DCAT Meeting August 2016

- Date & Time
- Dial-in
- Agenda
- Preparing for the call
- Meeting notes
  - History of DSpace statistics
  - Future of DSpace statistics
  - Performance
  - Housekeeping announcement
- Call Attendees

Date & Time

- August 9th **15:00** UTC/GMT - **11:00** ET

This call is a Community Forum call: Sharing best practices and challenges in the use of existing DSpace features

Dial-in

We will use the international conference call dial-in. Please follow directions below.

- U.S.A/Canada toll free: 866-740-1260, participant code: 2257295
- International toll free: http://www.readytalk.com/intl
  - Use the above link and input 2257295 and the country you are calling from to get your country's toll-free dial in #
  - Once on the call, enter participant code 2257295

Agenda

Community Forum Call: DSpace Statistics

Sharing best practices, challenges, and questions

- DSpace statistics
  - interpreting statistics
  - improving robot filtering & assessing robot traffic
  - exchanging which types of reports are being used for which purposes

Preparing for the call

Bring your questions/comments you would like to discuss to the call, or add them to the comments of this meeting page.

If you can join the call, or are willing to comment on the topics submitted via the meeting page, please add your name, institution, and repository URL to the Call Attendees section below.

Meeting notes

History of DSpace statistics

The first DSpace statistics, currently often referred to as the DSpace legacy stats, were based on DSpace logs. As this system does not take into account any traffic originating from bots, let alone they would filter out such traffic, it is highly discouraged to use these statistics. The lack of robot filtering would bias the results and make them uninterpretable.

The current DSpace usage statistics, introduced in DSpace version 1.6, is based on SOLR.

After the release further improvements and alternatives to the standard DSpace statistics have been developed on the initiative of several universities, institutions, and third party service providers.

An alternative to the DSpace statistics is google Analytics. Although this is an interesting tool to use in some use-cases, it does have some limitations. First of all analytics is a black box. You have to assume its robot filtering is working properly as it is unknown what filtering is used. Secondly, google analytics doesn't know DSpace's internal structure. It isn't familiar with the hierarchy of repository, communities, collections and items. This causes Analytics to be unable to create statistics on an aggregated level (e.g. the total item page views of all items in a collection).

Another alternative is the third party add-on Piwik. The DCAT was under the impression this system might provide skewed statistics. As piwik uses a client side javascript to collect statistics, only downloads made by clicking the DSpace download link are likely to be counted. Chances are high that downloads originating from outside DSpace, for example directly from google, are not logged.

Future of DSpace statistics

In the new User Interface it would be beneficial to enable SOLR to be queried directly through the centralized API instead of SOLR's REST API. This would allow to replace SOLR with another system, should a better data source arise. In the meantime, people developing to the DSpace statistics layer could more easily contribute their work to the community, as this would also be built upon this central DSpace API.

Performance

Some institutions noticed performance issues caused by the overhead created by SOLR. Harvard university has solved this issue by relying on web server logs. These logs are already made and therefor do not add additional load on DSpace. An other solution by a third party service provider was to use elasticsearch instead of SOLR, which appeared performant.

There are some opportunities to reduce the overhead load created by SOLR. It is for example not required to run SOLR on the same server as DSpace. It is possible to create a separate SOLR server. Another way of reducing the load is by creating a sharded SOLR core (for example per year). One third party service uses a SOLR caching mechanism to balance the load SOLR puts on DSpace's performance, this way there should not be a noticeable difference.

Housekeeping announcement

Up to now the name of DCAT itself, the 'DSpace Community Advisory Team', sounds rather formal. This may scare people off to join the conversations. For that reason there will be meetings called 'Community Forum calls'. We hope this name indicates the call is open to the entire community.

Discussion topics for the next DCAT calls are already listed on the DCAT meeting notes page. Next month's topic of interest will be the DSpace standard Data model and DSpace-CRIS.

Call Attendees

- Maureen Walsh - The Ohio State University
- Bram Luyten (Atmire)
- Ignace Deroost - Atmire
- David Corbly - University of Oklahoma
- Mariya Maistrovskaya - University of Toronto
- Terrence W Brady - Georgetown University
- Valerie Collins - University of Minnesota
- Marianne Reed - University of Kansas
- Monica Rivero - Rice University
- Iryna Kuchma - EIFL
- Peter Dietz - Longsight
- Elias Tzoc - Miami University
- Daniel Draper -Colorado State University
- Pauline Ward - University of Edinburgh
- Filipe Furtado - University of Minho
- Susan Borda - Montana State University
- Felicity Dykas - University of Missouri–Columbia
- Joseph Greene - University College Dublin
- Irene Berry - Naval Postgraduate School