DevMtg 2018-11-28

Developers Meeting on Weds, November 28, 2018

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)
    - Please note that all meetings are publicly logged

Agenda

Quick Reminders

Friendly reminders of upcoming meetings, discussions etc

- DSpace 7 Working Group (2016-2023): Next meeting on Thursday, Nov 29 at 15:00 UTC.
- DSpace 7 Entities Working Group (2018-19): Next meeting on Tues, Dec 4 at 16:00 UTC
- DSpace Developer Show and Tell Meetings: On hold until interesting topics arise.

Discussion Topics

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

1. (*Ongoing Topic*) DSpace 7 Status Updates for this week (from DSpace 7 Working Group (2016-2023))
2. *(Ongoing Topic)* DSpace 6.x Status Updates for this week
    a. 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.
3. Solr upgrade discussions. See Upgrading Solr Server for DSpace
    a. PR https://github.com/DSpace/DSpace/pull/2058
4. Brainstorms / ideas (*Any quick updates to report?*)
    a. (*On Hold, pending Steering/Leadership approval*) Follow-up on "DSpace Top GitHub Contributors" site (Tim Donohue): https://tdonohue.github.io/top-contributors/
    b. Follow-up on Curation Task Reporting (PR 2180)
    c. Bulk Operations Support Enhancements (from Mark H. Wood)
    d. Curation System Needs (from Mark H. Wood)
        i. PR 2181 implements per-run task parameters. Ready for review.
        ii. PR 2180 improves reporting. Ready for review.
5. 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. 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

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 CST)

```
Tim Donohue [9:00 AM]
@here: It's time for our weekly DSpace DevMtg.  Agenda for today is at https://wiki.duraspace.org/display/DSPACE
```

/DevMtg+2018-11-28
Let's do a quick roll call to see who is able to join the discussion today

Mark Wood [9:00 AM]
Hi

Terry Brady [9:00 AM]
hello

Pascal Becker [9:00 AM]
hi
will be lurking only and be distracted from time to time

Alexander Sulfrian [9:01 AM]
Hi

Tim Donohue [9:02 AM]
Hi all, welcome. Today's agenda is likely a bit light (it's mostly quick updates), so we should have time for other topics or discussions
Jumping right in though... last week was a slow week overall for DSpace meetings/efforts, cause of the USA Thanksgiving holiday. So, most meetings, including this one didn't happen.
But, we have a few meetings coming up to be aware of.... our DSpace 7 Meetings resume tomorrow (Nov 29) at 15UTC. And the next Entities Working Group meeting is next Tues, Dec 4 at 16UTC
On the DSpace 7 front, I don't have any specific updates to share with this team...so, I'd recommend joining our next DSpace 7 meeting tomorrow (if interested), and/or I'll share any updates next week.
As noted last week, the DSpace 7 rough/estimated schedule is now posted up on 7.0 Status page: https://wiki.duraspace.org/display/DSPACE/DSpace+Release+7.0+Status
Same goes for DSpace 6.x. No updates to share today.
Before we move into actual topics, any other quick updates to share? (or questions?)

Terry Brady [9:06 AM]
If time permits, I'll have an item to add at the end

Tim Donohue [9:07 AM]
Ok, moving along. We talked about this a bit last week... but, Upgrading Solr (in DSpace 7) is a big priority, and one that this general meeting can keep moving along. @mwood has kindly started up a wiki page on this: https://wiki.duraspace.org/display/DSPACE/Upgrading+Solr+Server+for+DSpace

Pascal Becker [9:07 AM]
There is a press release about the DSpace Konsrotium Deutschland. See #general about that.

Terry Brady [9:08 AM]
@mwood, let me know if it would be useful for you and I to meet over the next week to chat about Solr next steps.

Tim Donohue [9:08 AM]
Any updates or discussion we'd like to have around the plans to Upgrade Solr (Client & Server)? I'd like to ensure we can continue to move this along, as ideally we'd have this ready for the "preview" release of DSpace 7 in late January / early Feb
I'll gladly support this Solr Upgrade effort as well (where ever I can), but I'm assuming @mwood and @terrywbrady are taking more of a lead.

Mark Wood [9:10 AM]
How deeply should DSpace code be involved with installing and managing cores in a Solr instance that it doesn't control? Should we start out by just telling folks where to find the cores and suggesting how to find where they should be copied?

Tim Donohue [9:12 AM]
@mwood: I guess I'd turn that question around and ask what the "best practices" (if any) seem to be from a Solr standpoint? Is it recommended for "cores" (and their data/configuration directories/files) to sit under the main Solr directories? Or is there a way to tell Solr...this core is over _here_ under [dspace]/solr/ (like you can with Tomcat webapps)?

