DevMtg 2019-04-17

Developers Meeting on Weds, April 17, 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)

Agenda

Quick Reminders

Friendly reminders of upcoming meetings, discussions etc

- DSpace 7 Working Group (2016-2023): Next meeting is Thurs, April 18 at 15:00 UTC. Agenda: 2019-04-04 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. (*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.
    b. TAMU Sprint - Fix
3. 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
4. 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
5. 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)
6. 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
7. 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)
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. 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 [10:00 AM]
@here: It's DSpace DevMtg time.  Agenda is at https://wiki.duraspace.org/display/DSPACE/DevMtg+2019-04-17

Let's start with a quick roll call, as always

Mark Wood [10:01 AM]
Here!

James Creel [10:01 AM]
James here with my fellow developers Ryan and Kevin.

Bill Tantzen [10:01 AM]
Just lurking...

James Creel [10:02 AM]
We're just wrapping up a local sprint and want to give updates and discuss.

Terry Brady [10:02 AM]
Here

Tim Donohue [10:02 AM]
Welcome all!  Thanks, @jcreel256.  We'll make time for that today.  I think most of this meeting is simply
"quick updates" anyhow, so jump in when you feel it may be most appropriate

James Creel [10:03 AM]
I got us on the agenda at the end

Tim Donohue [10:03 AM]
Let's get started, since we have a lot on the plate...
On the DSpace 7 side, I'll keep things brief.  The release is progressing, and the first "Preview" release is
planned for late April (Preview releases won't be feature-ful, but will highlight major new features).
There will be a second "Preview" before OR2019.  The Beta is now likely later in the Summer, with Final Release
in the Fall.
The primary delays at this point have been in code review.  Development is moving along decently, but some of
the major features have taken much longer to code review (as you might expect) and have to sometimes go through
several iterations of review/test, fix, review/test, fix, review/test, fix, etc
That's the basics... but if you want all the details, obviously the weekly DSpace 7 dev meetings (every Thurs)
are a place to listen in for all updates, etc
Any questions/comments on DSpace 7 updates?

James Creel [10:07 AM]
So the UI and backend meetings are combined?

