lack of awareness of many researchers about existing resources is discouraging and should be counteracted with a planned, strategic marketing effort. This outreach campaign could be done in conjunction with similar efforts for other projects – the recently passed Open Access policy, the Faculty Profile System GTScholar, and the still in development but soon to be widely marketed Data Access Policy.

A marketing plan helps facilitate four different goals: 1. Researchers will be better informed about the policies that affect their research and work with data; 2. Researchers will know what resources and services are available to them in order to help them become in compliance with these policies; 3. The Data Stewardship Group can assess the value of existing data services and of their marketing efforts. Marketing of services has been ad hoc and done almost exclusively by the Library, which complicates our ability to understand what services are desired or what services require different or improved marketing efforts; and 4. The more faculty and researchers know about the issues related to research data, the more readily we can foster communication between those working with data and those creating services and resources.

In addition to the marketing plan, a central website with information regarding research data should be constructed and maintained by the Library. Information regarding all services on campus related to data management and archiving should be maintained on the website, so that researchers need only refer to one location for information about existing guidelines, resources, standards, and policies. This site should be integrated with other relevant units on campus, such as the Office of Sponsored Programs.

6.4 CREATE REPOSITORY OF GEORGIA TECH DATA MANAGEMENT PLANS

The NSF has been criticized for not providing more direction or feedback on submitted data management plans, and respondents in our assessment indicated that more information about plans

would be helpful. The Library has offered to review DMPs for the last few years, but the service is not widely used. Because researchers do not think to ask the Library for guidance on a grant application, creating a resource or service that more directly addresses their need is required. Developing a curated collection of DMPs written by Georgia Tech PIs is one way the Institute can provide guidance in the data management planning process. While it would be inappropriate for grant applicants to copy a plan verbatim, access to example plans helps researchers understand how to appropriately frame their own DMP and provides them with examples of practices and resources they themselves could be incorporating or using in their own work with research data. Curating this type of collection is well in line with the work of libraries and archives, and if hosted by the Office of Sponsored Programs or the Office of Research Integrity Assurance, this collection would be located where researchers are most likely to look for assistance with questions about their grants and research obligations. If the plans are separated by discipline, links to these examples could be included in the different templates in the DMPTool.

6.5 INCREASE DATA MANAGEMENT TRAINING, PARTICULARLY FOR GRADUATE STUDENTS

Until the tools that will do all data management for researchers are invented, much of the data management work must be done manually. This requires that researchers are aware of potential data management issues and that they stay up to date on the best practices in their discipline. Training in data management is critical for raising awareness, ensuring that research data is properly managed, and for educating the next generation of world-class researchers. While training should be offered to everyone who works with research data, given that graduate students are so often working closely with the data, they are an obvious first group to target with instruction. Already the Library has offered training that counts towards the Responsible Conduct for Research requirement, and work is underway to develop a data management boot camp for graduate students; however, more can done, including specialized data management instruction, integrated with coursework or lab work, as well as requiring data management education as part of the Responsible Conduct of Research training requirement[11].

[11] All Georgia Tech doctoral students who have an admit date later than Fall 2011 or master's students pursuing a research-related degree are required to complete Responsible Conduct of Research Training. Data Management

6.6 CREATE AND UPDATE THE NECESSARY AND APPROPRIATE INSTITUTIONAL POLICIES

Potentially the most challenging recommendation to implement is to develop the necessary institutional policies to guide researchers to work with and manage their research data ethically, responsibly, and in accordance with federal, state, and local law. In some cases, this only requires adjusting existing policies to account for the growing interest in treating research data as an important scholarly output, in and of itself (for ex. GT Intellectual Property Policy [33]). In other cases, new policies must be developed, such as policies governing responsibility for management and retention of all types of research data[12]. Peer institutions who have adopted comprehensive research data policies include Johns Hopkins University, Stanford University, and the University of Virginia[13]. While Georgia Tech values the entrepreneurial spirit of its researchers, and maintaining the balance between the needs of the Institution and the need for researchers to conduct efficient and creative research will continue to be important, the ambiguous nature of research data and the rapidly changing research environment leaves Georgia Tech exposed to potential liabilities. Additional exploration into the appropriate research data policies to be developed and adopted by the Institution is needed.

---

workshops are currently optional electives that count towards the in-person training requirement. More information on the policy can be found at http://www.rcr.gatech.edu/resources/.

[12] Currently, the Board of Regents Retention Schedule indicates that data related to Human subjects, agricultural products, and animals should be kept either 3 years after the completion of the project (for animals), 70 years after the completion of the project (for humans and agricultural products), or forever (for animals, humans, and agricultural products) [34].

