Reliable Facts from Unreliable Figures: Comparing Statistical Packages in DSpace

Proposal for a General Paper
Open Repositories 2011

Bill Anderson, Chris Helms, and Sara Fuchs – Georgia Institute of Technology
Andy Carter – University of Georgia

To ensure that a web-based institutional repository is fulfilling its mission, repository managers and staff must have some way of gauging usage. It's easy to see how many items are submitted to a repository, but more difficult to see how these items are used. Failing to capture and display accurate download statistics can make it difficult to convince faculty members to deposit. These numbers "provide a powerful tool for repository managers and librarians to sell the importance of this innovative technology to faculty and funding authorities" (Organ, 2006). Collecting such statistics requires the use of some sort of statistical package.

Web-based statistics, however, must be treated with care. Available statistical packages – whether open-source, proprietary, or home-cooked – can take radically different approaches to evaluating, classifying and displaying data. In addition, logging and filtering methods can have a large impact on statistical accuracy, as can differing interpretations of such terms as "hit," "page view," and even "item." The Georgia Institute of Technology's repository team decided to test and compare a variety of statistical software packages and methodologies. Our goal is to determine a set of procedures and best practices for collecting and disseminating DSpace statistics." At an early stage, this effort was expanded to include the University of Georgia Knowledge Repository (UGA KR), with the possibility of including the GALILEO Knowledge Repository, a grant-based initiative to develop a statewide metadata repository. This project is now underway, with a goal of having preliminary results by June of 2011. We hope that a report on (and discussion of) these results will be a useful addition to the OR 2011 program.

SMARTech, Scholarly Materials and Research at Tech, is the Georgia Institute of Technology's institutional repository. Started in 2004, the IR houses over 32,000 items, and is ranked in the top 100 of repositories worldwide, and the top 20 of repositories in the United States and Canada (Ranking Web of World Repositories, January 2011). In using SMARTech, we've relied primarily on the built-in DSpace statistics package; this package has never been particularly flexible or robust, and our practices in using the statistics generated have not been as rigorous as we might have wished. When an upgrade problem resulted in incomplete statistics for several months, we attempted to fill the gap with AWStats, which was installed on the server. We noticed that some of the counts we got from AWStats weren't consistent with previous results. We had similar issues when we implemented a Google Analytics account for the repository. (We have not yet ironed out performance difficulties with DSpace's relatively new SOLR statistics package, but anticipate including it in our study once we finish fine-tuning it.) An example is provided by comparing the statistics for one of our more popular items, a

1999 Georgia Tech thesis used as the basis of a current $2.8 billion redevelopment project along a railroad corridor in Atlanta ("Belt line - Atlanta: design of infrastructure as a reflection of public policy" located at http://smartech.gatech.edu/handle/1853/7400):

	March 2007	April 2007	May 2007
Google Analytics	18	16	3
AWStats	96	88	25
DSpace Statistics	144	122	71

The University of Georgia Knowledge Repository (UGA KR), operating on DSpace 1.6.2, launched in August 2010 and is now home to almost ten thousand items, the majority of which are Electronic Theses and Dissertations. Compared to SMARTech, the UGA KR is but a babe in the life cycle of institutional repositories, with neither the scale nor hands on experience to compare. We share, though, the need to better comprehend how our collections are accessed and used. A preliminary glance at the available statistical packages, and their differing results, revealed gaps in our understanding of what was being measured, how it was being measured, and what it meant in relation to our users. The ability to leverage quality statistics is crucial for the UGA KR's nascent attempt to integrate itself meaningfully into the flow of scholarly communications at the University of Georgia. The difficulties faced when attempting to solicit faculty interest in IRs are well documented. We can be more responsive to the concerns of faculty by making available to them data that is meaningful (relevant to their contexts) and trustworthy (relevant to any context). High quality metrics are also a cornerstone when constructing the argument for sustainability during these tough economic conditions.

The University of Georgia and the Georgia Institute of Technology are partners, with several other colleges and universities, in the **GALILEO Knowledge Repository** initiative, an IMLS-funded grant to build a statewide IR system in Georgia. GALILEO (GeorgiA LIbrary LEarning Online) is the state's virtual library, providing access to a wide range of databases and other educational resources. Members in this project are building and populating their own repositories (some of which will be hosted by Georgia Tech). All IRs developed in this grant will be harvested and searchable through a central site to be known as the GALILEO Knowledge Repository. The GKR initiative is in the early stages of being implemented, but we hope initial experience with the project will add another perspective to our study.

Our project will focus on three questions:

1) What is the best way to capture statistics for a DSpace repository?
2) What statistics do we want to capture?
3) How do we best display these statistics to the end user?

To better understand the type of statistical data being collected, several controlled experiments will be executed. Each experiment will start with a vanilla DSpace 1.6.2 installation and will then be populated with material by both batch ingest and user submitted content. "Bots" consisting of both human and machine will be utilized to browse, search, and view data presented publicly

via the DSpace 1.6.2 instance. Within each scenario a combination of the following tools will be deployed and analyzed: SOLR, system statistics (log based), AWstats, and Google Analytics.

Scenario I will include DSpace 1.6.2, Tomcat, and the default SOLR package. Scenario II will then compare the existing system statistics package before and after conversion of log-based statistics in to a SOLR database. Scenario III will again employ DSpace 1.6.2 along with Tomcat+Apache (mod_proxy_ajp) and AWstats. Lastly, Scenario IV will take a close look at Google Analytics. Our goal will be to analyze what is known via the controlled experiment against what statistical data is presented from each package. Inconsistencies in statistical data will be marked and followed by recommendations to improve upon the methodologies employed. A package or combination thereof will then be selected for a follow up user study to gain more insight into the usage of statistics and the scholarly impact of an institutional repository.

We know that our repositories are being used, and generating a fair amount of traffic; but the numbers we generate are not answering some important questions. Who are our users? How much time do they spend navigating the site? How many of our users are from our institutions, and how many are not associated with our campus? Is material deposited into our repositories more likely to be found and used than the same material on another university website, on a faculty member's personal web site, or in a proprietary database? Do users browse SMARTech, or discover its assets through an external search engine and then leave? In order to answer these questions, we must have confidence in the reliability, accuracy, and flexibility of our statistics software and procedures. The aim of our project is to ensure that confidence, and provide guidelines and best practices for other DSpace users.

References

Organ, Michael. "Download Statistics—What Do They Tell Us? The Example of Research Online, the Open Access Institutional Repository at the University of Wollongong, Australia", D-Lib Magazine 12 (11) (November, 2006) http://www.dlib.org/dlib/november06/organ/11organ.html (accessed January 17, 2011).