# Stewardship of Research Data

Georgia Tech Library & Information Center

Sponsor: Catherine Murray-Rust

Project Lead: Susan Wells Parham

## Executive Summary

The Stewardship of Research Data Proposal recommends the development of an institute-wide data stewardship framework, with a focus on providing long-term access and preservation of Georgia Tech research data. Georgia Tech is, and should continue to be, recognized globally as a source for quality research data. To continue in this capacity, we require the institutional ability to capture, manage, and retain the valuable data sets produced by our researchers.

The goal of this proposal is to present the rationale for a coherent set of research data services, including institutional policies and technical infrastructure, with coordination among campus service providers. Proposal recommendations include areas for improvement; an institute-level working group to establish responsible research data stewardship; a central web presence for research data information; and a strategy for the development of a data repository with external storage and replication. Financial resources are required for research data curation and preservation technology and services.

## Introduction & Background

The February 2011 special issue of *Science* Magazine is devoted to the subject of the collection, curation, and access to research data. Journal editors note that these data practices are essential to addressing the broad demands placed upon research communities to improve public health and welfare, spur economic recovery, and provide transparency of research operations.[1] In addition, responsible data management preserves confidential information, protects data from unethical practices such as falsification and fabrication, and can clarify the ownership of intellectual property rights.

The proper stewardship of research data facilitates new advances in scientific discovery by increasing the visibility, validity, and re-use of data. As federal funding agencies and other sponsors increasingly require researchers to provide evidence of sharing, preserving, and properly managing their data, Georgia Tech has a strategic opportunity to invest in our researchers by supporting these efforts.

A stated mission of Georgia Tech is for members of its community to become leaders in improving the human condition around the globe. One ambition put forth in the Georgia Tech strategic plan, *Designing the Future* is to "increase the quality of the research at Georgia Tech in a manner that commands the attention of the world."[2] The inclusion of research results and data in widely searched (and well maintained) repositories will increase the prominence of Georgia Tech research,

[1] "Introduction: Challenges and Opportunities" *Science.* 331(6018), 692-693. http://www.sciencemag.org/content/331/6018/692.short

[2] Designing the Future: A Strategic Vision and Plan, p. 12 http://www.gatech.edu/vision/sites/kraken.gatech.edu.vision/files/Georgia_Tech_Strategic_Plan.pdf

demonstrate its relevance, and encourage further scholarly inquiry. It will help answer the question: "What does Georgia Tech think?"

An institute-wide framework for research data stewardship requires institutional policies, technical infrastructure, and complementary data services. Developing this framework will require a consolidation of expertise from across campus, including input from researchers, administrators, technologists, and librarians.[3]

Stewardship activities include:

- Data management planning prior to research
- Maintaining data management best practices during the collection or creation of data
- Data security, storage & back-up while research is on-going
- Secure retention & disposal of selected data after research completion
- Data curation throughout the data lifecycle
- Sharing & publishing of research data for validation and re-use
- Archiving data & providing long-term access
- Implementation & development of metadata standards
- Compliance with institute policies, legal requirements, and ethical standards

The Georgia Tech Strategic Technology Investment Collaboration (STIC) provides a method for vetting proposed IT programs at the institute level. The establishment of a data stewardship framework at Georgia Tech requires both the strategic, institute-wide development of policies and programs, and the related resource allocation for IT infrastructure and tools. STIC presents the necessary means to prioritize IT projects, allocate resources, and increase stakeholder engagement.[4]

One aspect of data stewardship, *data curation*, is defined as "the active and on-going management of research data through its lifecycle of interest and usefulness to scholarship, science, and education; curation activities and policies enable data discovery and retrieval, maintain data quality and add value, and provide for re-use over time."[5] The Georgia Tech Library has a history of effectively curating GT scholarship and research in both print and digital form, including eighty-six

[3] Data stewardship framework and graphic on p. 3 adapted from "Creating a Data Management Framework" *Australian National Data Service* http://ands.org.au/guides/dmframework/data-management-framework.pdf

[4] See Tom Maier's EDUCAUSE presentation "Is Agile IT Governance an Oxymoron in Higher Education?" http://www.educause.edu/sites/default/files/library/presentations/E10/SESS089/EDUCAUSE%2B2010%2Bupload.pdf

[5] From the course description for *LIS 590DC: Foundations of Data Curation* at the University of Illinois Graduate School of Library and Information Science

years of theses and dissertations, as well as conference proceedings and the technical reports from sponsored research.[6] Research data are an additional aspect of this scholarship.

| Research Data Stewardship | | |
|---|---|---|
| **Policies** | **IT Infrastructure** | **Services** |
| • Data management<br>• Data ownership<br>• Data security<br>• Data retention & disposal<br>• Data sharing & reuse<br>• Intellectual property<br>• Ethical standards & relevant laws | • Data & metadata storage and backup<br>• Network connectivity<br>• Identity management, authentication & authorization<br>• Software development & support<br>• Data visualization<br>• Collaborative research environments<br>• High performance computing | • Data management planning<br>• Identification of datasets<br>• Metadata services<br>• Data & metadata formats and standards<br>• Data discovery & access<br>• Data sharing & publishing<br>• Data curation<br>• Data preservation |

## Improvement Opportunity

This initiative will improve the quality of data management and curation at Georgia Tech, the availability of policies regarding data retention and security, and the access to and preservation of research data.

