

DCAT Meeting April 2017

Date & Time

- April 11th **15:00** UTC/GMT - **11:00** ET

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 Performance

Open discussion on DSpace performance challenges, exchanging best practices for analysing and resolving performance problems. How to involve users & repository managers in adequately reporting performance issues.

Preparing for the call

In preparation of the call, you could do the following:

- List any performance problems you may have with DSpace. Make it clear which version you are using
- List any specific performance improvements or hacks you have made
- List any monitoring tools/diagnostics you have experience with

Meeting notes

DSpace 5 vs. DSpace 6 comparison

In the DSpace 6.0 release the performance enhancing efforts were not entirely successful. However, in the release of DSpace 6.1 these should be fixed, making DSpace 6 in general terms more performant than DSpace 5.

To test this statement it would be good if we could set up two identical server environments on which we deploy respectively a DSpace 5 and a DSpace 6. If these repositories are then populated with the exact same content we can make an objective comparison of the performance of DSpace 5 and 6.

Multiple collections issue

In DSpace 6.0 JSPUI, when a repository has many communities and collections this can cause a performance issue. In such repository, during the collection selection step in the item submission process, the collection list takes a long time to load. This issue is currently under investigation.

During the call there were some other issues reported which are related to the above. For example, for repositories with many communities and collections performance appeared to be decreasing when upgrading to newer DSpace versions for one participant. This attendee also notices performance issues in indexing repositories with many items.

The fact that these issues were not detected during the testing phase of DSpace 6.0 reflects a more general issue with DSpace performance testing. This testing is currently done on the DuraSpace Demo repository (demo.dspace.org). This repository however is usually populated with only limited amounts of communities, collections, and items. At this point we are not testing DSpace's performance on large repositories. It would be good if we could set up such testing environment for future releases.

Monitoring infrastructure for early signs of performance issues

One popular proprietary tool for server monitoring is New Relic. It can detect significant changes in the use of resources and send alerts when this happens. It also lets you know at which time an issue occurs. New Relic is also capable of pinpointing lines of code which may have caused the performance issue.

A low tech way of doing basic test of your repository's performance is by using your in-browser developer tools, which are included in many modern browsers. In most cases you can access these tools by right-clicking in your browser, and selecting an option such as 'inspect' or 'developer tools' which should pop-up a pane at the bottom of your browser screen. This pane will likely have a network tab, in which you can monitor the loading times of pages in DSpace while you are testing features. This will provide you with hard numbers you can use to compare your performance over time.

Configuration

There are several configurations which may impact your repository's performance.

Apache Tomcat

One Tomcat configuration setting you can use to increase performance is the crawler session manager, which can restrict the number of sessions for a crawler user agent. If bot traffic generates performance issues limiting the maximum amount of sessions for those bots may help.

Database

The standard PostgreSQL settings are not ideal for repositories with much traffic. For these repositories it is better to increase the maximum database connections.

During the call it was also not certain why the default PostgreSQL settings allow for an unlimited number of idle connections.

Apache Solr

Solr is memory intensive, and runs alongside DSpace in the tomcat application server. This means it will have to share its available memory with DSpace.

As solr is recording all the DSpace usage events (item page views, bitstream downloads, search queries), the memory usage of solr is related to the usage of the repository. Repositories with much usage may also require more memory for their solr.

One way of limiting the memory usage of solr is not writing any robot traffic to the solr core.

Load testing

One tool which can be used for load testing is loadimpact.com, the free tier should already suffice for most repositories. It is advised to be cautious when using this tool, as increasing the load on your DSpace may eventually lead to a failure.

Another tool used by a call attendee is Apache JMeter (<http://jmeter.apache.org/>). This tool is free and has the capability of capturing browser settings.

How to contribute solutions back to the community

Codebase-fixes can be contributed just like any other code-fix. However, there seems to be a need to centralize more information regarding environment-specific optimizations:

- Tomcat config
- Postgres config
- SOLR Config (mixed, because solr config does live within the codebase to some extent)
- Apache HTTPD config (caching?)
- Operating system config (Linux vs Windows)
- ...

Call Attendees

- [Bram Luyten](#) (Atmire)
- [Maureen Walsh](#) (Ohio State University)
- [Ignace Deroost](#) (Atmire)
- Andrew McLean (Imperial College London)
- Agustina Martinez (University of Cambridge)
- Nicholas Webb (Mount Sinai Health System)
- [Iryna Kuchma](#) (EIFL)
- Felicity Dykas (University of Missouri)
- Emilio Lorenzo (Arvo Consulting)
- [Pauline Ward](#) (University of Edinburgh)
- Valerie Collins (University of Minnesota)
- [Pascal-Nicolas Becker](#) (The Library Code / TU Berlin)
- [Terrence W Brady](#) (Georgetown)
- [Suzanne Chase](#) (Georgetown)
- [Michael Marttila](#) (Georgetown)