Tim Donohue [10:08 AM]
Yes, it's one meeting now, and we've restructured that meeting into (1) a 15 minute Sprint-like "Standup", (2)
a 30-minute discussion period, and (3) a 15 minute planning period for the next week.
Here's tomorrow's agenda which shows you the new structure: https://wiki.duraspace.org/display/DSPACE/2019-04-
18+DSpace+7+Working+Group+Meeting
You'll also notice the agenda includes a detailed list of all the features/PRs/tickets being actively worked
on.  See the "Current Work" section of the agenda
So, even if you don't attend DSpace 7 meetings, the last week's agenda is a great place to look for PRs to help
```

review, or just to see exactly what we are working on right now
Any other questions/comments?
Ok, moving along for now. Please do reach out though if you have DSpace 7 questions. The team is moving quickly, but we are always glad to get others involved (whereever we can). Get in touch if there's interest in doing so
On the DSpace 6 side, I have no updates to share at this time. We still plan for a 6.4 release, but currently we have no one who has volunteered to lead/organize that release. So, until someone is found (or someone's time frees up), that release is "on temporary hold"
Any questions/comments to add on DSpace 6.x side?

James Creel [10:13 AM]
Our last sprint involved fixing a bug with harvesting in 6.x which we can detail at the end

Tim Donohue [10:14 AM]
@jcreel256: Honestly, if it was 6.x specific, I'm OK with you detailing the sprint info now. The other items on this agenda are "hold over" topics from last week...so, it'd just be "quick updates" on their status

James Creel [10:14 AM]
Ryan over here had some trouble building the 6.x branch, but has gotten some feedback on Slack.

Tim Donohue [10:15 AM]
Let's talk your 6.x Sprint now then :slightly_smiling_face:

James Creel [10:15 AM]
We need to get the build working, verify the fix, and then we'll craft a PR to add to the existing multitude.
Currently, our local build has the fix in the additions.

Ryan Laddusaw [10:16 AM]
joined #dev-mtg.

Tim Donohue [10:16 AM]
@jcreel256: was the build issue in out-of-the-box DSpace 6.x? I saw a question from Ryan on #dev , but that seemed like a build problem caused by local tweaks/additions (unless we misunderstood it)

Ryan Laddusaw [10:17 AM]
Yeah, i just haven't had a chance to follow up on that.
And it was an out of the box build

Terry Brady [10:19 AM]
I'll make a pitch here... as you are testing changes, consider building and testing within Docker:
https://github.com/DSpace-Labs/DSpace-Docker-Images/blob/master/documentation/run.CommonTasks.md#building-dspace-code

Tim Donohue [10:19 AM]
Ok, I think we'd definitely like to get some more info then (when you get a chance). The error shared in #dev channel mentioned something that wasn't "out-of-the-box". But, it sounds like your deeper dive found other issues.
So, it seems appropriate to (obviously) get a JIRA ticket created to describe the bug, and then a PR to followup (once you verify the fix)

James Creel [10:20 AM]
The other big thing we were looking at was normalizing the language qualifiers on metadata values. They were all over the place, nulls, empty strings, en, en_US, etc. and it annoyed curators when you ended up with lots of extra columns in exports. @terrywbrady had provided some SQL to normalize these, and we expanded on that quite a bit to fit local requirements.
Still wrapping that up, but will be happy to share when it's complete.

Tim Donohue [10:21 AM]
With regards to normalizing, are you fixing code to achieve this? Or are you just building a SQL to clean it up (where it exists)? Either way, sounds like something useful to share

James Creel [10:22 AM]
It is strictly SQL to do cleanup.

Tim Donohue [10:22 AM]
Ok, good to know. Still would be worth sharing on the Wiki somewhere. Longer term we'll have to find where these oddities are introduced (in the code), which is harder to track down.

Mark Wood [10:24 AM]
Certainly we should not permit null or 0-length metadatavalue text. IMHO the proper way to say "this field has no value" is to have no metadatavalue row for it at all.

Terry Brady [10:24 AM]
I think the UI and the SAF ingest processes may have (or may have had) different default language settings which can cause these issues to creep in from time to time.

Tim Donohue [10:26 AM]
If we don't have these in JIRA yet, obviously would be good to note any obvious bugs we are aware of : slightly_smiling_face: I admit, I haven't dug in this area yet myself

James Creel [10:26 AM]
To be clear, @mwood we're just talking about the language column that sits alongside the text value column in the metadatavalue table.

Mark Wood [10:26 AM]
Ah. That shouldn't be 0-length, but null should be okay.

Tim Donohue [10:26 AM]
Any other updates to share from the sprint @jcreel256?

James Creel [10:27 AM]
Kevin over here had been looking at some of the outstanding PRs when he had time. That's it for us.
Thanks for the tip on the language column, @mwood

Tim Donohue [10:28 AM]
Ok, thanks. If you are able to review 6.x PRs, please do add comments/results of your testing. We definitely *will* get those PRs merged in the future (and I've been tracking them). It's just been "quiet" on 6.x PRs from Committers as the active ones are mostly currently working hard on 7.x

Kevin Day [10:30 AM]
joined #dev-mtg.

Tim Donohue [10:30 AM]
As soon as we find a 6.4 coordinator though, I expect many 6.x PRs to get quickly merged in prep for a 6.4 release. It's just a matter of finding someone with time to help coordinate that process -- it will happen, just not sure if it'll very soon, or more like after OR2019
Any final thoughts / comments / questions related to 6.x (in general)? If there are, this is a good time to bring it up...as the rest of the topics here are primarily geared towards 7.x related work.
Ok, not hearing any. So, let's move along
Next up, I wanted to note that the effort to upgrade our (embedded) Handle Server is ready for reviews: https://github.com/DSpace/DSpace/pull/2394
This is for 7.x, and it updates us to the latest version of the Handle Server (v9). It was contributed by the CNRI team, and I recently pushed the latest Handle Server JARs to Maven Central (with their approval)
So the PR now passes all tests & builds properly. It just needs code reviews / testing
@mwood you probably already saw, but I assigned you as a reviewer :wink:

Mark Wood [10:34 AM]
It's on one of my too many todo lists.

Tim Donohue [10:34 AM]