Opportunities include:

- Increased visibility of Georgia Tech research
- Increased availability of research data for validation and re-use
- Advancement of scholarship and innovation through long-term data sharing
- Strengthened support for researchers in meeting compliance requirements, providing them more time to focus on intellectual discovery
- The capture and retention of irreplaceable institutional assets
- Strategic resource deployment and decreased redundancy of effort institute-wide

Over the last several years, multiple federal agencies and professional societies have published reports on the potential for data-intensive research and the need for cyberinfrastructure initiatives (see references). In addition, current research shows that academic data management practices tend to be ad hoc, and that there is great risk of data loss due in part to the predominant amount of

[6] See, for example, the following digital collections in SMARTech:
*Theses and Dissertations* http://smartech.gatech.edu/handle/1853/3739/browse?type=dateissued
*OSP Technical Reports* http://smartech.gatech.edu/handle/1853/9273/browse?type=dateissued

"small science" sponsored research in academia.[7] Combine these trends with the advent of funding agency provisions such as the NSF Data Management Plan (DMP) requirement, and a complex data environment emerges.

## IMPROVEMENT AREAS

Researchers must now provide detailed strategies on matters generally considered within the domain of library and information science and computer science, such as: metadata and content standards; data security; intellectual property; data publishing and licensing; data storage and backup; and digital preservation and long-term access.

Georgia Tech has an exciting opportunity to enhance the future of research by improving data stewardship practices on campus – by developing a robust data curation infrastructure, the Institute will remove from faculty the additional administrative burden placed upon them by data management and sharing mandates, allowing them to focus on innovative scholarship and research. The following categories are adapted from the Cornell RDMSG Planning Group document.[8]

### *Metadata:*

Metadata provides the context needed to find research data, use data accurately, preserve data, and assess the integrity of data. Metadata tracks the circumstances of an experiment or study (e.g., equipment used), making data meaningful. Because research data sets may go through many stages of processing and transformation, metadata provides an important management tool and is crucial at multiple points in the research process. Metadata types are extremely diverse, ranging from administrative information such as when and how data are generated, to descriptive information such as subject headings and keywords. One component of the NSF DMP is an explanation of the data and metadata standards to be used during research.

Because individuals with expertise in metadata content and format standards are needed, we should identify consultants willing to fill this capacity within their areas of discipline expertise. The addition of a science metadata librarian to the library staff to assist with metadata consultation and data repository information design is recommended. (***resources required)***

### *Protection of data:*

This area includes several related issues: intellectual property rights (copyright and usage rights); commercialization and licensing issues; and confidentiality and privacy issues. Individuals on campus with expertise in these areas should be identified and asked to serve on a rights committee.

[7] From Stewarding Excellence @ Illinois IT (SEI-IT) Project Team Report, which cites: Borgman, C.L., Wallis, J.C., & Enyedy, N. (2007) Little science confronts the data deluge: habitat ecology, embedded sensor networks, and digital libraries. *International Journal of Digital Libraries*, *7*(1-2), 17-30; Heidorn, P.B. (2008) Shedding light on the dark data in the long tail of science. *Library Trends*, *57*(2), 280-299; and Cragin, M.H., Palmer, C.L., Carlson, J.R., & Witt, M. (2010). Data sharing, small science, and institutional repositories. *Philosophical Transactions of the Royal Society A*, *368*(1926), 4023-4038.
[8] Meeting Funders' Data Policies: Blueprint for a Research Data Management Service Group (RDMSG)

A service gap exists on campus in the area of intellectual property rights, and the issue of research data ownership and usage rights can be confusing. Researchers often assume that the data they collect or generate belongs to them, particularly because they have conducted research based upon their own ideas and hypotheses. In addition, researchers initially take primary responsibility for data stewardship activities such as data management, retention and disposal. However, in most cases, the institution owns the rights to the data, either because researchers are working for hire or because federally sponsored project agreements are executed between the sponsor and the institution. A question to be addressed via published policies is the latitude Georgia Tech gives researchers with regard to the retention and use of research data. Undergraduates, graduates, and postdoctoral fellows often ask if they are allowed to take copies of their data upon departing Georgia Tech.

Uncertainties about what researchers are legally allowed to do with data can inhibit sharing and reuse. Creative Commons and its offshoot, Science Commons, have developed a number of innovations in the area of licensing aimed at facilitating open dissemination, sharing, and use of a wide variety of information, including data. The addition of a designated intellectual property expert on campus, either in the library or elsewhere, is recommended.

***Access:***

Providing access to research data requires the appropriate contextual metadata, software, and hardware. During the 2010 Fall semester, the library conducted a research data assessment survey, with sixty-three results from faculty, academic professionals, and graduate students from across campus. When asked to identify useful services, 67% of respondents chose "tools for sharing research data."

