DevMtg 2019-04-24

Developers Meeting on Weds, April 24, 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, April 25 at 15:00 UTC. Agenda: 2019-04-25 DSpace 7 Working Group Meeting
- DSpace 7 Entities Working Group (2018-19): Next meeting is TBD
    - 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. **Tim is unavailable next week** (May 1). Will be at a staff retreat. Should we cancel the meeting?
2. (*Ongoing Topic*) DSpace 7 Status Updates for this week (from DSpace 7 Working Group (2016-2023))
3. *(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.
4. Upgrading Handle Server: Unable to locate Jira server for this macro. It may be due to Application Link configuration. (READY for Reviews!)
    a. PR: https://github.com/DS
5. Upgrading Solr Server for DSpace
    a. **Auto-reindexing in Solr**: Unable to locate Jira server for this macro. It may be due to Application Link configuration.
        i. Should this only h need to reindex?
    b. e `authority` core. Unable to locate Jira server for this macro. It may be due to Application Link configuration. Or should we use s
6. DSpace l p (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)
7. DSpace Docker and Cloud Deployment Goals (old) (Terrence W Brady)
    a. Update sequences on initialization
        i. https://github.com/DSpace/DSpace/pull/2362 - update sequences port
        ii. https://github.com/DSpace/DSpace/pull/2361 - update sequences port
8. 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 CDT)

```
Tim Donohue [3:01 PM]
@here: It's DSpace DevMtg time.  The agenda for today is at https://wiki.duraspace.org/display/DSPACE
/DevMtg+2019-04-24 (mostly a copy from last week)
Let's do a quick roll call to see who is joining today

Mark Wood [3:02 PM]
Here.

James Creel [3:03 PM]
Here

Tim Donohue [3:04 PM]
Looks like we are a small group today, but we may as well get started.  We'll see where this agenda takes us
I'll admit, on the DSpace 7 side, I don't think I have any specific updates to share this week.  I'm starting
to feel like a repeating record (in various meetings), so I also don't recall what I mentioned last week :
wink:  The basics are that we are working towards an initial Preview by end of April (hopefully).  A second
Preview by OR2019.
Beta & Final are less nailed down (hoping for Summer then Fall), but we're working on better estimating
remaining work in the DSpace 7 meetings (starting tomorrow).  Once we have better estimates of what's
remaining, we'll be able to more easily schedule those.
I think that's it, but if you have questions or feel unclear on something, please ask. I'm glad to clarify
anything DSpace 7 related
DSpace 6.4 is the same as recently as well. I don't have any updates to speak of at this time.  Still waiting
on volunteer(s) to help move this forward (or someone to free up elsewhere) (edited)
Any questions on those quick/brief updates?
Not hearing any
Oh, I skipped #1 on the agenda.  I forgot. I'm out of the office much of next week (Mon-Weds). That means I
won't be available for this usual meeting.  I was leaning towards cancelling next week (as this has turned into
an "update meeting"), but if others disagree, you are welcome to organize without me :slightly_smiling_face:

Mark Wood [3:10 PM]
OK

Tim Donohue [3:11 PM]
Thoughts?  Should I just cancel it?

James Creel [3:11 PM]
It's ok with me to cancel

Mark Wood [3:11 PM]
If there are folks who think that something needs discussion sooner than 2 weeks out, I could moderate next
week.
Otherwise cancellation is okay here.

Tim Donohue [3:12 PM]
I don't have any specific topics for next week (and don't anticipate any).  But, you are welcome to keep this
slot open and ask more on #dev for topics.
As noted, the agenda for this meeting has been very static recently (just updates).  So, it might depend on
whether others have updates, honestly
I'll leave the meeting on the calendar for now, @mwood.  If you find topics you (or others on #dev) want to
discuss with others, maybe it's worth touching base.  I just don't know of any off the top of my head
```

Mark Wood [3:15 PM]
OK.

Tim Donohue [3:15 PM]
Moving along for now...next is the usual updates on ongoing work/effort
First up, the PR to upgrade our (embedded) Handle Server to version 9 is ready for testing/reviews: https://github.com/DSpace/DSpace/pull/2394
So, this is just a friendly reminder that that could use more eyes.
I don't have any other updates beyond that, but if anyone has questions or wants to volunteer to take a closer look, this would help out the DSpace 7 team (and you don't need to know DSpace 7 really to review this)
That's it for that update
Next up is an update on the ongoing Solr Server upgrade work (from @mwood). Anything to mention this week?

