DevMtg 2018-11-14

Developers Meeting on Weds, November 14, 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 15 at 15:00 UTC.
- DSpace 7 Entities Working Group (2018-19): Next meeting on Tues, Nov 20 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. Solr as a prerequisite, installed/managed/updated separate from DSpace.
    a. PR https://github.com/DSpace/DSpace/pull/2058
    b. dspace-devel thread: https://groups.google.com/forum/#!topic/dspace-devel/XkYgGgVyGhs
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 CDT)

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

/DevMtg+2018-11-14

Let's do a quick roll call to see who is able to join today

Mark Wood [9:00 AM]
Hi.

Terry Brady [9:01 AM]
hello

Tim Donohue [9:01 AM]
Quite the small group today, huh :wink: I'll assume others may be listening in or joining shortly (as there's a lot of folks showing as "online" in Slack at least)
So, jumping into our agenda...upcoming meetings are listed in the "Quick Reminders". DSpace 7 meeting tomorrow (15UTC), Entities mtg coming up next Tues (16UTC).
An update on DSpace 7 (tentative) release schedules... a small subgroup of our DSpace Leadership Group met yesterday to plan out a tentative schedule
First off, we decided we will *not* do *any* (preview/alpha) release in 2018. We just aren't ready enough, and there's too many features that'd get left out
However, the goal is to have an "early preview" (not feature complete) release in early 2019 (likely late January / early Feb). This release will be made to show off the new/upcoming Configurable Entities feature (s). So, that's our concentration in the coming months.
It likely will *not* be an alpha or beta though, as (at least at this time) we don't feel we'll have all features complete at that time. However, we are tentatively looking for a beta in March/April (will be feature complete), and aiming for the final release in late May (prior to OR2019)
All of these dates are considered "goals" (and therefore they may change), but at least currently they seem doable. We could still use help though (especially in code reviews/testing, and especially of *angular UI* efforts), as more help will keep us on schedule and/or ahead of schedule.
That's the basic summary (as of now). Questions/feedback are welcome, obviously. And we'll talk more on this in tomorrow's DSpace 7 meeting as well

Mark Wood [9:08 AM]
Thanks. I agree: that schedule seems doable.

Terry Brady [9:10 AM]
I am still eager to clear space to make some meaningful DSpace 7 contributions.

Mark Wood [9:10 AM]
The one worrisome bit is the need to get Entities tucked in soon, given that we are approaching a cascade of major holidays.

Tim Donohue [9:10 AM]
If anyone @here is willing to chip in on DSpace 7, as noted, the ways to do so are relatively simple. We primarily could use PR reviewers / testers. While testing/reviewing on the Java side helps, we could use extra help on Angular testing (cause we are lacking reviewers/tests). So, this is a good opportunity to start to get your feet wet in Angular by helping us test out PRs.
(And no, in my opinion, you need not be an Angular expert to help test our Angular PRs. Obviously, as you test more and more, you may get more familiar with concepts and help in code review. But, right now, even just basic testers are welcome)
@mwood: yes, Entities is very high priority to get moving forward. Ideally we do get that merged into master or *close* before end of 2018...as that will ensure we are ready to show it off in a preview release in late Jan
Ok, any other questions/comments on DSpace 7 tentative scheduling?
Not hearing any...we'll move along
I don't have any updates on DSpace 6.x to share. The status is the same, I'm sure a 6.4 will happen at some point, but I don't have an estimate as to when (cause DSpace 7 is highest priority at this point)
So, let's move on to topic #3... Solr upgrade discussions, and specifically treating Solr as a prerequisite, installed/managed/updated separate from DSpace (which is required, as you can no longer install Solr as a "webapp")
This relates to this PR: https://github.com/DSpace/DSpace/pull/2058
And the discussion started by @mwood at: https://groups.google.com/forum/#!topic/dspace-devel/XkYgGgVyGhs

Mark Wood [9:16 AM]
Slightly related, since 2058 is the client side. But we need both upgraded.

Tim Donohue [9:18 AM]
Well, they really *are* related, as we cannot really run a Solr v7 client against a Solr v4 backend (that's not really a recommended setup). But, yes, there is a discussion here of client vs server
@mwood: would you like to kick off this discussion based on your dspace-devel comments?

Mark Wood [9:18 AM]
[Re-reading my comments]

So, how to determine the range of Solr service versions that we support?

Terry Brady [9:20 AM]
If I have half a dozen shards, how will they be ported into this new service? How will shards be managed?