Since the NSF DMP requirement went into effect in January 2011, multiple researchers from the Colleges of Sciences, Engineering, and Computing have requested library support in meeting the NSF data sharing policy (in addition to assistance in developing data management plans). Unfortunately, we have not been in a position to meet all of these inquiries, including a recent request to provide discovery and long-term access for approximately 50 gigabytes of chromatograms and mass spectra. Because the average time to award notification is 300 days, the requests for data curation services based on NSF data requirements should rise in Fall 2011.

Discovery & Access Solutions:

- SMARTech, Georgia Tech's institutional repository, is an appropriate vehicle for providing access to small-scale data sets which meet specific criteria (e.g., data are in final form and can be made openly accessible)

- External, domain-specific repositories (when available) are possible solutions for satisfying access requirements[9]
- The development of a Fedora-based research data repository hosted by the library (along with additional supporting personnel) is recommended (***resources required)***

***Preservation:***
Costs for short-term storage and backup (i.e., while research is on-going) at the college-level might reasonably be written into research grants. However, storage alone is not sufficient to ensure the long-term usability of digital data. Data curation activities such as implementing the appropriate technologies and standards, monitoring the integrity of stored content, migrating media and file formats, replicating and distributing content, etc. must be implemented for effective digital preservation. The full cost of digital preservation is not well understood, and costs for preserving digital data will continue to be incurred well after a research grant ends.

For small-scale data sets housed in SMARTech, the MetaArchive provides a solution for preservation. The MetaArchive Cooperative is an international effort for the preservation of electronic scholarly materials through the Library of Congress' National Digital Information Infrastructure and Preservation Program (NDIIPP) using the LOCKSS (Lots of Copies Keep Stuff Safe) technology. The Library is currently a partner of the MetaArchive Cooperative.

73% of respondents to the library's research data assessment survey chose "data storage and preservation" as a useful service. Development of a library hosted data repository with external storage and replication for preservation is recommended. Storage solutions include external storage and replication using Georgia Tech's high-performance, private network capabilities. ***(resources required)***

## Desired Goals & Specific Outcomes

This project addresses Goal 2 of the Georgia Tech Strategic Plan:

**Goal 2—Sustain and enhance excellence in scholarship and research.**

*Strategy 4: Demonstrate relevance and vitality by investing in faculty and infrastructure*

Building the institutional programs and infrastructure to address funder requirements regarding research data will increase researcher competitiveness for future awards. In addition to planning for data management within proposals, researchers will also be judged on compliance with data management plans when competing for future awards.

In addition, the inclusion of research data in widely searched data repositories will increase the prominence of Georgia Tech research, demonstrate its relevance, and encourage further scholarly

[9] DataCite maintains a list of discipline-specific research data repositories: http://datacite.org/repolist

inquiry. Georgia Tech should be recognized as a source for quality research data. To be recognized as such, we require the policies and technical infrastructure to capture, maintain, and provide access to valuable data sets produced by our researchers

*Desired goal:* a responsible research data stewardship framework with an alignment of data services, policies, and technical infrastructure among campus service providers. Specialized data assistance available for researchers as the need arises.

## Risk Assessment

In FY 2010, just over 20% ($115,240,436) of the federal research grants and contracts received by Georgia Tech were awarded by NSF and the U.S. Department of Health and Human Services, both of whom require the sharing of research data.[10] As proposed in the UIUC SEI-IT Project Team Report, we can weigh the potential impact of inadequate institutional capacity to meet granting agency data requirements against the total research funding from a given agency.

| Agency | FY 10 Amount | % Federal Sponsored Funding |
|---|---|---|
| **NSF** | $83,952,428 | 15% |
| **U. S. Department of Health and Human Services** | $31,288,008 | 5.6% |

While difficult to predict specific at-risk dollar amounts, it is evident that Georgia Tech will be less competitive in the pursuit of research grants and contracts without the appropriate stewardship of research data. Authors of the UIUC report also note the likelihood of unprepared institutions trailing behind their peers in the evolving global network of research data resources as additional research funders enact data management and sharing mandates. In late June 2011, the NEH Office of Digital Humanities announced a new grant program that includes a data management plan requirement based upon the NSF requirement.

Another measure of risk is the missed opportunity to enhance the Institute's reputation and the prominence of Georgia Tech scholarship. Without the services and infrastructure in place to support broad data sharing, Georgia Tech research outputs will not be as widely communicated, and will thus be less widely recognized and utilized world-wide.

Risks if no data stewardship actions are taken:

- Loss of original intellectual property & scholarship, especially in the case of "small science" (unfunded research or research with awards of less than $300,000)

[10] Figures from http://factbook.gatech.edu/content/awards-summary. NIH currently requires proposals for the amount of $500,000 and above to include a DMP, however, it is expected that they will soon require all proposals to share data.