Mark Wood [3:19 PM]
I have been trying out the solr-export-statistics command for dumping/restoring the statistics core to get it rebuilt.
There are some observations of the dumping process over on #dev.
Now I'm trying to load the dumps using solr-import-statistics. I had a little side-trip to make a gadget for repairing missing 'uid' values, also mentioned on #dev.

Tim Donohue [3:21 PM]
Should some of those observations be tracked more formally? Or turned into early docs/notes on the Wiki? I'll admit, I've been a bit swamped this week, and haven't been able to follow the entire #dev discussion to pull out the important points

Mark Wood [3:21 PM]
This process is going to take several hours.
Yes, probably I should capture my observations in more durable form.
I will drop them somewhere on one of the "what's going on with Solr" wiki pages for now. (edited)

Tim Donohue [3:23 PM]
I'm assuming though that this process could be run "behind the scenes" (while DSpace is up and running)? I guess I'm wondering if this is slightly painful (just takes a while to get to completion) vs quite painful (DSpace would need to be down the entire time)

Mark Wood [3:24 PM]
DSpace should be inaccessible while dumping. Otherwise we won't get all of the stat.s records. It could be fully available while loading, I think, but the stat. displays would be wrong until the process completes.

James Creel [3:25 PM]
That happens when redoing indexes anyway - probably not awful

Tim Donohue [3:25 PM]
Why wouldn't you "get all the stats records" if DSpace is running?

Mark Wood [3:25 PM]
Perhaps one could dump them online, then go inaccessible and re-dump the last day, trim off the duplicates from the first dump run, and load in background.
There would be new records going in while the dump is being written out.

Tim Donohue [3:27 PM]
Or, you could just say: "Sorry, you may lose stats records during this 1-2 hour upgrade period, but all past records will be available after reloading is complete."

Mark Wood [3:27 PM]
True.

Tim Donohue [3:27 PM]
I don't think it's horrible to be honest that the upgrade takes time, you can choose to keep the site up, but the stats during that period may not be "captured" fully.

Mark Wood [3:27 PM]
If losing some usage records for a few hours is okay then the whole process can be done in background.
We'll need to document the trade-off.

Tim Donohue [3:28 PM]
That's probably a better message then.. "you will need to keep your DSpace offline for 1-2 hours to fully upgrade stats". I see that being more of a "no go" for some sites

Mark Wood [3:28 PM]

Depending on how we fetch the documents from Solr, the actual loss could be quite small.

Tim Donohue [3:29 PM]
But, yes, we could document the trade-off here and let each site decide which way they want to go

Mark Wood [3:29 PM]
I will have another look at that code.

Tim Donohue [3:29 PM]
Sounds good. In any case, it's good to understand the scope of the upgrade/migration of stats
Another quick question. Did you ever determine if the `solr-export-statistics` (and import) also work for the authority records, @mwood? I see that's still a quesiton in our agenda.

Mark Wood [3:31 PM]
It is supposed to. I was able to dump and restore the authority core, but since we don't use authority here there were no documents. So it wasn't much of a test. I'll see if I can rig up enough authority stuff to get a nonempty core for testing.

Tim Donohue [3:32 PM]
You should be able to generate some authority records (a handful) pretty easily by enabling ORCID, then creating some documents where you lookup some authors via ORCID.

Mark Wood [3:33 PM]
Noted. Thanks. (edited)

Tim Donohue [3:33 PM]
I'm not sure if anyone has a bigger test set of authority (I'm sure there are some out there) but it could be something to ask about in #dev or in a DSpace 7 meeting if someone can just share a core

Mark Wood [3:33 PM]
Also a possibility.

Tim Donohue [3:34 PM]
Ok, thanks for those updates. It all sounds fine overall. We'll just need to nail down the process, and be honest about the options/balance between site downtime or running the upgrade in the background

Mark Wood [3:35 PM]
I will write up what I have, and continue working toward a full end-to-end test on both cores.

Tim Donohue [3:35 PM]
Sounds great!

Mark Wood [3:36 PM]
As noted above, some sites may have stat. documents that need repair before they can be imported. Somehow we have some with no 'uid' field. There is now a tool to fix that.

Tim Donohue [3:37 PM]
That's good to note as well. It sounds like you could start an 'early draft' of an upgrade process/procedure using what you've learned so far. Then as it gets more into a "good draft" we could have others test it out as well

Mark Wood [3:38 PM]
IIRC there is already a very early draft in the DS7 doco. I will keep it updated as I proceed.

