DevMtg 2019-07-10

Developers Meeting on Weds, July 10, 2019

Today's Meeting Times

- DSpace Developers Meeting / Backlog Hour: 15:00 UTC in #duraspace IRC or #dev-mtg Slack channel (these two channels sync all conversations)

No meeting on Weds, July 3. Tim is out of the office from July 3-7, returning on Monday, July 8.

Agenda

Quick Reminders

Friendly reminders of upcoming meetings, discussions etc

- DSpace 7 Working Group (2016-2023): Next meeting is Thurs, July 11 at 14:00 UTC.
- DSpace 7 Entities Working Group (2018-19): Next meeting is Tues, July 16 at 15:00 UTC.
- DSpace Developer Show and Tell Meetings: On hold until interesting topics arise.
- 2019 DSpace North American User Group Meeting (Sept 23-24) Call for proposals out!

Discussion Topics

If you have a topic you'd like to have added to the agenda, please just add it.

1. Quick Updates from other meetings
    a. DSpace 7 Status Updates for this week (from DSpace 7 Working Group (2016-2023) or DSpace 7 Entities Working Group (2018-19))
    b. DSpace 6.x Status Updates for this week
        i. 6.4 will surely happen at some point, but no definitive plan or schedule at this time. Please continue to help move forward / merge PRs into the dspace-6.x branch, and we can continue to monitor when a 6.4 release makes sense.
2. Ongoing Work
    a. Upgrading Solr Server for DSpace (
        i. **n Solr**: Unable to locate Jira server for this macro. It may be due to Application Link configuration.
        s only happen for **major** releases? Should it be configurable? Can we find a more precise trigger? When **do** we need to reindex?
        ii. ty core. Unable to locate Jira server for this macro. It may be due to Application Link configuration. Or shou
    b. DSpace l ls (old) (Terrence W Brady)
        i. Update sequences on initialization
            1. https://github.com/DSpace/DSpace/pull/2362 - update sequences port
            2. https://github.com/DSpace/DSpace/pull/2361 - update sequences port
        ii. DSpace Launcher Dashboard - Deploy a PR on AWS for Testing
            1. There is a 2 minute video that illustrates this proposal
3. For Discu
    a. Unable to locate Jira server for this macro. It may be due to Application Link configuration.
    b. ext" object had a separate DB connection. In DSpace 6+, each **thread** has a separate DB Connection (but Context objects may share a DB connection if they share the same thread).
4. Tickets, Pull Requests or Email threads/discussions requiring more attention? (*Please feel free to add any you wish to discuss under this topic*)
    a. Quick Win PRs: https://github.com/DSpace/DSpace/pulls?q=is%3Aopen+review%3Aapproved+label%3A%22quick+win%22

Tabled Topics

These topics are ones we've touched on in the past and likely need to revisit (with other interested parties). If a topic below is of interest to you, say something and we'll promote it to an agenda topic!

1. Brainstorms / ideas
    a. (*On Hold, pending Steering/Leadership approval*) Follow-up on "DSpace Top GitHub Contributors" site (Tim Donohue): https://tdonohue.github.io/top-contributors/
    b. Bulk Operations Support Enhancements (from Mark H. Wood)
    c. Curation System Needs (from Terrence W Brady)
2. Management of database connections for DSpace going forward (7.0 and beyond). What behavior is ideal? Also see notes at DSpace Database Access

a. In DSpace 5, each "Context" established a new DB connection. Context then committed or aborted the connection after it was done (based on results of that request). Context could also be shared between methods if a single transaction needed to perform actions across multiple methods.
b. In DSpace 6, Hibernate manages the DB connection pool. Each **thread** grabs a Connection from the pool. This means two Context objects could use the same Connection (if they are in the same thread). In other words, code can no longer assume each `new Context()` is treated as a new database transaction.
    i. Should we be making use of `SessionFactory.openSession()` for READ-ONLY Contexts (or any change of Context state) to ensure we are creating a new Connection (and not simply modifying the state of an existing one)? Currently we always use `SessionFactory.getCurrentSession()` in HibernateDBConnection, which doesn't guarantee a new connection: https://github.com/DSpace/DSpace/blob/dspace-6_x/dspace-api/src/main/java/org/dspace/core/HibernateDBConnection.java
c. Bulk operations, such as loading batches of items or doing mass updates, have another issue: transaction size and lifetime. Operating on 1 000 000 items in a single transaction can cause enormous cache bloat, or even exhaust the heap.
    i. Bulk loading should be broken down by committing a modestly-sized batch and opening a new transaction at frequent intervals. (A consequence of this design is that the operation must leave enough information to restart it without re-adding work already committed, should the operation fail or be prematurely terminated by the user. The SAF importer is a good example.)
    ii. Mass updates need two different transaction lifetimes: a query which generates the list of objects on which to operate, which lasts throughout the update; and the update queries, which should be committed frequently as above. This requires *two* transactions, so that the updates can be committed without ending the long-running query that tells us what to update.