- Missed opportunity to enhance the Institute's reputation and the prominence of Georgia Tech scholarship
- Hidden costs, such as unnecessary replication of experiments or redundant IT expenditures
- Risk to future grants that require detailed data management plans and subsequent fulfillment
- Liability related to audits or other legal actions

## Dependencies

There has been little institute-level planning for a coordinated data stewardship framework; however, a number of stand-alone services and technologies related to research data curation are currently available on campus:

**OIT: Centralized Storage**

- $1.70 per GB per year (current cost)
- Storage allocated according to pre-agreed requirement and invoiced on an annual basis
- OIT provides all backend maintenance
- Customer is responsible for data management and client (host/server) maintenance and setup

**OIT: High Performance Computing**

Partnership for an Advanced Computing Environment (PACE) provides a core suite of HPC services and resources available at no direct cost to faculty

- Infrastructure - data center space, power, cooling, racks, 24x7 data center operations
- Commodity Services - ethernet, high-speed scratch storage, back-ups, recovery, security
- Software - compilers, applications, license management
- Technical Services & Support - user support, hardware management, consulting, procurement

**College IT Services**

- Infrastructure
- Commodity Services - Data storage and backup during research cycle
- Software
- Technical Services & Support

**Library**

- Institutional repository (SMARTech) that may be used to meet data access & preservation requirements for small-scale data sets meeting specific criteria (e.g., open access, in final form)
- Responsibility for long term preservation and stewardship of the Institute’s intellectual assets
- Data management planning assistance
- Subject librarians with domain expertise
- Scholarly communication & digital services

- Assistance publishing to external repositories
- Hosted website containing a single point of contact for data services and information (recommended)

**Office of Executive Vice President for Research**

- Administration leadership for all research, economic development, and related support units

**Office of Sponsored Programs**

- Specialized educational, informational, and technological support to research administrators and faculty

Other dependencies include faculty members with specialized expertise in related areas, such as GTRI Research Scientist Bill Underwood and Spencer Rugaber in the College of Computing.

## Recommendations

Data management decisions should be made at the most immediate appropriate level–individual investigator, research group or lab, academic department; however, an element of central coordination is essential for the development of an institutional data stewardship strategy. Support for institutional data stewardship is a distributed multi-unit activity, a partnership among researchers, technologists, librarians, and other campus professionals which combines bottom-up and top-down approaches.

As evident from the *Dependencies* section above, the Georgia Tech Library is but one piece of the data stewardship framework. This proposal provides recommendations at both the institute and library levels.

### INSTITUTE

We recommend the creation of a virtual organization,[11] the Research Data Stewardship Group, sponsored by the GT Library, OIT, and the Office of the Executive Vice President for Research to coordinate data management services and provide a single web presence for all issues related to research data management and curation.[12] It is recommended that this group propose institutional policy to promote research integrity as it relates to the management and preservation of research data. Because Georgia Tech consists of diverse communities of practice with shifting requirements, institutional policies and services will continue to develop over time.

Proposed structure for the virtual organization:

[11] For a definition of *virtual organization* and *virtual team* see *Reference for Business*: http://www.referenceforbusiness.com/management/Tr-Z/Virtual-Organizations.html

[12] See http://data.research.cornell.edu/ or http://rci.ucsd.edu/ as examples. Our institute-level recommendation relies heavily on the Cornell example (citation in references).

- Joint sponsorship by the Executive Vice President for Research, the Dean and Director of Libraries, 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 & day-to-day operations
- Implementation teams to carry out specific tasks

Researcher awareness of data services is of great importance–as noted in a recent UK study, data loss and other data storage issues are often avoidable, but researchers are unaware of the services available to them.[13] A central web portal will serve as an outreach vehicle for data management planning and related services, as well as a single point of contact for those requiring assistance. A process should be developed for routing researchers' data requests. The site will be based on the Drupal content management system and hosted by the library, with content maintained by campus partners.

| Research Data Stewardship Group | | | | |
|---|---|---|---|---|
| **Sponsors & Advisors** | Executive Vice President for Research | Chief Information Officer | Dean and Director of Libraries | Faculty Advisory Board |
| **Management** | Management Council | | Coordinator | |
| **Implementation Teams (examples)** | Services Assessment | Web Presence | Sustainability | Outreach |
| **Service Providers (examples)** | Library | OSP | OIT | College IT |

## LIBRARY

To meet the rapidly growing number of researcher requests for data sharing and archiving capabilities, we recommend that the library develop a program specifically for the curation of research data. This program will augment successful digital curation strategies and programs already in place, such as the library's curation of nearly 16,000 Georgia Tech theses and dissertations in electronic form. To meet this building need, the library will require significant IT assets, including software, hardware, and personnel. We recommend a phased approach over a 3-5 year period, during which time we will conduct research and development regarding tools and best

[13] Jones, S., Ball, A. & Ekmekcioglu, C. (2008). The data audit framework: a first step in the data management challenge. *The International Journal of Digital Curation*, *2*(3), 112-120.

practices to facilitate data sharing, publishing, and archiving. A data curation program will consist of 1) a data repository and related services; and 2) personnel in support of the repository:

1. A Fedora-based Research Data Repository for providing long-term access to selected research data.

   Current digital curation technology strategies:
   - SMARTech: DSpace repository for digital Georgia Tech scholarly publications such as theses & dissertations, conference proceedings, and technical reports; DSpace is open source software supported by the DuraSpace organization[14]
   - Tiered storage: Oracle SAM/FS managed 20 T disk, SL500 (30 slots) LTO @ 24T