Tim Donohue [3:38 PM]
Ok, let's move along now. Next up is updates on the "One Webapp" backend PR: https://github.com/DSpace/DSpace/pull/2265
Just wanted to note here that the PR is now at +2, and I hope to merge it in the near(ish) future, pending a final approval of the DSpace 7 team. That said, I'll also note that with @terrywbrady's help, it looks like we've hit some oddities in how this gets deployed via Docker. I don't think they are of fault of the code in this PR though, rather they are part of the Docker configs
Still looking at the Docker side of things (and I'm really liking Docker so far, finally took the plunge last week), but I think this PR could go in as-is and Docker stuff could be cleaned up later
That's really all the updates I have. I'm just hoping to get this done soon, as I know there's followup work to be done :wink:
So, I think we can move along
I know Terry isn't here today, but I thought I'd add in a reminder that his two backports of the new `dspace database update-sequences` command need reviews: https://github.com/DSpace/DSpace/pull/2361 and https://github.com/DSpace/DSpace/pull/2362
These PRs aren't really Docker specific (they are adding a commandline tool), but Terry built them cause they make Docker initialization a bit easier.

Honestly, everything else on this agenda is a bit outdated at this time...there are Brainstorms / ideas still on here, but they all need revisiting at some time in the future (as we're all concentrating on other activities)
We have about 15 minutes left. Any other topics / discussion for today from anyone?

James Creel [3:46 PM]
Just thought I'd mention I'm just starting to use OpenRefine to help librarians with some of their curation tasks. Currently looking for duplicates in the existing IR and in a big incoming batch.
In combination with the batch metadata edit, I think it is extremely powerful

Mark Wood [3:46 PM]
We probably need more discussants, but I would like for us to work out just what are legal values for metadatavalue.text_lang. Are they Java language tags, IETF tags, or something else?

James Creel [3:47 PM]
Mark, I think there are some ISO standards or something. If you like, I can dig up some of the intelligence our librarians gathered.

Tim Donohue [3:47 PM]
@jcreel256: I admit, I'm not familiar with OpenRefine. But that process sounds like it could make a good webinar / workshop and/or "how to" on the Wiki.
@mwood: I think they should be ISO language codes
"they" being "text_lang" info

Mark Wood [3:48 PM]
ISO 636 defines language codes. IETF BCP47 references it and adds national/regional variants (plus other stuff I've never seen used).
Pure ISO 636-1 2-letter codes? No variant suffix?

Tim Donohue [3:50 PM]
Because our official example in dspace.cfg has a variant suffix, it seems like it was meant to be ISO 639-1 *with variant suffix*
https://github.com/DSpace/DSpace/blob/master/dspace/config/dspace.cfg#L52
The official default example is "en_US" :point_up:

Mark Wood [3:51 PM]
OK, so that's the Java form.

Tim Donohue [3:52 PM]
But, as noted in #dev , I don't know if we actually validate these values in any way. We expect them to be ISO codes, but they could say anything

Mark Wood [3:52 PM]
Our repo. is living proof that they *can* be anything. :rage:

James Creel [3:53 PM]
We're in the same boat, empties, nulls, etc. But I'm working on it.
nulls are fine, though

Tim Donohue [3:55 PM]
There probably should be more "built in" validation in our API layer. But, I'm also not sure if it should be super strict, or throw warnings -- super strict gives better data, but could also be frustrating for bulk ingests if language codes are not all 100% correct
In any case, I agree we need to improve this

Mark Wood [3:56 PM]
OK, probably needs a bit more discussion in a future meeting (but not *too* future!)

Tim Donohue [3:58 PM]
Seems reasonable. Though, at the same time, I think we have a good idea of what it should be validating *for*. It seems doing simple validation (throw warnings) could be a quick win. I think the question is just whether strict validation will cause more problems than it's worth, or if we can make it work too.
And cleanup tools sound like they are desperately needed across the board
In any case, I see we are at the top of the hour, so I'd suggest we wrap up for today.
As noted, next week (week of April 29) I won't be around Mon-Weds, but will be back Thurs & Fri. Have a good rest of your week!

James Creel [4:00 PM]
take care

Mark Wood [4:01 PM]

Thanks, all! 'bye.
Idea for later discussion: warn on ugly values, accept as written, provide a curation task to generate reports on metadata that should be inspected by an admin.