Open Repository 2013 Proposal - DSpace Track

Peter Dietz
Software Developer at Ohio State University Libraries

Usage Statistics, powered by Elastic Search

We have decided to replace the backend implementation of Usage Statistics for DSpace to give us a system that allowed for more flexible queries so that we could more easily ask for answers to our questions about how our repository was being used, and what was being used. The end result of looking deeper into the default SOLR statistics, and the available alternatives lead us to building a new statistics system powered by Elastic Search, and we are quite pleased with the results. The new system is fast enough for our needs, and we contributed it to DSpace 3.0.

We display the statistics on a Dashboard page for each Community, Collection, or Item in our repository. The default report shows you a hybrid of Growth Statistics and Usage Statistics. The Growth Statistics comes straight from the DSpace database, but we use it to give insight into the Collection. The data output shows on the Dashboard in a variety of forms, either a table of numbers, or visualized by Google Visualization API through either line charts, or Geographical heat maps. We also allow the data to be downloaded in CSV format. The default time range we allow for statistics queries is the Last Five Years, but we have a date picker where we allow the user to choose their date ranges, the granularity of the output data is only currently given out in month intervals, however there’s no technical restriction of how granular we could output.

In the process of coming up with our set of data for our usage statistics, we analyzed our various log files that we have been storing since we went online with DSpace in 2004. No one log file was complete from that early beginning to present, so we were considering trying to weave all of the different data sources together, indeed we built the tools to extract the information, but we had diminished confidence in the number once stitched together since we were getting odd spikes and gaps. We ultimately decided to stick with using our dspace.log files as the data source as it was continuous from late 2007 to present. Further analysis we did on our usage history was to look for the anomalies in the data. There is a basic robots filter list, but it is no means comprehensive, just catching a few of the big search engine crawling robots. Thus we had massive amounts of use to the system which we didn’t want to count. Some of our other techniques for “bot hunting” included scoring each unique visitor (IP address) to the system, and scoring it for the total amount of hits it made to the system, and then sorting it to reveal very active visitors (certainly bots), next down to our administrative staff who were interacting with the system many hours a week, then down to an end user who only downloaded a single document from our system. All the “certainly bots” users we then researched their IP address, and user agent to determine if indeed they were a robot. We then added these robots to our robot filter-out lists, and then display our usage with those robots filtered out.

