DevMtg 2019-03-13

Developers Meeting on Weds, March 13, 2019

Today's Meeting Times

- DSpace Developers Meeting / Backlog Hour: 20:00 UTC in #duraspace IRC or #dev-mtg Slack channel (these two channels sync all conversations)

Agenda

Quick Reminders

Friendly reminders of upcoming meetings, discussions etc

- DSpace 7 Working Group (2016-2023): Next meeting is Thurs, March 14 at 15:00 UTC. Agenda: 2019-03-14 DSpace 7 Working Group Meeting
- DSpace 7 Entities Working Group (2018-19): Next meeting is TBD. On hold until Entities PR to `master` is released (coming soon)
    - Last meeting notes at 2019-02-05 DSpace 7 Entities WG Meeting
- 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. Upgrading Solr Server for DSpace (Mark H. Wood)
    a. PR https://github.com/DSpace/DSpace/pull/2058
    b. Docker configuration for external Solr
        i. https://github.com/Georgetown-University-Libraries/DSpace/commit/7115173d61776dd2455690518f5c9809cd0f28d4
            1. The Dockerfile creates a new solr instance with 4 cores. It then overlays the schema and config changes in PR 2058.
            2. I attempted to create my branch so that I could create a PR back to Mark's branch, but some other changes from master seem to be showing up if I create a PR.
        ii. This will need a small change to our docker compose files to invoke the external solr service. https://github.com/DSpace-Labs/DSpace-Docker-Images/pull/79
4. DSpace Backend as One Webapp (Tim Donohue)
    a. PR: https://github.com/DSpace/DSpace/pull/2265 (PR is finalized & ready for review)
    b. A follow-up PR will rename the "dspace-spring-rest" module to "dspace-server", and update all URL configurations (e.g. "dspace.server.url" will replace "dspace.url", "dspace.restUrl", "dspace.baseUrl", etc)
5. DSpace Docker and Cloud Deployment Goals (old) (Terrence W Brady)
    a. Creating default AIP dataset for DSpace 7 docker load
        i. Tim shared a link to the entities WG dataset. This dataset contains no bitstreams. How should we handle this for the AIP's?
    b. https://github.com/DSpace/DSpace/pull/2362 - update sequences port
    c. https://github.com/DSpace/DSpace/pull/2361 - update sequences port
    d. Add Docker build/push to Travis
        i. This make sense to consider after 2307 is merged
        ii. https://github.com/DSpace/DSpace/pull/2308
6. 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. Bulk Operations Support Enhancements (from Mark H. Wood)
    c. Curation System Needs (from Terrence W Brady)
7. 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
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 CDT)

```
Tim Donohue [3:01 PM]
@here: It's DSpace DevMtg time.  Agenda is at: https://wiki.duraspace.org/display/DSPACE/DevMtg+2019-03-13
Let's do a quick roll call to see who is able to join today

Mark Wood [3:02 PM]
Hi.

Terry Brady [3:02 PM]
here

Tim Donohue [3:02 PM]
Looks like it's the usual crew for this meeting :slightly_smiling_face:  Hi Mark & Terry
Let's go ahead and dive in.  As usual the topics are mostly updates...we'll see where it takes us, and we can
always close up the meeting earlier if needed.
First up, I should mention here...the big news of today is that LYRASIS and DuraSpace are merging by July 1.
https://duraspace.org/lyrasis-and-duraspace-announce-merger-expanding-the-capacity-of-the-global-scholarly-and-
scientific-research-ecosystem/ (edited)
That's really just an FYI for those here. I don't expect anything will change from your perspective (except
I'll eventually be sending you emails from a Lyrasis email address, etc)
If questions do come up though, let me know.  Again, I personally don't expect any changes to affect the goals
/plans for DSpace in the coming months. I just may get pulled away (briefly) for merger related stuff in the
coming 3-6 months.
Moving right along though...on the DSpace 7 side, I don't really have any updates to share specific to this
meeting.  We're still pushing hard for DSpace 7 Preview Release (as soon as possible).  Much more in the DSpace
7 meetings (next one is tomorrow)
On the DSpace 6 side, the updates are also the same.  I expect a 6.4 will happen at some point, but no exact
timelines yet (waiting really on someone to be "freed up" enough to take a lead on the release)
Any questions on any of that before we move along?

Terry Brady [3:08 PM]
no questions

Mark Wood [3:08 PM]
Nope.

Tim Donohue [3:09 PM]
moving right along then. :wink:
@mwood: any updates you'd like to share on the Solr upgrade side of things? https://wiki.duraspace.org/display
/DSPACE/Upgrading+Solr+Server+for+DSpace or https://github.com/DSpace/DSpace/pull/2058

Mark Wood [3:10 PM]
The DSpace 7 WG will be asked to look it over tomorrow, I believe?  Hoping to merge the fresh_install part soon.

Tim Donohue [3:11 PM]
Yep, DSpace 7 WG will look at the PR again tomorrow.  So, I hope PR#2058 (the first stage) will be merged soon
(hopefully even tomorrow)

Mark Wood [3:12 PM]
Several of us looked at how to upgrade cores, and after several suggestions were examined, it looks to me like
rebuilding indexes is the way.  I've begun work on a tool to do that for 'authority'.  The other cores can be
dump/restored or regenerated from the database already.
```