Ticket Summaries

1. Help us test / code review! These are tickets needing code review/testing and flagged for a future release (ordered by release & priority)

key	summary	type	created	updated	assignee	reporter	priority	status	fixversions

> Unable to locate Jira server for this macro. It may be due to Application Link configuration.

2. Newly created tickets this week:

key	summary	type	created	assignee	reporter	priority	status

> Unable to locate Jira server for this macro. It may be due to Application Link configuration.

3. Old, unresolved tickets with activity this week:

key	summary	type	created	updated	assignee	reporter	priority	status

> Unable to locate Jira server for this macro. It may be due to Application Link configuration.

4. Tickets resolved this week:

key	summary	type	created	assignee	reporter	priority	status	resolution

> Unable to locate Jira server for this macro. It may be due to Application Link configuration.

5. Tickets requiring review. This is the JIRA Backlog of "Received" tickets:

key	summary	type	created	updated	assignee	reporter	priority

Unable to locate Jira server for this macro. It may be due to Application Link configuration.

Meeting Notes

Meeting Transcript

Log from #dev-mtg Slack (All times are EDT)

```
Mark Wood [11:01]
Hello and welcome to the weekly DSpace developers' meeting. Tim is unavailable today, so I'm filling in as
moderator. Who is here today?

James Creel [11:02]
I'm here!

Terry Brady [11:03]
hello

Mark Wood [11:03]
OK, we can start, and if others join in, so much the better.
There was no DSpace 7 WG meeting last week, so nothing to report.
The DSpace 7 Entities WG had ongoing discussion of UI mockups and how metadata namespaces should be used, and
brief discussion of clarified labels for relationship fields.
I'm not aware of any recent movement on DSpace 6.4.
Is there anything to add to any of these?
Moving along then: ongoing work.
I must once again report that I've had little time to devote to the Solr upgrade. I'm still working to assemble
a realistic set of test data for trying out the DSpace upgrade process.
Is anything happening in the DSpace Docker and Cloud Deployment area?

Terry Brady [11:11]
I will meet with Tim sometime in the next couple weeks to look over the DSpace launcher dashboard to see if the
project would be interested in adopting it.
Otherwise, no DSpace work on my end.

Mark Wood [11:11]
Thank you.

Terry Brady [11:11]
http://bit.ly/dspace-launcher-dashboard
terrywbrady.github.io
DSpace Launcher Dashboard
Deploy a DSpace PR in Docker on AWS for Testing

Mark Wood [11:13]
Next on the agenda is "Brainstorming how to improve DSpace database usage (now that we use Hibernate)." Is this
something we can discuss today?

Terry Brady [11:13]
sounds interesting to me

Mark Wood [11:13]
I should have given a link to the agenda:
https://wiki.duraspace.org/display/DSPACE/DevMtg+2019-07-10

James Creel [11:14]
```

Bugs have cropped up here and there for us around the Hibernate session. It's hard to trace them with the XMLUI stacktraces.

Mark Wood [11:17]
My impressions: Historically, DSpace has carried the database connection around inside the Context object, and treated it like an expensive resource. Hibernate (and JPA) seem to treat the Session (which manages DB connections) like a cheap resource, to be created and destroyed freely and kept close to the code that uses it. DSpace keeps a single Session open for an entire operation (like a web request, or a command-line command) and uses it for all database activity. This may not be a good fit to the way that Hibernate manages DB connections.

James Creel [11:19]
The db connection pool tends to fill up and crash the service. We've had to up the size.

Mark Wood [11:20]
Interesting. Hibernate should be managing those under the covers, but DSpace may be hampering its efficient use of connections.
Is that the general case, or have you noticed specific activities that fill up the pool?

James Creel [11:22]
This is unscientific. It could be that our traffic patterns have just changed. But this became an issue a couple years ago. We have not analyzed the activities except to say that it occurs under load.