   Future research data curation technology strategies:
   - Fedora (Flexible Extensible Digital Object Repository Architecture) – open source repository architecture supported by DuraSpace
   - Trial three data curation solutions with Fedora backend:
     - Islandora – open source framework combining Drupal & Fedora
     - EULfedora – Python library with optional Django integration
     - Hydra – open source Ruby on Rails framework for repository applications
   - DuraCloud software – open source service supported by DuraSpace
   - External storage and replication via Southern Crossroads Storage as Service (two-tier architecture):
     - On premise cloud (Atmos hardware & software infrastructure)
     - Distributed cloud access services layer (at Georgia Tech & BoR in Athens)

2. Permanent personnel to support research data consultation, curation and preservation:
   - 1 FTE Metadata Librarian – hire in year 1
   - 1 FTE System Administrator – hire in year 2
   - 1 FTE Data Curation Librarian – hire in year 2
   - 2 FTE Programmers – hire in years 1 & 3
   - 1 GRA Programmer – ongoing throughout project

## Approach

The STIC Data Curation Virtual Community Task Force will play a critical role in determining the requirements for a data stewardship framework at Georgia Tech and the methods for developing this framework. Membership expertise will be vital to establishing key areas for technical development, increasing stakeholder feedback, and planning financial sustainability and assessment at the Institute level.

[14] See http://www.duraspace.org/

## INSTITUTE

It is recommended that the Research Data Stewardship Group serve as a coordinating body for data stewardship at Georgia Tech. If implemented, the group's management council will play a critical role in evaluating researcher and funder requirements, aligning existing services with requirements, and recommending additional services as needed. It will lead the development of a centralized web presence to provide data stewardship information for researchers. The council should also consider long-term sustainability and financial planning to support the ongoing costs of research data preservation and long-term access. According to a poll conducted of 1700 *Science* peer reviewers, 80% of respondents do not have sufficient funding within their lab or research group for data curation.[15]

Specific policy and service areas which the group may review or provide information about include:

- Data Management Planning & Best Practices
- Secure Data Storage & Backup
- High Performance Computing
- Data Access & Discovery
- Data Curation & Preservation
- Metadata Services
- Privacy & Confidentiality
- Responsible Conduct of Research
- Intellectual Property

The faculty advisory board will provide the management council and sponsors with feedback regarding requirements and services for the institute. An initial board of at least three members, ideally involved with data-intensive research and with an interest in data stewardship and curation, should be recruited by the Executive Vice President for Research.

## LIBRARY

As part of the recent strategic planning process conducted by all Tech academic units, the Library designated data curation as a significant direction for the organization, in line with both the Georgia Tech strategic plan and the Library's responsibility for the long term preservation and stewardship of the Institute's intellectual property. The library's description of this action plan is to "develop campus partnerships to collect, manage, share, and preserve Georgia Tech digital research data."[16]

In preparation for this initiative, the Library Research Data Project Team has taken a number of steps toward building awareness of data stewardship issues, assessing research data-related needs, and developing campus partnerships. Actions taken thus far include:

[15] "Introduction: Challenges and Opportunities" *Science.* 331(6018), 692-693. http://www.sciencemag.org/content/331/6018/692.short

[16] From the internal document "GT Library Strategic Planning Action in Progress"

1) A campus-wide research data needs assessment survey, with sixty-three responses from across the colleges and research centers.[17] Initial results from the survey revealed that a number of faculty members create data sets that exist in accessible formats, are relatively small in size, and can be made publicly available for an indefinite amount of time. These criteria fit our collecting policies for the GT institutional repository SMARTech, so repository managers have made recommendations and policies for the data sets the library is able to accept at this time. Over half of the survey respondents asked to participate in follow-up interviews to discuss their needs regarding research data management, access, and preservation. 73% of respondents expressed a need for data storage and preservation; 67% a need for tools for sharing data. http://library.gatech.edu/research-data/

2) Promotion & outreach regarding data management requirements and best practices targeted to faculty, graduate students, and research administrators. Outreach included presentations at meetings, workshops, articles, web guides, and print materials. Training and consultation provided to researchers developing data management plans. For example: http://libguides.gatech.edu/content.php?pid=123776&sid=1514980

3) Collaborations with other campus units with regards to research data stewardship: Office of Sponsored Programs – workshops on the NSF DMP requirement and Data Stewardship RCR webinar (responsible conduct of research); Graduate Research Ethics Program – RCR Ethical Data Management lectures; OIT and College IT directors – initial conversations regarding researchers' requirements for data storage, backup, and long-term access and preservation. Collaborations with colleagues at peer institutions regarding data curation best practices.