Terry Brady [9:13 AM]
For everything but the statistics, I suppose we would want to recreate those in the new location. I imagine the stats would need a migration tool.

Mark Wood [9:13 AM]
I have the impression that the preferred way is to place them in Solr's home directory tree and let it discover them.

Tim Donohue [9:13 AM]
Ideally, we let folks manage Solr more directly (don't do as much for them). But, we will still have to provide Solr Schemas & core(s) configuration, and I'm not sure how best to "give that" to Solr. (edited)
@mwood: Ok, then that implies we could either give detailed instructions for how to create these cores, or provide scripts/commands that do so (during `ant fresh_install` for example)

Mark Wood [9:15 AM]
I will experiment a bit with just copying cores, both new and used.

Tim Donohue [9:15 AM]
That'd be good
I think we'll want to have a balance here of "hand holding". Ideally, we want DSpace to be "out of the box" (that's always the goal), so if we can ease the creation/copying/configuration of cores that's *ideal*. But, at the same time, we should stop doing stuff that Solr does better...so, for example, maybe upgrading cores needs to be more manual (as it's part of your Solr upgrade process)

Mark Wood [9:16 AM]
We can distinguish new installs and existing sites. A new, empty core seems to be just a directory structure with a few configuration files in it. We should be able to just tell people "copy this to there".
Copying a core with records in it, from one version of Solr to another, may require more.
Yes, another distinction that we need to keep in mind is "Solr changed" vs. "DSpace changed the way it uses Solr."

Tim Donohue [9:18 AM]
@mwood: Right, I was trying to first figure out a clean "fresh install". But, yes, an upgrade from DSpace 6 -> 7 may require more... If you recall though, we have Solr Core "upgrade process" already built into `ant update` which could be of use here perhaps

Pascal Becker [9:18 AM]
Does something like Flyway for Solr exists?

Tim Donohue [9:19 AM]
@pbecker: not aware of any...but, we have a custom Ant process to upgrade Solr cores already
Here's our custom Solr core update process.. `ant update_solr_indexes`: https://github.com/DSpace/DSpace/blob/master/dspace/src/main/config/build.xml#L951

Terry Brady [9:19 AM]
I see the following use cases
- One DSpace 6 stats shard
- Multiple DSpace 6 stats shards (uuid migration complete)
- One DSpace 5 stats shard (unmigrated)
- Multiple DSpace 5 stats shards (unmigrated)
- No existing cores (new install) (edited)

Tim Donohue [9:20 AM]
All that Ant knows how to do though is simply run the Solr upgrade core scripts... it's not checking schemas or anything...just ensuring the core itself is compatible with later versions of Solr.
But, that still could be useful for a basic "move & upgrade" of existing Solr cores.

Pascal Becker [9:21 AM]
@tdonohue doesn't ant needs access to those indexes? Do we have it if solr runs standalone? We have to think about institutions that run solr like databases on a separate server than DSpace.
I lately had a tender where I was asked specifically if it is possible to use the already existing standalone installation of solr instead of DSpace's embedded one.

Tim Donohue [9:22 AM]
I'm *only* suggesting using Ant for a DSpace 6 to 7 upgrade (to update the Solr indexes to a later version so they can be moved elsewhere)

Pascal Becker [9:22 AM]
I see.

Tim Donohue [9:23 AM]
I *don't* think we should continue to use Ant after that... Solr has it's own upgrade processes once you are on an external Solr

Pascal Becker [9:23 AM]
perfect.

Mark Wood [9:23 AM]
Assisting with the transition to a separate Solr makes sense. After that, we should depend on Solr's own tools

and procedures for "Solr changed" and provide support ourselves only for "DSpace changed."

Alexander Sulfrian [9:24 AM]
There is an API to manage the solr schema. I think the ant stuff could be rewritten to use the API to check (at least) the schema version.
https://lucene.apache.org/solr/guide/6_6/schema-api.html

Tim Donohue [9:25 AM]
@sulfrian: yes, we had noticed that Solr has a new Schema API last week: https://lucene.apache.org/solr/guide/7_0/solr-configuration-files.html#configuration-files

Do you have any experience with it? We were not sure how much work this would be to move all our schema.xml files to the API
this was the link to the actual API: https://lucene.apache.org/solr/guide/7_0/schema-api.html#schema-api

Alexander Sulfrian [9:27 AM]
No, I do not have experience with it, but it should be easy to check if the schema was installed correctly.