Terry Brady [11:25]
It would be really useful to curate some reference datasets (SQL + assets) for verifying performance of specific actions (search, re-index, etc)

Mark Wood [11:26]
I think that one area in which we are not using Hibernate well is in bulk operations. Using the same Session to identify perhaps thousands of objects for modification, and for making those modifications, creates long-lived transactions that build up a lot of uncommitted work in memory. We've added work-arounds to flush the cache at intervals, but they've proven to be tricky to use correctly. I think that we need to have small, short-lived writable Sessions separated from a long-lived read-only Session that drives the work. Thoughts? (edited)
@terrywbrady yes, that should be helpful.

James Creel [11:27]
The batch SAF import, the media filter, and the batch metadata update are all less functional now, as they don't commit changes incrementally, but wait until the very end. We've needed to divide jobs into batches to get them finished.

Terry Brady [11:28]
The changes you are describing would be really useful to add to a major release (vs a point release). It would be great if it could go out with the 7.0 release.

Mark Wood [11:29]
Oooh, I don't know if there is time to develop and test something so fundamental, for 7.0. It's an attractive target, though.
I wouldn't rule it out.
Oh, one thing I've seen is that DSpace has always assumed that each Context represents a distinct DB connection. Hibernate doesn't work that way: you get one default Session per *thread*, so a thread which creates multiple Contexts will be using the same Session in all of them. This leads me to think that we may want to remove the embedded Session from Context.
I'm not sure how any of this is related to pool exhaustion. We need to work out how to understand what Hibernate is doing with Connections and how DSpace's use of Hibernate affects that.

James Creel [11:35]
Yeah, I feel like I'm on the hook to provide a reproducible procedure before we can act on that. It's been all about just getting it working when it crashes.

Mark Wood [11:36]
I know that feeling.

James Creel [11:37]
What I'll do is make a card for that on our internal sprint board. That way we won't forget it's an opportunity when time becomes available.

Mark Wood [11:37]
It may be hard to reproduce. Another approach would be to work out how to log the right stuff, without having it be invisible in a flood of the wrong stuff.
Thank you.

James Creel [11:37]

No DSpace sprints on the calendar at present though.

Mark Wood [11:38]
It's good just to realize that we may have another problem somehow related to these issues.
So, do we have a way forward on any of these issues?
We need more data about the pool exhaustion, and may get them as opportunities arise.
We need to look at all of the bulk operations, to see if we can improve their use of transactions.
Anything else?
Any thoughts on whether it would be worth the effort, to rip the Session out of Context and move Session creation down into the code that needs Sessions?
Or thoughts on better ways to use and manage Sessions?

James Creel [11:43]
Is that a more established practice with ORMs? Would DSpace have been done that way if using an ORM from the start?

Mark Wood [11:43]
I suspect that it would have been, yes.

James Creel [11:44]
Perhaps it could be tried out for a small, isolated feature set before a larger effort was made.

Mark Wood [11:44]
Good idea. Let's see, where's a handy victim....:slightly_smiling_face:
Maybe we should just think about that a bit, and come back to it when an idea surfaces. I want to leave a few minutes to discuss Issues and PRs.
Final thoughts on Connections, Sessions and Contexts?
OK, are there PRs or Jira Issues that need attention?

Terry Brady [11:52]
There are still my 2 migrate-sequences PR's that would be a very quick thing for someone to test.
https://github.com/DSpace/DSpace/pull/2362
https://github.com/DSpace/DSpace/pull/2361

Mark Wood [11:53]
Ah, those are ports from master. I'll try them out. I think I already tested the original, maybe I can remember how I did it.

Terry Brady [11:54]
Thanks. Give me a shout if you need a hand with the testing.

Mark Wood [11:54]
Will do.
We have 28 open Quick Win PRs, many with one approval already. https://github.com/DSpace/DSpace/pulls?q=is%3Aopen+review%3Aapproved+label%3A%22quick+win%22
Shameless plug: I'm still looking for another review of #1992 (DS-3872, Velocity for email templates).
Other Issues or PRs to discuss?
Any other topic that needs a quick mention?
It's grown quiet. Shall we close up the meeting?

James Creel [12:00]
I've got to run to another. Bye!

Terry Brady [12:00]
Thanks for running the meeting. Have a good week.

Mark Wood [12:00]
Bye, then, and thank you all for coming. Meeting adjourned.