Data Management Position Paper

Clark Atlanta University
Peter Molnar, Michelle Y. Kassorla
{pmolnar,mkassorla}@cau.edu

Statement of the Problem

With the implementation of NSF/NIH grant requirements to supply a data management plan for grant proposals we recognized the need to establish a modern data management plan at Clark Atlanta University that would address current and future data archiving needs, privacy considerations for human subject tests, and data sharing.

Clark Atlanta University had not developed a data management plan, and had a loose system of storage, backup, and sharing systems based in departmental or research-group protocol. At this point there was no unified system that would insure the integrity, backup, archiving, and sharing of research data sets.

We saw a need to develop a comprehensive data management solution. This system must integrate the needs of various stakeholders, including researchers from various disciplines, Instructional Technology, the Division of Research and Sponsored Programs, and the Woodruff Library (a research library shared between Clark Atlanta University, Spelman and Morehouse Colleges, and the International Theological Seminary.)

Existing Data Management

A critical element of developing a data management infrastructure is the understanding of the researchers data needs. Our next step was to collect information about the way that data was currently stored at Clark Atlanta University.

We developed a survey that we sent to Faculty to assess their current and future data management needs.

From that survey, we found the following information: (This information will be updated when survey is fully completed.)

- 35.3% of respondents were aware of the NIH/NSF data management requirements.
- 52% had completed research.
- 50% of data was stored on computer hard drives, 28.6% in removable hard drive, 71% in a flash drive, 7.1% is on a departmental server, 41% on a CAU computer, 28% on
personal computers, and 28% stored on other medium.
• 38% share their raw data, 23% share data after removing personally identifiable information, and 38% do not share data.

The survey was deliberately kept short two only 8 qualitative questions. We plan to conduct a much more detailed survey to assess the requirements for backup and archiving, as well as network bandwidth requirements. Various studies about data management requirements and behaviors have been conducted. The literature suggests that these kind of surveys are best done in interview form.

Analysis of Data Management Plans in Submitted Proposals
We looked at the NSF guidelines, and analysed how well they were addressed in the data management plans.
1. the types of data, samples, physical collections, software, curriculum materials, and other materials to be produced in the course of the project;
2. the standards to be used for data and metadata format and content (where existing standards are absent or deemed inadequate, this should be documented along with any proposed solutions or remedies);
3. policies for access and sharing including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements;
4. policies and provisions for re-use, re-distribution, and the production of derivatives; and
5. plans for archiving data, samples, and other research products, and for preservation of access to them.

The result of the 6 data management plans of NSF and NIH proposals since January 18, 2011 are tabulated as follows.

<table>
<thead>
<tr>
<th>Proposal#</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Type of Data</td>
<td>student data</td>
<td>N/A</td>
<td>survey, test, video</td>
<td>Lab notes, exper-</td>
<td>research, student</td>
<td>student data</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>mental student data</td>
<td>data</td>
<td></td>
</tr>
<tr>
<td>2. Standards</td>
<td>Excel</td>
<td>N/A</td>
<td>jpeg, html, SPSS</td>
<td>(paper), Word, Excel, CIF, ASCII</td>
<td>jpeg, html,</td>
<td>N/A</td>
</tr>
<tr>
<td>3. Policies Sharing</td>
<td>CAU Policies</td>
<td>Pub: journals, conferences</td>
<td>public Project Website</td>
<td>Pub: journals, conferences</td>
<td>website, journals</td>
<td>project website</td>
</tr>
<tr>
<td></td>
<td>CAU plc Copyright</td>
<td>N/A</td>
<td>public Project Website</td>
<td>CAU plc Copyright</td>
<td>CAU plc Copyright</td>
<td>N/A</td>
</tr>
</tbody>
</table>


While most PIs addressed the five required items of the data-management plan, their interpretation varied. None of the proposals addressed the measurable amount of data that would occur, or indicated requirements on storage capacity. Almost all of the plans referred to university policies without specifying them.

**Status of research cyberinfrastructure**

There are three groups with computational research access to high performance computers, and one group with access to a DNA sequencer. Three systems are located the Science Research Center data center, and the DNA sequencer is located on a separate floor in the same building with its own air conditioning unit. The systems have combined about 100 TB of scratch disk space, but no long-term archiving solution.

The research cyberinfrastructure is part of the academic campus network, but it is not supported by the IT staff. Faculty PIs and their staff are responsible of the operations and maintenance of this equipment. Some of the systems operate in their own virtual subnet, but all data still shares the campus network infrastructure, and passes through firewalls and PC-virus scanners.

The IT department does not allow adjustments to the firewall for the GlobusOnline file sharing systems. Currently, IT is in discussion with researchers about the possibility of an independent research network that would exist outside of the current CAU firewall system. This research network would allow CAU to participate in XSEDE Campus Bridging, and utilize off-site storage and archiving solutions.

**Public Data and Document Repository: Digital Commons®**

The Atlanta University Center Robert W. Woodruff Library (AUC Woodruff Library) assists in the discovery, access, and preservation of the collective body of research emanating from the Atlanta University Center. By utilizing technology that facilitates the creation, publication, and preservation of research AUC Woodruff Library is able to promote, provide access to, and preserve the scholarship produced by the faculty and students of the Atlanta University Center. AUC scholarship can be deposited or created into the AUC Woodruff Library Digital Commons e-scholarship institutional repository. The e-scholarship repository uses the DigitalCommons software and hosting services provided by bepress. The e-scholarship repository operates in a virtual web environment and can be populated by the contributions of the faculty and students of the Atlanta University Center. The repository will support deposits of digital representations of text, media, presentations, data sets, etc.
Established in 2005, the AUC Woodruff Library Digital Commons e-scholarship institutional repository currently hosts approximately 800 items. Content hosted on the AUC Woodruff Library e-scholarship repository includes theses and dissertations, journals, newsletters, and bibliographies. Google Analytics provides an overview of website traffic including hits, visits, and locational statistics. As an example over the past 18 months there have been almost 17,000 unique visits to the e-scholarship repository. Specialized usage software permits tracking of full text downloads to the title level. Since 2005, there have been almost 53,000 full text downloads of resources. Benefits of contributing to AUC Woodruff Library e-scholarship repository include:

1. Support of open access initiatives;
2. Global exposure and discovery of research through search engines such as Google and Yahoo;
3. Promotion of new forms of scholarly publishing including born digital peer-reviewed journals;
4. Enhanced access to ephemeral documents, data sets, media files, technical reports, speeches, and student papers;
5. Provision of a tool to archive scholarly research.

Conclusion

Like many other institutions Clark Atlanta University has just begun to address Data Management, and to develop a DM infrastructure and policy. The first step was to establish a DM task force that includes research faculty, IT staff, CAU administration, and the Woodruff Library.

We need to gain a clear and detailed understanding of the researchers needs, and the requirements of funding agencies to develop a roadmap for infrastructure and policies.

The implementation and operation of additional infrastructure, such as a separate research network, or shared backup and archiving systems, will have an impact on the university budget.