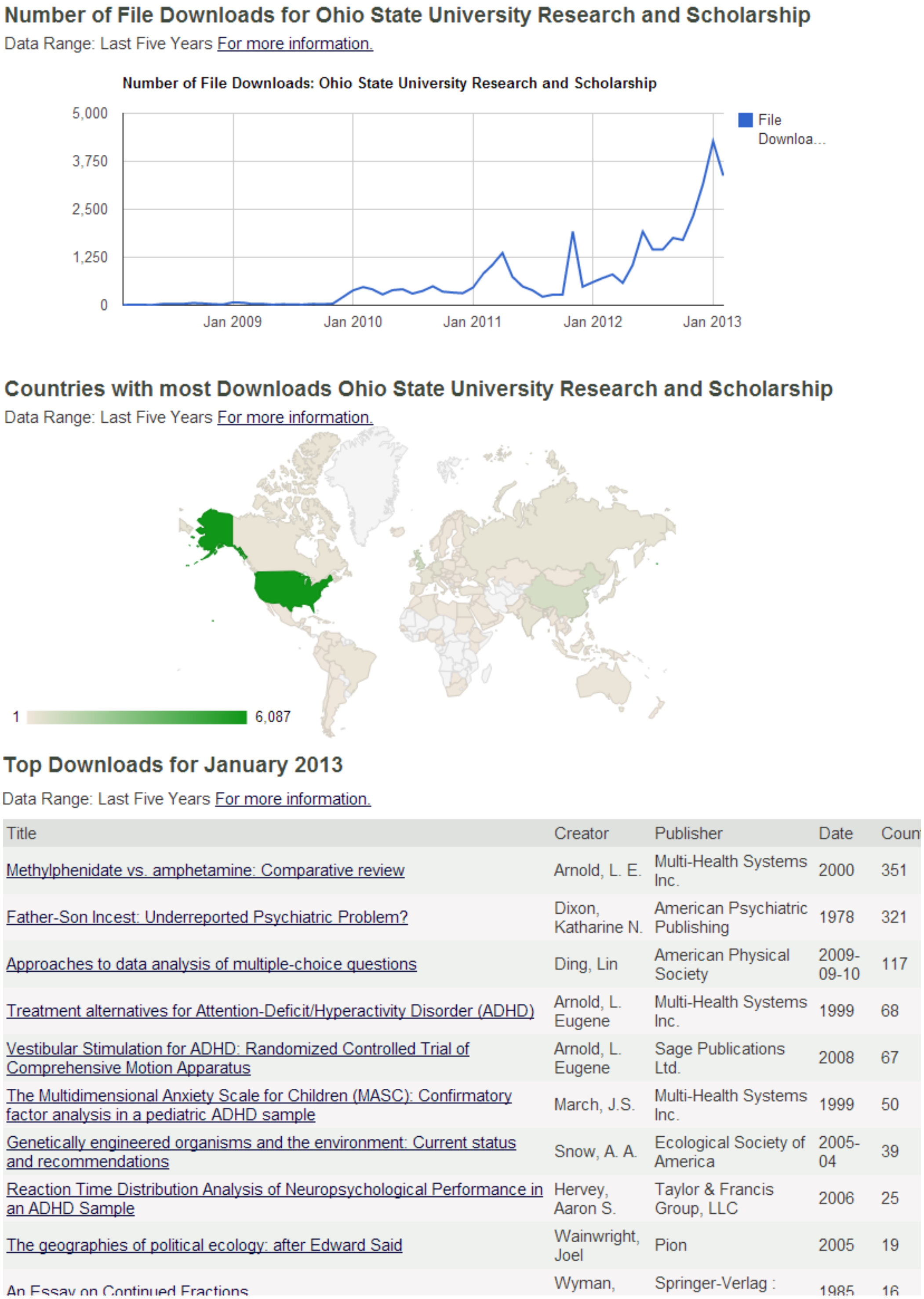
Once we had a trustworth data set for statistics, we then felt comfortable re-using this data to answer our questions about who is using the collections, where are they coming from, which objects are most consumed, what actions can we do to increase usage of our materials, and at the same time reducing our future workload when Communities ask for annual statistics, that now we can just hand them the keys and let them poke around.

Open Repository 2013 Proposal - DSpace Track

Peter Dietz
Software Developer at Ohio State University Libraries

Usage Statistics, powered by Elastic Search

Included is a screenshot of the Usage Statistics powered by Elastic Search in use at kb.osu.edu



Number of File Downloads for Ohio State University Research and Scholarship

Data Range: Last Five Years For more information.

Countries with most Downloads Ohio State University Research and Scholarship

Data Range: Last Five Years For more information.

Top Downloads for January 2013

Data Range: Last Five Years For more information.

Title	Creator	Publisher	Date	Coun
Methylphenidate vs. amphetamine: Comparative review	Arnold, L. E.	Multi-Health Systems Inc.	2000	351
Father-Son Incest: Underreported Psychiatric Problem?	Dixon, Katharine N.	American Psychiatric Publishing	1978	321
Approaches to data analysis of multiple-choice questions	Ding, Lin	American Physical Society	2009-09-10	117
Treatment alternatives for Attention-Deficit/Hyperactivity Disorder (ADHD)	Arnold, L. Eugene	Multi-Health Systems Inc.	1999	68
Vestibular Stimulation for ADHD: Randomized Controlled Trial of Comprehensive Motion Apparatus	Arnold, L. Eugene	Sage Publications Ltd.	2008	67
The Multidimensional Anxiety Scale for Children (MASC): Confirmatory factor analysis in a pediatric ADHD sample	March, J.S.	Multi-Health Systems Inc.	1999	50
Genetically engineered organisms and the environment: Current status and recommendations	Snow, A. A.	Ecological Society of America	2005-04	39
Reaction Time Distribution Analysis of Neuropsychological Performance in an ADHD Sample	Hervey, Aaron S.	Taylor & Francis Group, LLC	2006	25
The geographies of political ecology: after Edward Said	Wainwright, Joel	Pion	2005	19
An Essay on Continued Fractions	Wyman,	Springer-Verlag :	1985	16