Tim Donohue [3:12 PM]
Sounds good.

Mark Wood [3:12 PM]
It appears that the new field implementation classes that we had to move to would interfere with upgrade-in-place.
Still to be thought about: sharding. I think that dump/restore will take care of the data there, too, but we have new options for how we do this that make it much more automatic.

Terry Brady [3:15 PM]
If someone has _already_ sharded, they will need to run the export from every shard. Fortunately, the process is the same for each repo. (edited)

Mark Wood [3:15 PM]
A question just occurred to me: if sharding is mostly hands-off, do we even want to make it optional anymore?

Terry Brady [3:16 PM]
I think that we should let Solr handle it and we may no recommendations in that area.

Tim Donohue [3:16 PM]
can we even control sharding anymore, if the Solr index is external? Or is that controlled at the Solr level?

Mark Wood [3:16 PM]
I need to read up on it. I think there are a few settings in the collection that affect this, but I'll work it out when I get there.

Terry Brady [3:17 PM]
For usage reporting, we need to retain the ability to access the older records. Hopefully solr makes that invisible to the client.

Tim Donohue [3:17 PM]
Sounds good. At the very least we'll likely want to learn enough to *document how to now do sharding* (even if those docs mostly just point at Solr sharding docs). But, yes, it sounds like we may need to dig more in this area

Mark Wood [3:17 PM]
Yes, it should be invisible.
Ah, I just remembered: Time Routed Aliases *require* cloud mode, so maybe this will still be optional. I'll try to come up with tradeoffs and recommendations.
In general, we're going to do a lot more pointing at the Solr documentation and less Telling You How.
We should document what DSpace requires, and point toward Solr doco. on how to provide that.
I will try to restrain my tendency to mention every possible option, and stick to basics. Sites with fancy needs can read the Solr manual to see how to meet them. (edited)

Tim Donohue [3:21 PM]
yes, I agree. We just need to document the basics, and let Solr document the rest. Just like we do with Tomcat, etc.

Mark Wood [3:22 PM]
I think that's all I have, unless there are questions.

Tim Donohue [3:22 PM]
I've got a (potentially side-tracking) novice question about the `authority` index. Do we know where/how this index is being used within DSpace? Is it just used for ORCID, or are there other use cases?

Mark Wood [3:23 PM]
ORCiD is one source of authority, but I believe there are a few others, and sites can add their own.
But I don't know much about it. We don't use it here.

Tim Donohue [3:24 PM]
The reason I ask, is that I'm starting to wonder if Entities will provide us with a way to *reindex* these authorities. We should be able to store more of these identifiers on Entities now (so that primary storage is no longer a Solr index). However, this is a very "half baked" idea in my head...

Mark Wood [3:24 PM]
That *is* attractive.

Tim Donohue [3:24 PM]
and honestly, this half baked idea likely won't happen for DSpace 7. But, *maybe* for DSpace 8

Mark Wood [3:24 PM]
Yes.

Tim Donohue [3:25 PM]
Ok, I just wanted to plant this idea in your heads too. Again, I'm not yet sure if it's really plausible (for all use cases of the `authority` index), but It'd be really nice to get these out of Solr being "primary storage"

Mark Wood [3:25 PM]
Don't let's forget that. We wouldn't need the dump/restore tool if that index were a cache, as Solr was meant to be.

Terry Brady [3:25 PM]
I think there is some optional controlled vocabulary thing that can save some allowed values in the authority index.
It pulls a set of LOC author or publication names...
I have no idea how actively that feature is used

Tim Donohue [3:26 PM]
@mwood: exactly. I don't want this to sidetrack the dump/restore tool idea though, as I think we likely will still need that for DSpace 7.

Mark Wood [3:26 PM]
Oh, yes, we will.