[13] JHU's "Policy on Access and Retention of Research Data and Materials" can be found at http://jhuresearch.jhu.edu/Data_Management_Policy.pdf. Stanford's "Retention of and Access to Research Data" Policy can be found at http://doresearch.stanford.edu/policies/research-policy-handbook/conduct-research/retention-and-access-research-data, and UVa's "Policy Statement on Recording and Storage of Laboratory Data" is located at https://policy.itc.virginia.edu/policy/policydisplay?id=RES-002.

Finding	Recommendation	
Data management plans are still a frustrating burden for most researchers	· Develop repository of GT data management plans · Training · Marketing Campaign	Example plans will provide critical guidance in developing DMPs, as will data management training. Further, researchers who are aware of resources on campus are better prepared to plan for data management.
Lack of guidelines, resources, standards, and policies	· Research Data Stewardship Group · Develop repository of GT data management plans · Institutional Policy · Marketing Campaign	The Research Data Stewardship Group should identify, create, or adopt guidelines, resources, standards, and policies to assist researchers in data management. Targeted marketing will raise awareness of these resources.
Disconnect between expectations of Principal Investigators and Graduate Assistants	· Institutional Policy · Marketing Campaign · Training	The disconnect between PI's and students could be alleviated by increasing awareness about the issue, through more formal policy statements about data management responsibility, a marketing campaign, and increased data management instruction. Further, additional training for graduate students would help address the issue of "data gatekeepers."
Researchers recognize the importance of documentation and metadata, but few capture this information adequately	· Training · Repository Development	Until the tools are developed to automatically capture and generate metadata, training is one of the only ways to encourage proper usability metadata creation. The proposed data repository would assist researchers in the creation of discoverability metadata.
Sharing data with collaborators outside Georgia Tech is challenging	· Research Data Stewardship Group · Marketing Campaign	Examining how the Institution can facilitate data sharing between Universities should be one of the first tasks for the Data Stewardship Group. Once solutions have been identified or created, the marketing campaign will help communicate this to the Georgia Tech community.
Researchers are willing to share their data, but the conditions under which they are willing to do so vary widely	· Marketing Campaign · Repository Development	Increasing researcher awareness about available services to support data sharing will help those researchers interested in sharing. Developing a data repository that is flexible enough to address researcher concerns (such as requiring end users to agree to Terms of Use written by the data creator) will help encourage additional sharing.
Little planning for final disposition of data	· Institutional Policy · Training · Repository Development · Marketing	Institutional policy addressing data archiving and sharing is necessary to ensure that data are dealt with suitably at the end of a study. Training researchers to care for their data and plan for data archiving is also critical, as is providing the services and infrastructure necessary to support long-term archiving. Finally, researchers must be aware of relevant polices and services.
Very few researchers deposit data into repositories	· Institutional Policy · Training · Repository Development · Marketing	Establishing and marketing the policies, training, infrastructure, and services to support data archiving in data repositories is the best way to safeguard research data and preserve them into the long-term.

Table 3: Mapping between key findings, the recommendations stemming from those findings, and the explanation for how the proposed recommendation addresses the finding.

7. REFERENCES

1. Borgman, C.L., *The conundrum of sharing research data.* Journal of the American Society for Information Science and Technology, 2012. 63(6): p. 1059-1078.

2. Holdren, J.P., *Memorandum for the Heads of Executive Departments and Agencies: Increasing Access to the Results of Federally Funded Scientific Research*, O.o.S.a.T. Policy, Editor 2013: Washington, D.C.

3. MIT Libraries. *What is Data?* Data Management and Publishing n.d. 05/09/13]; Available from: http://libraries.mit.edu/guides/subjects/data-management/what.html.

4. Office of Management and Budget, *Circular A-110 Revised 11/19/93 As Further Amended 9/30/99,* , 1999.

5. Shreeves, S.L. and M.H. Cragin, *Introduction: Institutional repositories: Current state and future.* Library Trends, 2008. 57(2): p. 89-97.

6. Hey, T., S. Tansley, and K. Tolle, *The Fourth Paradigm: Data-intensive Scientific Discovery*. 2009, Remond, WA: Microsoft Research.

7. Committee on Science, E., and Public Policy, *Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age*. 2009: The National Academies Press.

8. Stodden, V., *The Legal Framework for Reproducible Scientific Research: Licensing and Copyright.* Computing in Science & Engineering, 2009. 11(1): p. 35-40.

9. Marchionini, G., et al., *Curating for Quality: Ensuring Data Quality to Enable New Science: Final report: Invitational Workshop Sponsored by the National Science Foundation*, 2012. p. 119.

10. Whitlock, M.C., *Data archiving in ecology and evolution: best practices.* Trends in Ecology & Evolution, 2011. 26(2): p. 61-65.

11. Buckingham Shum, S., et al., *Towards a global participatory platform.* The European Physical Journal Special Topics, 2012. 214(1): p. 109-152.

