

## **Data Management at Georgia Tech**

Dr. Lew Lefton, Assistant Dean of Information Technology  
College of Sciences, Georgia Institute of Technology  
2-3-2016

### **I. An incomplete view of the GT data landscape**

Data (broadly interpreted) is a fundamental asset of Georgia Tech, and is arguably as important as our facilities and personnel. Our researchers produce and consume data in order to better understand things. This understanding leads to innovation, and our students learn and gain opportunities as we share our knowledge and experience around the world. Our business and financial operations depend on data to succeed and to measure our performance. But, like the story of the blind men and the elephant, different units of campus view data (and its management) differently. The attached "mind map" is an incomplete attempt to view the whole landscape at a high level (suggestions and reactions welcome).

OIT provides core infrastructure to support and manage campus level data (networking, some centralized services, etc.), however a significant amount of data (especially research data) is decentralized in labs and units where it is produced and consumed. Both OIT and the unit level CSRs often view data in terms of services and technologies. IT managers want to deliver the data to the proper audience, and preserve and protect it, too. The IT view of data is on the left side of the mind map.

Business units and campus leadership view data as Institute property which is subject to policies, contractual obligations, and other compliance requirements. Their view includes not only research data but also enterprise data (HR and Finance), student data (Registrar, LMS), and operational data (facilities, institute communications). The business view of data is on the right side of the mind map.

Researchers often view data as their professional lifeblood. They work hard to gather it and analyze it. Some want to share their data far and wide and others want to carefully control access. Regardless, they often expect simple, high performance access to their data anywhere, anytime, on any device. The researcher view of data is on the bottom of the mind map.

The top part of the mind map isn't really "owned" by a single unit on campus. This is where the management of the data lives. Budget for data (including systems, storage, etc.) flows through the usual channels from ELT to unit heads. Communication about who is doing what on campus is ad hoc at best. Different units handle data acquisition, deployment of data solutions, enforcement of data policies using their own internal workflows and practices.

NOTE: I am not necessarily proposing any changes at this point. This framework is more of a personal exercise to help me better understand the situation. It is quite possible that "doing nothing" is the best path forward since one could argue that we have all locally optimized and major change could cause as much harm as it does good.

## **II. Scope**

Research data is the primary focus of this document. The broader scope of enterprise data, student data, communications data, etc. may still be able to leverage some of the same technologies and ideas laid out here.

## **III. Considerations and steps toward a plan**

### **Identify Needs**

As our campus data footprint grows, we would be well served to identify both immediate and future needs. This is probably already known but the information is anecdotal and widely distributed around the researchers and IT professionals. One way to identify needs is to capture a collection of user stories. For example:

1. A researcher has a large dataset that she needs to share externally and internally.
2. GTRC signs a contract which includes use of HIPPA protected data.
3. New storage technology emerges which shows promise and should be piloted.
4. Small collaborative projects and teams need a convenient and flexible space to share data.
5. IT professional needs to build/grow systems to store and backup a variety of data, but is limited by budget/expertise.

### **Identify Solutions**

Once needs are known, we can determine which can be met in a centralized way with scalable solutions, and which needs are best handled by individual units. Ideally, we could develop a collection of solutions which have different attributes (see below) that will accommodate a variety of use cases, perhaps in combination. For example:

1. Data/Technology centric attributes of solutions (probably choose some but not all)
  - Handles large data
  - Security model (Accessible/Available vs restricted)
  - Speed/Performance
  - Reliable
2. Business centric attributes of solutions
  - Policy and compliance monitoring
  - Cost
  - Matches appetite for risk
  - Data Classification
3. User centric attributes of solutions
  - List of available resources
  - Communication of policies, guidelines
  - Clear understanding of options (pros, cons, features, costs)
  - Technical support and documentation for moving data in and out and sharing

**Identify resources and implement**

The next step is to develop a cost model (investment, charge back, direct and indirect costs) so that resources flow to meet needs (both centralized and unit based), driven by strategic priorities. This may (likely will) require additional investment. In addition, there may need to be some resource reallocation.

Finally, we define who is responsible for implementing the model. Perhaps this falls under a campus project management office, or perhaps a task force/governance body has oversight. There are top down approaches (e.g. hire a "Data Czar" with budget and authority) or bottom up approaches (e.g. create incentives for units to implement desired solutions). It may be that a technical layer of data management (IT) is separated from a logical layer of data access control and services (business and researchers) similar to a model like PACE uses for HPC.