Mark Wood [9:28 AM]
We may just want to remember that for the future, when schema changes may be required.

Tim Donohue [9:28 AM]
Oh, good point. We could possibly use the Schema API to verify things "look good". Maybe eventually we could find a way to use the Schema API to also do schema updates (i.e. automate those), but I'm not sure if the latter is a massive amount of effort

Alexander Sulfrian [9:29 AM]
It should be one REST request for each field and it should be possible to create a tool to incrementally update the schema (something comparable to flyway).

Mark Wood [9:29 AM]
Consider also that, in a large shop, the people who run Solr may not allow the people who run DSpace to make changes themselves.

Tim Donohue [9:31 AM]
@mwood: but DSpace is going to need to be able to manage the Solr Core in some way anyhow. I'm not sure why it shouldn't be allowed to manage the schema of the cores that it "owns". But, it's probably something we can consider post-7.0 anyhow
In any case, trying to bring this back to useful next steps... It sounds like we need more testing/analysis of how to copy over existing cores to an external Solr
We also need to investigate what a new "fresh_install" would look like (how do cores get configured in the first place)

Mark Wood [9:33 AM]
Should we look at this as a two-part effort? (1) Move to a separately managed Solr in the simplest way; (2) look into new opportunities for making core management easier.
Part 1 for DSpace 7; part 2 for "later".

Terry Brady [9:34 AM]
Mark, if I want to experiment with Solr as a standalone service (vs a DSpace webapp), do you have a PR or some other setup suggestions?

Tim Donohue [9:34 AM]
@mwood: I'm not sure all of part 2 should be "later".... it depends on whether part 1 makes the DSpace 7.0 upgrade "extra complex" or not

Alexander Sulfrian [9:35 AM]
replied to a thread:
We want to migrate the authority core, too. It cannot be recreated easily, because it contains data (f.e. orcid) that is not available at any other place.

Mark Wood [9:35 AM]
I have a patch that rips out dspace-solr. I should make certain that it is represented by a work-in-progress PR for ease of experimentation. I haven't tried setting up Solr separately with DSpace cores.

Tim Donohue [9:35 AM]
@mwood: So, I guess I agree with part 1, assuming we split it into "fresh install" & "upgrade" (from DSpace 6 to 7). But, then we'll need to see if parts of that can be eased, or not.

DSpaceSlackBot (IRC) APP [9:36 AM]
qwebirc89139 has quit the IRC channel

Terry Brady [9:37 AM]
When you have that PR running, I will give it a test and then try to create a Docker representation of the separate services to make it easier to test.

Mark Wood [9:37 AM]
@sulfrian Then we probably should have a dump/restore tool for the authority core for disaster recovery, and it could be used for migration as well.

Tim Donohue [9:37 AM]
I suspect we need a running list of all the "cores" here...and various needs in the upgrade. I'm worried these discussions aren't always making it back to a "TODO" list

Mark Wood [9:38 AM]
I will try to capture today's discussion in the wiki page.
At least, the core list and special considerations.

Pascal Becker [9:38 AM]
@mwood @sulfrian does such a tool exist? I'm not sure if we already have something to dump the autority core. I wish we do.

Tim Donohue [9:38 AM]
(And I'm not sure all cores need *reindexing* to upgrade them. That's only really needed when the schema changes...if the schema is the same, we should be able to just run the Solr index upgrade script without a full reindex)

Pascal Becker [9:39 AM]
Sorry, I have to run.

Mark Wood [9:39 AM]
I do not recall one, but if the data are not reproducible from other sources then we definitely should make one.

Pascal Becker [9:39 AM]
Good bye everyone!

Terry Brady [9:39 AM]
bye @pbecker

Tim Donohue [9:39 AM]
We have a script to dump & reindex Solr Statistics index. It likely could be extended to be used for Authority too.
Back to the TODO's here.. it seems like there are a lot of questions still. While it's useful to capture all the questions in a TODO list, I think we need to start small... we likely need to finish up the Solr Client upgrade, and get DSpace 7 working with a *new, fresh* external Solr Server
From there, we can more easily test & figure out an upgrade & fresh_install process

Mark Wood [9:41 AM]
That sounds like a good way to proceed.

Tim Donohue [9:41 AM]
Currently, I worry we are speculating a bit too much, without any code to test against / play with

Mark Wood [9:41 AM]
Yes.