Mark Wood [9:20 AM]
I took a quick look at a CentOS 7 system and it appears that Solr is not in the Red Hat package system, so we aren't tied down by that group of distro.s.

Tim Donohue [9:20 AM]
Well, there's a couple points there... How do we tell folks to install/configure Solr, what version? Do we give them advice on upgrading? and generally how do we configure/manage our Solr schemas with an external Solr. There's lots to discuss :wink:
I admit, I haven't done as much digging here on how to get us from "embedded, controlled Solr v4 webapp" to an "externally managed Solr v7 service"...but, I'd like to start discussing options

Terry Brady [9:21 AM]
Will we need to explicitly stop/start solr as a separate service?

Tim Donohue [9:22 AM]
@terrywbrady: yes. Solr (as of version 5) is no longer distributed as a webapp. It has it's own `solr` commandline tools to start/stop now

Mark Wood [9:22 AM]
Yes, Solr 5+ will be a separate system service. I had little trouble writing a System V service script for it.

Tim Donohue [9:23 AM]
So, literally, part of installing DSpace 7 will be... *You must install Solr first*, set it up (to be determined how), and ensure it is running so that DSpace can communicate with it

Terry Brady [9:24 AM]
Will we need to introduce any wait logic to the DSpace app to wait for Solr startup to complete, or will the Solr client library manage that for us?

Mark Wood [9:24 AM]
I think we should not try to give all-encompassing advice about setting up Solr. At most the simplest possible example: "this works, but ask the Solr community if you have special needs (or think you might)."

Tim Donohue [9:25 AM]
@terrywbrady: Good question, and I don't know (yet). We don't yet have DSpace fully running with a Solr v7 server. I think we'll want to see how DSpace behaves in this scenario and determine whether to throw a better error message or not.

Terry Brady [9:26 AM]
Since our tools have encouraged sharding in the past, we will need to provide good migration guidance for folks who have chosen to do that.

Tim Donohue [9:26 AM]
@mwood: I think my main question there though is how do we ensure your Solr (that you setup) has all the necessary cores configured & has loaded our custom schemas. There will be *some advice/instructions we have to give*. We also need to help people upgrade their current cores to the latest version

Terry Brady [9:26 AM]
I wonder if our command line tools are still appropriate, or if the upgraded Solr offers some equivalent functionality.

Mark Wood [9:27 AM]
Our tools for reindexing will, I think, have to be kept.
The Collection API offers some nice features that we might use for managing shards, instead of doing all the work ourselves.

Terry Brady [9:28 AM]
That makes sense. I was thinking of the import/export tools which might have some out of the box equivalents.

Tim Donohue [9:28 AM]
I wonder if we should start investigating each of these things in a wiki page... e.g. list out all our current Solr tools, and determine which must be kept versus which to replace.

Mark Wood [9:28 AM]
Actually 4.10 already has snapshotting. I used it recently. The documentation is woefully incomplete....
Wiki page is a good idea.

Tim Donohue [9:29 AM]
Plus, that'd help use figure out both an "fresh_install" setup for Solr and an upgrade process, from Solr v4 -> v7 (or whatever we require)

Mark Wood [9:29 AM]
There's two ways to handle a fresh-install: copy cores to the Solr home directory tree, or use the admin. API to hand them over.

Tim Donohue [9:30 AM]
Is this an effort (starting wiki page & brainstorms/links to resources) either of you would like to lead (or co-lead)? I'll gladly contribute, but I admit I feel like I only get to look at this off & on.
@mwood: I'd like to understand those two "fresh_install" options more, to be honest (which is where a wiki page might come in and describe them in more detail, with resources)

Terry Brady [9:31 AM]
I'll help. I think @mwood would need to be the leader on this since he has done more work.

Mark Wood [9:31 AM]
We may want to think a bit about sharding and whether we want to continue to (ab)use it as we do. More recent versions have some other ways to spread stuff out by timestamps that I've just started to study.
OK, I'll make a note to start a page.

Tim Donohue [9:32 AM]
I think we first need to figure out fresh_install & upgrade :wink: Sharding *is* important, but I'm hoping if we figure out how to upgrade/migrate to new Solr, that will also help us better understand/test how sharding works there

Terry Brady [9:32 AM]
If we can make shards invisible for folks, I think that is a good thing.

Mark Wood [9:33 AM]
The least chancy method of upgrading is to reindex, they say. With some of our cores that may not be practical. I don't yet have a good grasp of all the ways we use Solr in DSpace.