### *Five Year Library Action Plan:*

### Year 1:

1) **Hire librarian to focus on the metadata issues** related to the research data curation lifecycle, while providing metadata consultation to Georgia Tech faculty and graduate students. This position: works on initiatives related to the discovery and preservation of digital research data, and creates local documentation on metadata standards and metadata application guidelines; participates in the development of the research data repository and works to improve access to resources such as locally-created data sets and digital research collections; stays abreast of scientific research trends, data documentation tools, and standards important for data exchange, reuse, and interoperability; advises on digital preservation strategies, including metadata used for digital repositories; suggests methods

[17] Survey based upon the Data Asset Framework, a data assessment methodology developed in the UK: http://www.dcc.ac.uk/resources/tools-and-applications/data-asset-framework

for streamlining or automating metadata creation and management, using various tools for metadata manipulation and scripting. (Susan Parham)

2) **Hire programmer to assist in the exploratory phase** of the data curation program. This position: works on initiatives related to the discovery and preservation of digital research data by testing, developing and integrating software and web applications; works collaboratively with content curators as well as fellow technologists; shares advancements in standards, software development practices, and IT trends; constantly refines his or her skill set and applies new knowledge and techniques; has proficiency with MVC frameworks, multiple programming languages, and more than one programming language paradigm; has enthusiasm for staying informed about cutting-edge technologies for use with software development initiatives, and for engaging the broader digital library software development community; has excellent problem-solving skills. Desired experience includes knowledge of digital repository and content management services and understanding of semantic technologies such as descriptive metadata and RDF ontologies. (Susan Parham or Head IT&D)

3) **Develop infrastructure for the research data repository**: purchase and install servers; conceptualize and implement Fedora repository architecture. (Chris Helms & Keith Gilbertson)

4) **Trial three previously identified data curation technologies** (Islandora, EULfedora, and Hydra) with Fedora backend. (Keith Gilbertson, Chris Helms, Susan Parham, along with additional librarian and programmer, as hired)

5) **Conduct a research data management and curation use case** to determine cost and service models.[18] Study specific research data practices and workflows in partnership with a Georgia Tech research group, appropriate subject librarian, research data librarian (Susan Parham), digital technologies librarian (Keith Gilbertson), and network administrator (Chris Helms). Metadata librarian and additional programmer will participate as hired. Participating researchers will receive:
   a. Full audit of data assets and data management practices with recommendations for improvements[19]
   b. Metadata analysis – assistance with metadata creation to make data discoverable and available for re-use
   c. Data format analysis
   d. Library-initiated proof of concept using one of three data curation solutions

[18] As an example, see the UCSD pilot program: http://adminrecords.ucsd.edu/Notices/2011/2011-5-26-3.html

[19] Using the Data Curation Profiles Toolkit developed by Purdue University Libraries and University of Illinois GSLIS http://www4.lib.purdue.edu/dcp/

e. External storage and replication via Southern Crossroads

6) Members of the Library Research Data Project Team will **conduct additional IRB-approved research** (approval received) to determine gaps in data curation services provided to researchers. We will identify the main data issues faced by researchers, discover areas where data are at risk, and plan for future infrastructure requirements. This needs assessment will take the form of a) a series of interviews with Georgia Tech researchers who have already volunteered to participate; b) a revised version of the online survey previously conducted with a targeted group of researchers. (Susan Parham, Alison Valk and others)

**Year 2:**

1) **Hire systems administrator** to assist in installation, configuration, maintenance and support of software, systems and related networking infrastructure on physical and virtual servers in support of digital and data curation programs. This position: installs, rebuilds or migrates existing servers; configures hardware and virtual machines, applications, peripherals, services, networking, storage; creates and maintains user accounts, security, permissions, and file systems; performs ongoing support and maintenance for data curation systems and related applications, including, but not limited to hardware upgrades, networking, performance tuning, monitoring, alerting and backup systems; has experience managing physical and virtual server environments, maintaining web applications such as Apache, and basic database administration. (Chris Helms)

2) **Hire data curation librarian to focus on developing the information architecture** for the research data repository and to support work in the area of data curation project management. This position: works to increase the library's ability to collect, provide access to, and preserve research data; investigates user requirements for the data repository; develops repository work flows; serves as an agent between researchers and the library's data repository; contributes to the development of data management plans for funded projects and assists in data extraction, reporting, and monitoring compliance with established data management protocols; advises researchers on the management of data and provides technical support for use of analytical tools. The successful candidate will maintain competence with tools and methodologies for computationally centered, data-driven research (data mining, visualization, etc.) and have experience with relevant technologies, including one of the commonly used repository platforms (Fedora preferred). (Susan Parham)

3) Based on Georgia Tech case study and results from a planned study of costs throughout the data lifecycle, **define business and cost models for data curation services** at Georgia Tech Library. Continue research data management and curation use case, and identify

additional pilot partners for research data curation pilot program. (Susan Parham, et al.)

4) **Assess repository architecture and front-end solution**, making adjustments as necessary. Continue to review NSF and other funder requirements for data curation and preservation. As campus requirements change, modify systems and services in an agile manner. (Susan Parham, et al.)

**Year3:**

1) **Hire programmer to assist in the implementation phase** of the data curation program. This position: works on initiatives related to the discovery and preservation of digital research data by testing, developing and integrating software and web applications; works collaboratively with content curators as well as fellow technologists; shares advancements in standards, software development practices, and IT trends; constantly refines his or her skill set and applies new knowledge and techniques; has proficiency with MVC frameworks, multiple programming languages, and more than one programming language paradigm, enthusiasm for staying informed about cutting-edge technologies for use with software development initiatives, and for engaging the broader digital library software development community; has excellent problem-solving skills. Desired experience includes knowledge of digital repository and content management services and understanding of semantic technologies such as descriptive metadata and RDF ontologies. (Susan Parham or Head of IT&D)

2) **Further develop collection, preservation and other policies** for data repository and related applications. Develop promotion and outreach strategy regarding research data curation and preservation services. (Susan Parham, et al.)