Terry Brady [3:27 PM]
https://github.com/DSpace/DSpace/blob/master/dspace/config/dspace.cfg#L1434-L1444

Tim Donohue [3:27 PM]
@terrywbrady: good to know. It's possible we could still represent that in Entities though. If we have an Author Entity, it should be able to store an ORCID, an LOC identifier, etc (as metadata). Then we can just index that info into the `authority` index for Solr
thanks for that link. I forgot about that section of the dspace.cfg
In any case, there's nothing more that I wanted to say about that. I mostly wanted to plant this idea somewhere. It really only came to me this week -- so, again, we'll have to see how it "plays out" as Entities moves along. But, logically, it seems like an opportunity to cleanup our very odd Authority Control framework.

Mark Wood [3:30 PM]
Odd indeed. Consider it planted.

Tim Donohue [3:31 PM]
Any final thoughts/questions on Solr upgrade? Sounds like we've wrapped this up?

Mark Wood [3:31 PM]
Nothing more here.

Tim Donohue [3:31 PM]
Ok, moving right along to "DSpace Backend as One Webapp" (from me)
The main update here is that I've decided that we *might* want to keep the existing PR *as-is* in order to avoid major conflicts with other ongoing work. So, I now consider this PR "final": https://github.com/DSpace/DSpace/pull/2265
I plan to now follow it up with a PR to *rename* the `dspace-spring-rest` webapp to `dspace-server` (webapp), and collapse all the URL-based configs down to one (`dspace.server.url`) in dspace.cfg

Mark Wood [3:33 PM]
So you want to hold off until that other work lands, fix up conflicts, and then merge One Webapp?

Tim Donohue [3:34 PM]
I want to do that rename in a *separate PR* as renaming involves touching/moving a massive number of files (literally >400 files), and that's going to result in a large number of conflicts with any other work going on with `dspace-spring-rest`.

Mark Wood [3:35 PM]
Ah, I get it: "keep the existing PR as-is" = don't combine it with the rename.

Tim Donohue [3:35 PM]
@mwood: My inclination is to move PR#2265 forward as-is, as soon as folks are "ready for it". It results in a semi-odd `dspace-spring-rest` webapp that does more than REST. But, it ensures #2265 won't have massive conflicts with everything else (should be mostly "backwards compatible")

Terry Brady [3:36 PM]

Do you need to move the source files or could you just rename the exposed path to the service?

Tim Donohue [3:36 PM]
@mwood: exactly. #2265 will not change any further.

Mark Wood [3:36 PM]
That makes perfect sense.

Terry Brady [3:37 PM]
I suppose if you are ever going to move the files, it would be good to do it in the near future.

Tim Donohue [3:38 PM]
@terrywbrady: I'm going to need to move the files eventually. I agree, I'd rather do that sooner than later, but currently we have several big DSpace 7 PRs "in the works". If I throw in the rename now, it's gonna be painful

Terry Brady [3:38 PM]
Your timing makes sense.

Tim Donohue [3:38 PM]
So, my splitting this idea into two PRs is simply so that we can merge one "now" (ish), and merge the other in a week or two (once other PRs have been merged)

Mark Wood [3:38 PM]
I like it.

Tim Donohue [3:39 PM]
Glad to hear it makes sense. So, that's my main update. I've already started the rename process (no PR yet though). But, in the meantime, the current PR is "final" and just waiting on reviews: https://github.com/DSpace/DSpace/pull/2265
Any other questions/comments on this PR / effort in general?

Mark Wood [3:41 PM]
Nothing constructive. I have a prejudice against Spring Boot but not enough to object to this effort. I'll learn to deal with it.
I think that gathering up the various app.s makes sense.

Tim Donohue [3:43 PM]
I actually love Spring Boot :slightly_smiling_face: Sorry to hear you aren't as much in favor. I've found overall it's not much more than Spring MVC (which it borrows very heavily from). Beyond that, it's mostly a few "convention over configuration" ideas to get used to.
If we ever were to decide it was too much of a pain, I do think it's rather easy to back out and move to plain old Spring MVC though.
In any case, we'll move along to other topics
Next up is DSpace + Docker updates from @terrywbrady: https://wiki.duraspace.org/display/~terrywbrady/DSpace+Docker+and+Cloud+Deployment+Goals

Terry Brady [3:45 PM]
I had another thought since our conversation this morning about https://github.com/DSpace-Labs/DSpace-Docker-Images/pull/93
See my last comment.
Just an fyi... the ports of the update-sequences enhancement are
- https://github.com/DSpace/DSpace/pull/2362
- https://github.com/DSpace/DSpace/pull/2361
I may want to re-open a version of PR 93 in order to allow a map of env settings to be assembled. If I do that, I will plan to completely overwrite local.cfg.