That's really it for that update
(but much more info is in the PR description if it's of interest)
moving along...
Next up, any updates @mwood on the effort to upgrade Solr Server (to v7)? I know the initial PR has been merged (to master).

Mark Wood [10:36 AM]
No, I've been sidetracked by various issues. I need to look at using solr-export-statistics instead of the new exporter I wrote.

Tim Donohue [10:37 AM]
Ok, no worries. But, thanks for the update. Ideally, I'd like to see if we can get the "upgrade strategy" figured out for OR2019 (just to let you know timelines). This would be important to mention as part of the Workshop at OR2019

Mark Wood [10:38 AM]
I would like that too.

Tim Donohue [10:38 AM]
So, this should be scheduled for the 2nd Preview release (which would be released in late May / early June, just before OR2019)
We'll talk more about 2nd Preview Release planning at the DSpace 7 mtg tomorrow though

Mark Wood [10:39 AM]
Noted.

Tim Donohue [10:39 AM]
Moving right along (I'm going quickly through these updates, but if anyone has questions, jump in and type them!)
Another update that I think should be in the 2nd Preview Release... the "one webapp" backend changes (which are on my plate)
I'll just note that I've *finally* got back to fixing this up. I've started using Docker to test it out (with help from @terrywbrady who noted some issues in Docker). So, I'm hoping to fix any (minor) issues soon -- either late this week or early next
In the meantime, other reviews are welcome, but a few small fixes to RDF & OAI will be coming
That's it for that update too
Moving along... Docker updates from @terrywbrady (and I should say, I've enjoyed using Docker so far -- literally started yesterday though!)
Any updates you want to share on Docker, @terrywbrady?

Terry Brady [10:42 AM]
I have 2 very simple PR ports that need a review. https://github.com/DSpace/DSpace/pull/2362 - update sequences port

https://github.com/DSpace/DSpace/pull/2361 - update sequences port
The other Docker agenda items can be closed for now.
We are not yet going to build a system to auto-publish docker images for active PR's. I suspect we may want that in a few months as Docker is more widely adopted for testing.
For major features, we can create a "feature branch" and set that branch up to auto-build on docker hub.

Tim Donohue [10:45 AM]
Just for others following along, the two PRs mentioned above are *not* Docker specific. They are backporting tooling to kick off "update-sequences.sql" from commandline (to 5.x and 6.x). But, those backports were done to make Docker builds easier
So, if that idea is interesting to you (regardless of whether you use Docker), please take a look at PR#2361 and #2362 (linked above)

Terry Brady [10:46 AM]
I am swamped with other work, but I have an outstanding todo to verify that the entities features are demonstratable within Docker. We have confirmed that an instance can be spun up, but I have not verified specific features.
We currently rely heavily on docker-compose to initialize a usable container on startup (admin user, hierarchy and test content populated). The goal here would be to allow both developers and repository admins to test and verify new features using a local docker instances.
In the future, it would be great to make our images cloud-deployable, but we have a bit more work to do to get there.

Tim Donohue [10:50 AM]
As I get more comfortable with Docker, I also hope we can play with cloud deployments. I'd hope to eventually replace demo.dspace.org with a Docker cloud deployment. But, I don't expect I'll have time to look at that until after OR2019 (at the earliest)
Thanks for those updates though, @terrywbrady. I'd encourage others to try out the new Docker tools (they work for DSpace 4.x through 7.x), as they are pretty cool.

Terry Brady [10:51 AM]
I shared this with Tim. I am taking a class on AWS and I plan to build a DSpace PR test system as part of my final project. https://github.com/terrywbrady/CldAws230
I hope to eventually have a little demo to share.
(until my AWS credits run out)

Tim Donohue [10:52 AM]
thanks again!
Ok, realizing we are nearing 1 hour. The remaining topics on the agenda are old brainstorms (that I don't think anyone has gotten back to recently).
Therefore, are there any general questions / comments anyone here would like to bring up / discuss in the remaining 7 minutes?
Ok, I don't have any other topics to bring up for today either. So, I suspect we should just wrap up this discussion a few minutes early

Mark Wood [10:55 AM]
I did want to draw attention to https://github.com/DSpace/DSpace/pull/2397

Tim Donohue [10:55 AM]

I'll remind everyone that this agenda could use "refreshing" with new topics (at any time). I'm pretty much carrying over the last week's agenda into the next week (to keep track of ongoing work). But, other topics are *always* welcome. Pass them my way or add them to the agenda yourself

Mark Wood [10:55 AM]
If anyone has a large repository and concerns about Hibernate performance, it may be interesting. Our Hibernate caching setup is being ignored.

Tim Donohue [10:56 AM]
Thanks @mwood for referencing that PR. We should also add this to the DSpace 7 agenda (to draw attention tomorrow as well)

Mark Wood [10:56 AM]
I will do that.

Tim Donohue [10:56 AM]
thanks!
Ok, since we have nothing more to discuss, we'll wrap up for today. Have a good rest of the week all! Talk to you next week at 20:00UTC

Mark Wood [10:57 AM]
'bye, and thanks.

Mark Wood [11:10 AM]
I just ran a quick check on our IR looking for 0-length text_lang values. "select count(*), element, qualifier from metadatavalue join metadatafieldregistry using(metadata_field_id) where text_lang = '' group by metadata_field_id, element, qualifier order by count desc;" turned up 13,747 0-length contributor.author values, and 3-4 thousand date.issued, identifier.uri, date.accessioned, date.available along with 53 others in lesser numbers.

Terry Brady [11:11 AM]
dc.date fields do not have a lang. I am not certain if the same applies to a dc.contributor

Mark Wood [11:12 AM]
We have a total of 78,479 metadatavalues for contributor.author.
Yes, but the text_lang should not be empty; it should be null.

Terry Brady [11:12 AM]
There could be an inconsistency in the submission UI vs other ingest processes.

Mark Wood [11:13 AM]
Yup, this is just evidence that (a) there is an inconsistency somewhere, and (b) it's not just at one site.

Bill Tantzen [11:16 AM]
my results are of the same order...