3) **Implement multiple data curation projects** with research partners. Continue assessment of customer requirements, costs, and agreements. (Susan Parham, et al.)

**Year 4:**

1) With full-scale projects in production, conduct review of personnel time spent on the data curation program, evaluating time necessary for specific tasks, such as metadata analysis and customized programming. Conduct top down review of personnel roles, software, hardware, and customer needs. **Evaluate a tiered service model.** (Susan Parham, et al.)

**Year 5:**

1) **Conduct hardware renewal.** Prior to relegating servers and storage to a non-critical tier, staff time will be spent evaluating current needs and applying those towards selecting new equipment and technologies. (Chris Helms, et al.)

**2) Conduct outcomes assessment and full review.**

## Financial Analysis

### CURRENT

The Library already invests in providing long-term access and preservation services for sharing Georgia Tech scholarship and research. Library IT investments for these endeavors have thus far garnered a $99,630 allowance. This initial IT build out includes support for 100T of storage space at a rate of $0.97 per gigabyte. Maintenance, configuration, and support are provided by 2 FTE system administrators averaging $110,000 per year. Code development is supported by 2FTE programmers averaging $110,000 per year.

With a five year running cycle, another $2500 will be spent on maintenance renewals in year three and four. Renewal of equipment will occur towards the end of the five year lifecycle at which time approximately $50,000 will be required to invest in new hardware or services. Throughout this lifecycle, third party services may be acquired to offset renewal costs or provide enhanced preservation services. During the fifth year prior to relegating servers and storage to a non-critical tier, staff time will be spent evaluating current needs and applying those towards the culling of new equipment and technologies.

### PROPOSED

Approximate server costs with external storage and replication (at current cost) for Oracle X4170M2 averages $21,000 (see table below). Funding for year 2 and beyond would be subject to price per GB of data being stored.

Initial upfront costs include personnel costs, which will be phased in over a three year period. Two programmers will average a total $140,000 per year, with an additional systems administrator averaging $55,000. A graduate research assistant in years 1 & 2 of the project will require $20 per hour, 20 hours a week. Along with additional IT support, librarian expertise is required for metadata and data curation support; two additional FTE librarians will cost a total average of $100,000 per year.

Additional overhead costs are the support of facilities and operational costs such as power and cooling. Based on cost studies conducted at the University of Oxford, it is suggested that the cost of research data curation will decrease with economies of scale and over time, as the majority of costs are start up costs.[20]

At the institute level, participation on the Research Data Stewardship Group management council or on implementation teams is an inherent overhead cost to support research at Georgia Tech, and these activities should be considered within the normal scope of responsibilities of all participants.

[20] "Developing Infrastructure for Research Data Management at the University of Oxford" http://www.ariadne.ac.uk/issue65/wilson-et-al/

| **Startup Technical Costs*** | | |
|---|---|---|
| **Servers—option 1** | 2x Dell PowerEdge R410<br>Price per/unit: $7,663.00<br>5 year service contract | $15,326.00 |
| **Servers—option 2** | 2x Oracle X4170M2<br>Price per/unit: $8,932.00<br>2 year service contract* | $17,864.00 |
| **Storage—option 1 (currently available)** | OIT NAS Storage/Backup<br>2TB/1 year<br>1 year service agreement | $3,481.60 |
| **Storage—option 2 (proposed)** | SoX Cloud Storage<br>2TB/1 year<br>(based on proposal) | $983.40 |
| **Repository Architecture** | Duraspace Fedora Repository Architecture | open source |
| **Frontend** | Islandora, EULfedora, or Hydra | open source |
| **Total (with highest cost options)** | Option 2 Server & Option 1 Storage | $21,345 |

*Does not include overhead costs

| **Ongoing Personnel Costs*** | |
|---|---|
| **Metadata Librarian** | $50,000 |
| **Data Curation Librarian** | $50,000 |
| **Repository Programmer (2)**<br>**$70,000 / programmer** | $140,000 |
| **System Administrator** | $55,000 |
| **GRA** | $20x 20 hrs x 30 wks = $ 12,000 |

* Approximate salary first year of hire without fringe benefits

## Selected Peer Institutions

**Cornell**

http://data.research.cornell.edu/

Institute-level *Research Data Management Service Group* jointly sponsored by the Senior Vice Provost for Research and the University Librarian, with a faculty advisory board. Developed to present a coherent set of services to researchers; a unified web presence providing general information on data management planning, services available on campus, and standard language that may be used in data management plans in grant proposals; and a single point of contact that puts researchers in touch with specialized assistance as the need arises. Multiple repositories including Cornell University Geospatial Information Repository and DataStaR, a data staging repository.

MIT
http://libraries.mit.edu/guides/subjects/data-management/index.html
*MIT Libraries Data Management Team* provides a data management and publishing guide.