Tim Donohue [3:48 PM]
I'm not sure I understand your "single string is passed in" comment in PR#93. Could you expand on what you mean / how this works?

Terry Brady [3:49 PM]
Each override file has the opportunity to set JAVA_OPTS. The last value set will take precedence.
While it is unrealistic that some would need to test with both Oracle and RDF, lets imagine that scenario. The last set of JAVA_OPTS set will be the one that is passed to the container.

Tim Donohue [3:50 PM]
Can't we just append to current JAVA_OPTS from each? Something like `JAVA_OPTS=${JAVA_OPTS} new-value"?   (I'm forgetting the exact syntax, but I think you can append to environment variables like this)

Terry Brady [3:50 PM]

That is what makes it attractive to use each individual env variable.

Mark Wood [3:51 PM]
Modulo quoting, yes, that's how you append.

Terry Brady [3:51 PM]
The yaml works as an overlay. I do not believe the values are interpreted or computed.
I can create some sample files to "prove" my assumption.

Tim Donohue [3:52 PM]
oh, so the yaml is all combined first, then the environment variables are set (based on the combined yaml)
Yes, it might be nice to verify that is true....just so we know for certain how it is behaving

Terry Brady [3:52 PM]
Yep. That is why my initial implementation set 3 different variables allowing multiple points of override.
We could do a JAVA_OPTS1, JAVA_OPTS2, ... and then combine those.
I'll write up a little proof and then construct another option for you to review. Thanks again for thinking though this with me.

Tim Donohue [3:54 PM]
If it's easier to just completely overwrite the `local.cfg` with an "initial setup that works for Docker", honestly that's OK by me. It seems like (and correct me if I'm wrong) this main issue is all in the *initial spin up of Docker*. So, it might make sense to have that "initialize" the `local.cfg` with sane values too

Terry Brady [3:55 PM]
I think so. I'll package up some examples and ask you all for input. I want to anticipate what folks are going to want in the future.
Regarding the last item on the agenda, I want to continue to defer the travis work until people are ready to use it.
I still think we are building up the usage of Docker for testing.
We can take that off the agenda for now.

Mark Wood [3:57 PM]
That would be 5d.

Tim Donohue [3:57 PM]
Yes, I agree. It's good to keep in mind, but we still need to get Docker setup more "finalized" -- it still seems to be changing/improving a lot right now
Once the Docker setup stabilizes and is more widely used, then it would make sense to look more at Travis doing builds/pushes

Terry Brady [3:58 PM]
I am pretty delighted that our compose files can also be used to trigger a build of a docker image. See https://github.com/DSpace-Labs/DSpace-Docker-Images/blob/webinar/documentation/run.CommonTasks.md#building-dspace-code
I can build a docker image faster than I can run maven/ant locally.

Tim Donohue [3:59 PM]
nice :+1:

Terry Brady [4:00 PM]
I am unsure if the future need will be to (1)provide pre-built PR images or to (2)deploy a running version of the code for a PR.
There are endless opportunities to refine this stuff...

Tim Donohue [4:01 PM]
We're at the top of the hour here...so we likely should wrap things up. One (minor) PR I forgot to mention... I'm looking for a second +1 on this minor Travis / Coveralls reconfiguration: https://github.com/DSpace/DSpace/pull/2367

(Not high priority, but it fixes issues folks had recently when Coveralls.io was down one day)

Mark Wood [4:01 PM]
I looked it over today. It all makes sense, and obviously it didn't break anything.

Tim Donohue [4:02 PM]
@terrywbrady: Yes, the Docker stuff is really exciting. I may finally get a chance to start using it more heavily as I transfer to Lyrasis (as local tasks requiring Vagrant may start to be lifted off my plate) (edited)

Terry Brady [4:02 PM]

I have also been able to get it to run within an EC2 instance which is handy.

Tim Donohue [4:03 PM]
Ok, we're over time here. So, it's time to wrap things up. I don't have any more agenda items to call today (and I expect you both need to run off to other things)

Mark Wood [4:03 PM]
I just added my review to 2367.

Tim Donohue [4:04 PM]
Thanks for the discussion today & we'll talk again very soon (if not in the DSpace 7 mtg then in Slack, surely)!

Terry Brady [4:04 PM]
Have a good week!

Mark Wood [4:04 PM]
'bye all.