12. Newman, G., et al., *The future of citizen science: emerging technologies and shifting paradigms.* Frontiers in Ecology and the Environment, 2012. 10(6): p. 298-304.

13. Organisation for Economic Co-operation and Development, *OECD Principles and Guidelins for Access to Research Data from Public Funding*. 2007.

14. Roose, K., *Meet the 28-Year-Old Grad Student Who Just Shook the Global Austerity Movement*, in *New York*2013.

15. Barlett, T., *Former Harvard Psychologist Fabricated and Falsified, Report Says*, in *The Chronicle of Higher Education* 2012.

16. Bhattacharjee, Y., *The Mind of a Con Man*, in *The New York Times Magazine*2013.

17. Heidorn, P.B., *The Emerging Role of Libraries in Data Curation and E-science.* Journal of Library Administration, 2011. 51(7-8): p. 662-672.

18. Gold, A., *Data Curation and Libraries: Short-Term Developments, Long-Term Prospects*, 2010.

19. Monastersky, R., *Publishing frontiers: The library reboot.* Nature, 2013. 495(7442): p. 430-2.

20. Parsons, T., S. Grimshaw, and L. Williamson, *Research Data Management Survey*, 2013, The University of Nottingham.

21. Ribero, C. and M.E.M. Fernandes, *Data Curation at U. Porto: Identifying current practices across disciplinary domains.* IASSIST Quarterly, 2011. Winter: p. 14.

22. Provost's Task Force on the Stewardship of Digital Research Data, *Research Data Stewardship at UNC: Recommendations for Scholarly Practice and Leadership*, 2012.

23. Peters, C. and A.R. Dryden, *Assessing the Academic Library's Role in Campus-Wide Research Data Management: A First Step at the University of Houston.* Science & Technology Libraries, 2011. 30(4): p. 387-403.

24. Scaramozzino, J.M., M.L. Ramirez, and K.J. McGaughey, *A Study of Faculty Data Curation Behaviors and Attitudes at a Teaching-Centered University.* College & Research Libraries, 2012: p. 17.

25. Rice, R. and J. Haywood, *Research Data Management Initiatives at University of Edinburgh.* International Journal of Digital Curation, 2011. 6(2): p. 13.

26. Choudhury, G.S., *Case Study in Data Curation at Johns Hopkins University.* Library Trends, 2008. 57(2): p. 211-220.

27. Cragin, M.H., et al., *Data sharing, small science and institutional repositories.* Philos Trans A Math Phys Eng Sci, 2010. 368(1926): p. 4023-38.

28. Parham, S.W., *Testing the DAF for Implementation at Georgia Tech*, in *IDCC*2010: Chicago, IL.

29. Parham, S.W., J. Bodnar, and S. Fuchs, *Supporting tomorrow's research: Assessing faculty data curation needs at Georgia Tech.* College & Research Libraries, 2012. 73(1): p. 10-13.

30. Parham, S.W. and C. Murray-Rust, *Stewardship of Research Data*, 2011.

31. Jones, S., S. Ross, and R. Ruusalepp, *The Data Audit Framework: A toolkit to identify research assests and improve data management in research led institutions*, in *5th International iPRES Conference*2008.

32. Parham, S.W. and C. Doty, *NSF DMP Content Analysis: What Are Researchers Saying?* Bulletin of the American Society for Information Science and Technology, 2012. 39(1): p. 37-38.

33. Georgia Institute of Technology. *Intellectual Property Policy*. 2007 May 2, 2013]; Available from: http://www.academic.gatech.edu/handbook/general_institute_policies/50.1_intellectual_property_policy.htm.

34. University System of Georgia, R.M.a.A. *Records Retention Schedules*. 2010 May 6, 2013]; Available from: http://www.usg.edu/records_management/schedules/I/.

8. RESEARCH DATA PROJECT TEAM MEMBERSHIP

Bill Anderson*, Digital Library Developer

Mary Axford*, Librarian for International Affairs, Psychology, and Public Policy

Jon Bodnar*, Previous Librarian for Literature, Communication and Culture

Chris Doty, Reference & Subject Librarian (Materials Science and Engineering and Physics)

Sara Fuchs*, Previous Digital Initiatives Librarian

Katie Gentilello, Digital Projects Coordinator

Marlee Givens, Georgia Knowledge Repository Manager

Wendy Hagenmaier, Digital Collections Archivist

Chris Helms, Network Administrator

Heather Jeffcoat, Web Program Manager/ Systems Librarian

Lisha Li*, Librarian for Civil and Environmental Engineering

Lizzy Rolando, Research Data Librarian

Alison Valk, Librarian for Computer Science

Susan Wells Parham, Scholarly Communication and Digital Curation Services Department Head

*Indicates former group members.