**Purdue**
http://d2c2.lib.purdue.edu/DMP.html
Institute-level *Data Management Plan Working Group*, a collaboration between Purdue University Libraries, Information Technology at Purdue (ITaP), the Office of the Vice President for Research (OVPR), and Sponsored Program Services (SPS) to develop assistance to identify data management needs for the completion of a data management plan (DMP) as prescribed either generally by NSF or by other funding agencies; to create a self-assessment guide that PIs can use to translate their data workflow into a DMP with minimal assistance; to collaborate to develop a model for ingest and discovery of data sets identified as target data using the HUBzero™ platform to create the Purdue University Research Repository (PURR).

**University of California – San Diego**
http://rci.ucsd.edu/
Institute-level *Research CyberInfrastructure* provides the computing, network, and human infrastructure needed to create, manage, and share data. Principal investigators are encouraged to use the campus's RCI in addressing federal sponsors' existing and new data management requirements. Collaborators include San Diego Super Computer Center, UCSD Libraries, Office of Research Affairs, and Office of Contract and Grant Administration.

**University of Illinois**
http://blogs.cites.illinois.edu/datasteward/
Institute-level *Data Stewardship Initiative* to ensure that the campus carefully and responsibly manages research data – partnership of relevant campus units built on infrastructure and systems. Members of the group include representatives from the Libraries, Office of the Vice Chancellor for Research, the Office of the CIO, and the Graduate School of Information and Library Sciences.

## References

Association of Research Libraries. (2006). *To stand the test of time: long-term stewardship of digital data sets in science and engineering*. Retrieved from http://www.arl.org/bm~doc/digdatarpt.pdf.

Association of Research Libraries. (2007). *Agenda for developing e-science in research libraries: ARL joint task force on library support for e-science final report & recommendations*. Retrieved from http://www.arl.org/bm~doc/ARL_EScience_final.pdf.

Beagrie, N., Eakin-Richards, L., & Vision, T. (2010, September). Business models and cost estimation: Dryad repository case study. *Proceedings of the 7th International Conference on Preservation of Digital Objects.* Retrieved from http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/beagrie-37.pdf

Association of Research Libraries. (2010). *E-Science and data support services*: *a study of ARL member institutions*. Retrieved from http://www.arl.org/bm~doc/escience_report2010.pdf.

Atkins, D. (2003). *Revolutionizing science and engineering through cyberinfrastructure: report of the National Science Foundation.* Blue-Ribbon Advisory Panel on Cyberinfrastructure. Washington, D.C.: National Science Foundation. Retrieved from http://www.nsf.gov/od/oci/reports/atkins.pdf

Gold, A. (2007). Cyberinfrastructure, data, and libraries, part 1. *D-Lib Magazine, 13* (9/10). Retrieved from http://www.dlib.org/dlib/september07/gold/09gold-pt1.html.

Gold, A. (2007). Cyberinfrastructure, data, and libraries, part 2. *D-Lib Magazine, 13* (9/10). Retrieved from http://www.dlib.org/dlib/september07/gold/09gold-pt2.html.

National Academy of Sciences. Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age. (2009). *Ensuring the integrity, accessibility, and stewardship of research data in the digital age*. Retrieved from http://www.nap.edu/catalog.php?record_id=12615

National Science and Technology Council. Committee on Science. Interagency Working Group on Digital Data. (2009). *Harnessing the power of digital data for science and society: report of the Interagency Working Group on Digital Data to the Committee on Science of the National Science and Technology Council.* Washington: Interagency Working Group on Digital Data. Retrieved from http://www.nitrd.gov/about/Harnessing_Power_Web.pdf

National Science Foundation. Cyberinfrastructure Council. (2007). *Cyberinfrastructure vision for 21st century discovery*. Washington: Cyberinfrastructure Council. Retrieved from http://www.nsf.gov/od/oci/CI_Vision_March07.pdf.

National Science Foundation. National Science Board. (2005). *Long-lived digital data collections enabling research and education in the 21st century* (No. NSB-05-40). Washington: National Science Board. Retrieved from http://www.nsf.gov/pubs/2005/nsb0540/

RDMSG Planning Group. (2010). *Meeting funders' data policies: blueprint for a research data management service group (RDMSG).* [Cornell University]. Retrieved from http://data.research.cornell.edu/sites/rdmsg/files/RDMSG1007.pdf

Stewarding Excellence @ Illinois IT (SEI-IT) Project Team. *Report*. University of Illinois. Retrieved from http://oc.illinois.edu/budget/it_project_team_report.pdf

Wilson, J.A.J., Fraser, M.A., Martinez-Uribe, L., Patrick, M., Akram, A. & Mansoori, T. (2010, Oct 30). Developing infrastructure for research data management at the University of Oxford. *Ariadne, 65*. Retrieved from http://www.ariadne.ac.uk/issue65/wilson-et-al/

## Acknowledgements

Members of the Georgia Tech Library Research Data Project Team contributed to this paper, with significant input from Chris Helms and Keith Gilbertson, and reference assistance from Lisha Li. A number of Georgia Tech colleagues provided feedback on draft versions. The Stewarding Excellence @ Illinois IT (SEI-IT) Project Team Report and *Meeting Funders' Data Policies: Blueprint for a Research Data Management Service Group (RDMSG)* written by colleagues at UIUC and Cornell, respectively, provided models for much of this proposal.

S. Parham