DSpaceStatistics

Contents

Page outdated

DSpace contains functionality for statistics for which updated documentation can be retrieved in the official DSpace documentation on DSpace Statistics

Introduction

At the DSpace User Group Meeting 2005 in Cambridge there was some interest in developing statistics reporting for DSpace to a greater degree.

Initially DSpace came with a basic log file analyser which performed aggregation on the logged actions and produced a basic text report of system activity. Since then there have been a number of developments in statistics for DSpace, each with a different focus or methodology. Following discussions in Cambridge there seemed to be sufficient interest in stats development that we are proposing to work together to provide some sort of more advanced package or module to handle stats and possibly logging in DSpace. We would like to have the design process happen in public, and use this wiki page as the main point of contact for development.

- RichardJones (University of Bergen) developed some code which was posted here as DstatPackage, and which has subsequently been included into the DSpace 1.3 codebase. This code is basically an extension of the original log analyser written in Java with some sightly more advanced archive analysis and a UI component that allows generated reports to be published (publicly or privately), and manages navigation through general and monthly reports. It relies purely on periodic log file analysis.

- LeoMonus (ANU) developed some code which was demonstrated at the Statistics BOF group in Cambridge. Events are stored in database tables, currently capturing view_item and view_bitstream events. IP address logging was added in order to identify and filter harvesters. New reports can be added or existing ones modified by adding SQL queries to an XML configuration file. The default report format is HTML tabular format (although reports can be associated with any XSLT stylesheet) however by using a Cocoon-based extension graphical reports (line charts, piecharts, graphs) and spreadsheets are also available. A preliminary version is available for download.

- NathanSarr (University of Rochester) developed some code which was presented in the main auditorium in Cambridge. The statistics software we developed uses database tables as the storage mechanism for our statistical information. Currently we capture download information at the bitstream level. We have developed two different statistics packages. One that captures download counts per bitstream per month (less sophisticated but stable) and one that captures download counts per bitstream per day per IP address (more sophisticated but in beta).

It would seem sensible to take ideas and code from these and any other systems that people have developed to come up with something to meet needs.

Notes from Cambridge

During the statistics BOF session in Cambridge, a number of issues and thoughts came up regarding the challenges and requirements of DSpace statistics

We noted that there were at least 3 different types of statistics that may be interesting to people:

1. Activity statistics - file downloads, user logins, search requests, etc.
2. Archive statistics - number of items, number of types of items, etc.
3. Administrative statistics - how long items spend in submission/workflow, etc.

The discussion noted the number of different approaches already in use (as given above). The main common need raised was for statistics for file downloads and viewing of items and other usage (so category 1 stats). The possibility of having a table or tables containing the raw activity data was discussed, as well as how this would be populated, what the performance implications would be and how to eliminate erroneous robot accesses. Also raised was the possibility of opening up such data for OAI harvesting, thus allowing the possibility of cross instance analysis of usage.

Also discussed was the possibility of a more structured approach to logging events (into the db) that would allow simpler querying of logged data (the current methods, as listed above, requiring regexp-based logic either in the aggregation or at the querying stages).

Development Discussion

There are a number of points to consider before getting too deep into development. These are listed below, please add any comments/ideas/additional issues.

Log4j

Should we be thinking about replacing log4j, overloading/extending it, leaving it as it is?

In cases where log4j is writing to the same file from different VMs it would be more advised we have a socket based logging service. – MarkDiggory 08:31, 3 November 2006 (EST)

Combining and Improving Existing Statistics Packages

Is there a way we can address all 3 of the above statistics types in one package?

Achieving Modularity

We should make this package totally modular, in part as an experiment in modularisation of DSpace.

Can we use the new Plug-in Manager for configuring the Statistics module

Database Solution

If we replace log4j, and opt for a database and filesystem solution, what are the issues we may encounter?

An alternate solution would be to keep log4j and use http://www.dankomannhaupt.de/projects/index.html solution – MarkDiggory 08:43, 3 November 2006 (EST)

Related Thoughts and Design Issues

StatisticsProposalOne - Some design thoughts by RichardJones

StatisticsAndLog4jIdeas - some ideas about leveraging log4j functionality by LiamLynch

StatisticsFurtherSpeculations - some further ideas by RichardJones about implementation of logging, based on LiamLynch's above ideas.

AdministrativeStatistics - some thoughts about how certain administrative statistics might be handled by RichardJones

ReportGeneration - yet more semi-coherent thoughts about statistical analysis by RichardJones

Interested Parties

Please add yourself here if you are interested in this work, indicate what level of involvement you would like, and perhaps describe a little about your interests/requirements

- RichardJones - Happy to be involved, especially at the architecture stage, and to devote some development time to bringing this up to what we would like to use at Imperial College.
- LiamLynch - Happy to be involved in discussion of architectural aspects, and hopefully the requirements of the Open Repository project for usage stats can be met by a system that is globally applicable, so that we can get involved with moving the development of that forward generally.
- GaryPhillips - Interested in getting more useful usage statistics (view counts) for items and bitstreams by filtering out robots and crawlers, and displaying counts inside DSpace. Mostly lurking until I get more familiar with some possible solutions.
- [StuartLewis] - Would like to be involved, and is happy to contribute effort.
- Mark H. Wood- Interested in view counts and robot identification. Using the University of Rochester code now. Looking into alternate robot filtering methods.
- [PaulNeedham] - Would like to be involved, happy to contribute effort. Worked on the JISC/PAL3 PIRUS Project looking into producing COUNTER-compliant stats for Journal Articles. Developing own stats suite (very rough and ready in PHP) which allow for per author and per collection stats.
- [gfarrasb] - Interested in applying educational data mining algorithms to obtain information about user actions in DSpace.

Please add your thoughts, links to any stats work you may have done, design questions/solutions, requirements and so forth to this page.