Terry Brady [9:34 AM]
We can't re-index the statistics

Mark Wood [9:35 AM]
Statistics can be re-loaded *if* you are treating the statistics core as a cache and still have the logs (or loadable extracts).

Tim Donohue [9:35 AM]
We can actually reindex statistics. We have a commandline tool that does that (by exporting statistics and reimporting): https://github.com/DSpace/DSpace/blob/master/dspace/config/launcher.xml#L237

DSpaceSlackBot (IRC) APP [9:35 AM]
renilgh has quit the IRC channel

Mark Wood [9:35 AM]
But we *may* be able to just upgrade in place. We should test this.

Tim Donohue [9:36 AM]
I'd say again we start small... Let's see if we can figure out how to do the "search" core first. Then "statistics" can come next (hopefully based on what we learn from "search")

Terry Brady [9:36 AM]
I could argue terminology with you. For the search repo, we can rebuild from the db/bitstreams. The problem is different for stats.

Mark Wood [9:36 AM]
We may run afoul of stuff that has been deprecated, replaced and removed over three major updates. Some of the field types that we use have been replaced at least twice.
Yes, search is a good place to start. It is most like what Solr was invented for.

Tim Donohue [9:38 AM]
@mwood: Yes, I think we also may want to consider whether we "recreate" our schema based on a current version. I don't believe we created that many custom fields...and if we can identify those, we might be able to just recreate them in a modern Solr Schema (instead of doing the diff of our old schema versus the latest)
I know that could be a big "if", but we should be able to search the code to see which fields are actually *used* in DSpace code

Terry Brady [9:38 AM]
I have a 7x Jira ticket to consider changing field types for better search stemming. We should link that to this effort.

Mark Wood [9:39 AM]
I'm sure that we have some type definitions that aren't used in DSpace. There is a lot of gunk in our schemas that was just carried over from the stock examples.
Yes, @terrywbrady, this is a good opportunity to take up some improvements in our schema design.
My point is that we may have to make some of those changes anyway.

Tim Donohue [9:41 AM]
@mwood: for what it's worth, I also have noticed that in latest Solr, you don't even need to manage the `schema.xml` directly. It's now a `managed-schema` file that can be modified via the Schema API: https://lucene.apache.org/solr/guide/7_0/solr-configuration-files.html#configuration-files

So, we could consider whether it's easier to manage our schema via the new API, or if we'd rather stick with the "distribute our schema as a schema.xml".

Terry Brady [9:41 AM]
https://jira.duraspace.org/browse/DS-3691

Tim Donohue [9:41 AM]
I admit, I haven't looked into this Schema API in any great detail... so if we find it's too much work, we can just use the `schema.xml`. But, I wanted to note its existence
(And the Schema API has a big note about "Why is hand editing the managed schema discouraged?" to consider here: https://lucene.apache.org/solr/guide/7_0/schema-api.html#schema-api)

Mark Wood [9:42 AM]
I will have to take a look at that.

Tim Donohue [9:43 AM]
Sounds good. Again, I just stumbled on this yesterday as I was preparing for this discussion and reading up on Solr 7. So, I don't know much more than what I've said (so I have a lot to learn here too) :wink:
But, again, this seems like useful notes to gather on the Wiki page, so we can learn together & brainstorm solutions/options

Terry Brady [9:44 AM]
This seems like a place where I could help the 7x effort.

Tim Donohue [9:45 AM]
@terrywbrady: that'd be wonderful. I think both of you know more than I about Solr (in general). I'm glad to give advice here, but I'd love it if you two could "run with this" and see what makes sense for DSpace 7

Mark Wood [9:45 AM]
I did want to get people to think about whether we want to require SolrCloud mode. It is necessary if we want to use the Collections API. That may not be a strong enough reason. But a single-instance "cloud" is simple: add a no-value option to the startup command.

Terry Brady [9:46 AM]
Once we implement this change, we will no longer be able to run 6x and 7x with the same data. Currently, I will swap the code b/w 6 and 7 without altering the data stores.
What are the implications of SolrCloud mode?

Tim Donohue [9:47 AM]
@mwood: I think we should consider SolrCloud mode, but I need to learn more about how hard it is to install, manage/upgrade. I keep reading notes that say/imply that "It may be overkill for some setups". So, we need to understand whether it's overkill for *most DSpace users* or not

Mark Wood [9:48 AM]
A SolrCloud has one or more Solr instances, and also one or more instances of Apache ZooKeeper which orchestrates them. The add-an-option approach runs ZK internally. If you later want to scale out, you can set up external ZK instances and reconfigure the startup scripts to use them.
Yeah, it may be more than we want to take on, just to get the Collections API.