Tim Donohue [9:42 AM]
@mwood & @terrywbrady: Is this something you two could start with? Find time to finish the Client upgrade (existing PR). Then, get it working with an external Solr (perhaps Docker-ify it too)?
I'll gladly offer support, just want to be sure one (or both) of you can take the lead here

Mark Wood [9:44 AM]
Yes, I should get back to the client upgrade. It would help if I had a better understanding of how embedded Solr is started and managed in the tests. Is there a place I should start reading?

Terry Brady [9:44 AM]
We also need a bare minimum to disregard the solr webapp... perhaps that just requires deleting webapps/solr from the install directory.

Tim Donohue [9:45 AM]
@terrywbrady: Yes, I agree. I think that `webapps/solr` should be deleted.
@mwood: I think Spring REST is just using the `EmbeddedSolrServer` that comes with SolrJ for all tests. Here's

the `MockSolrServer`: https://github.com/DSpace/DSpace/blob/master/dspace-spring-rest/src/test/java/org/dspace/solr/MockSolrServer.java
That MockSolrServer looks to do setup/destroy
(That's literally all I know though...if you get stuck, maybe we can ask @tom_desair, who looks to have created this)

Mark Wood [9:47 AM]
Ah. Not really a mock, if I recall correctly. I will look more deeply into how it is used. Really we should have the test framework set up/purge/tear down Solr as required.

Terry Brady [9:47 AM]
Perhaps I can start to assemble some tests in a Docker setup that (1)drops webapps/solr from the tomcat dir and (2)starts an external solr

Tim Donohue [9:48 AM]
@mwood: I'd welcome you to refactor it however you see fit (and document/comment the heck out of it). I admit, I don't know how this works either... and I haven't had time to dig in to figure it all out

Mark Wood [9:48 AM]
That should be fairly simple. We could simply ignore webapps/solr at first, so long as it is not started.
I will try to spend some time figuring out that test failure. I think that is all that stands in the way of final review.

Tim Donohue [9:50 AM]
With regards to the Solr webapp, I think we can even turn it off in the Maven build process... something like `mvn -P!dspace-solr package` should disable that module from even building/compiling

Mark Wood [9:50 AM]
Do we know for a fact that SolrJ 7 is incompatible with Solr server 4.10?

Tim Donohue [9:51 AM]
@mwood: I don't know

Mark Wood [9:52 AM]
We want to move to a contemporary server version, yes, but during development it might be handy to stick with the included server until we know how we are going to abandon it.
I can try that, if it seems useful.

Tim Donohue [9:53 AM]
@mwood: perhaps, but I'm not confident a SolrJ v7 will work with a Solr v4. If it does (for development), then that's fine. But, I think the sooner we get to an external Solr the *better* (as that's really the end goal).

Mark Wood [9:54 AM]
I won't spend time on that mix unless it looks like *saving* time.

Tim Donohue [9:54 AM]
So, I only see that as a very very temporary stopgap...one that isn't really worth much time/effort looking at, since we'll abandon it soon anyhow :wink:
Ok, so it sounds like there's a general plan here. I'm seeing we are nearing the top of the hour. Any final questions on Solr? (Otherwise, we'll keep this on the agenda and check back in later)

Mark Wood [9:55 AM]
Nothing here.

Tim Donohue [9:55 AM]
Ok, @terrywbrady you mentioned having another topic? Do we have time today? Or for next week?

Terry Brady [9:55 AM]
When we have an external solr, will we access it as [root]/solr, or is that url format not necessary?

Mark Wood [9:56 AM]
It will likely be on another port, so we can do as we please.

Terry Brady [9:56 AM]
(My other item can wait_
Oh, so we do not need to drop webapps/solr in order to test... we just need to update the solrUrl and verify that the new cores are being used
Keeping the old cores accessible might be useful for migration

Tim Donohue [9:58 AM]
Looking at my calendar, I just realized I can only make the first 1/2 of this meeting next week (my daughter

has a holiday party that afternoon). So, next week's meeting (Dec 5 at 15UTC) we'll try to keep to 30 mins, if it goes over, I'll have to turn discussion over to someone else

Mark Wood [9:58 AM]
OK,we'll cope with it.

Tim Donohue [9:59 AM]
Well, since we've finished up Solr discussions for today, let's wrap up the meeting. We'll check back in next week, and if you need more help/support @mwood on Solr, just ask on #dev
Thanks all!

Mark Wood [9:59 AM]
Will do. Thanks, all.

Mark Wood [10:51 AM]
https://wiki.duraspace.org/display/DSPACE/Upgrading+Solr+Server+for+DSpace updated from today's meeting.