Terry Brady [9:49 AM]
What does "collections api" mean here?

Tim Donohue [9:50 AM]
@mwood: it is a good question to ask, as we might want to at least come up with the answer to the question: Can I run DSpace 7 with SolrCloud mode? (I'm sure some institutions might be interested in that, if they use

SolrCloud heavily elsewhere... I'm just not sure whether requiring SolrCloud is necessary or not)

Mark Wood [9:50 AM]
Collections is attractive to me because it would allow us to scrap our custom dump/restore code and just send a SPLITSHARD operation, watch the status, and follow with a DELETESHARD to clean up the unsplit copy.
Collections API is used to manage shards, replicas, etc.

Terry Brady [9:51 AM]
It sounds like it implements the cloud scaling

Mark Wood [9:52 AM]
I might mention that current documentation still describes "legacy" shard management, in a tone of "you can still use this if you want, but we're not paying it much attention anymore."

Tim Donohue [9:52 AM]
It sounds like we need more info here in general. It's attractive to consider, but we need to understand whether it makes simple setup/upgrades harder or not? I.e. what are the downsides to this approach for users of DSpace who just want a simple, small repository

Mark Wood [9:53 AM]
We do need more info., but I suspect the answer is that if you want a simple, small repository you just start Solr with this option so DSpace can talk to it and that's that.

Terry Brady [9:53 AM]
And how difficult would it be to migrate to SolrCloud after your simple instance becomes more complex

Tim Donohue [9:55 AM]
If we find SolrCloud is super easy to setup/upgrade, then yes, let's just require that mode. I just don't have clarity into whether there are downsides to SolrCloud that we are missing (i.e. why does this legacy mode still exist, is it just for backwards compatibility, or is there something that is difficult in SolrCloud)
So, if we just do some fact finding/testing of SolrCloud mode, and find it's easy enough, then that's good enough for me.

Mark Wood [9:56 AM]
Cloud mode does have more moving parts. You need a ZK instance that wasn't there before. There isn't a lot of advice about running it internally in a single-instance "cloud". What I've seen assumes that you will be running multiple instances in production.
I don't think that's *required* but we may be a bit lonely if we run single-instance in production.
ZK can take charge of the Solr configuration files. One of the things it does is handle configuration distribution.

Tim Donohue [9:58 AM]
So, we're hitting up against our time here (nearly top of the hour). So, we probably need to wrap-up with "next steps"
It sounds like the major next step is to start a Wiki page on DSpace 7 Solr Service discussions/consideration.
@mwood will you start that up for us?

Mark Wood [9:59 AM]
I'm just glad to know whether this is "think some more" or "don't bother, we aren't going that way."
I will set up the page.

Tim Donohue [10:00 AM]
Thanks! I think the other thing is that this should probably be an ongoing discussion in this meeting. So, once the page is created, we can add it as an ongoing topic to this agenda

Mark Wood [10:00 AM]
OK, thanks.

Tim Donohue [10:01 AM]
Obviously, we can also make the DSpace 7 WG aware of this effort, but I think this seems like a topic for this DevMtg (as the DSpace 7 WG are more concerned with REST API & Angular)
Any other final notes/topics before we wrap up for today? I think this has been a good discussion on Solr, and I hope to learn more as we dig deeper on these brainstorms, etc

Terry Brady [10:03 AM]
I suspect the migration to this set up will be disruptive to a dev environment, so a fair amount of coordination will be needed with the 7x team.

Mark Wood [10:03 AM]
I updated a number of PRs for merge conflicts against the Log4J and Configuration patches. If anyone is interested in reviewing, that would be helpful.

Tim Donohue [10:04 AM]
@terrywbrady: yes, definitely, once we are ready for that disruption (i.e. have a plan and PR ready), we'll need to closely coordinate with the DSpace 7 team
@mwood: I'd recommend linking those PRs into #dev (or here), just as a reminder. I'll see if I can get to them myself

Mark Wood [10:05 AM]
OK. They're things that have sat around for a while -- that's how they got clobbered by other patches.

Tim Donohue [10:06 AM]
Ok, sounds good
Let's wrap this up for today though. Thanks for the great discussion today, and looking forward to more Solr discussion in future meetings! Thanks all!

Mark Wood [10:07 AM]
Yes, thanks all.

Terry Brady [10:07 AM]
have a good week