DevMtg 2018-12-12

Developers Meeting on Weds, December 12, 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 is tomorrow, Thursday, Dec 13 at 15:00 UTC.
- DSpace 7 Entities Working Group (2018-19): Next meeting on Tues, Dec 18 at 16:00 UTC
    - Last meeting notes/video at 2018-12-04 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 (Any status updates?)
    a. PR https://github.com/DSpace/DSpace/pull/2058
4. DSpace Docker and Cloud Deployment Goals (old) (Terrence W Brady)
5. 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 Mark H. Wood)
        i. PR 2180 improves reporting. Ready for review.
6. 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+2018-12-12
```

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

Pascal Becker [9:00 AM]
hi!

Mark Wood [9:00 AM]
Hello

Alexander Sulfrian [9:00 AM]
hi

Terry Brady [9:01 AM]
hello

Tim Donohue [9:01 AM]
Hi all, welcome. We'll go ahead and dive into quick updates (while others join), and take it from there
On DSpace 7 side of things, the only update of note is that we've begun planning which features will be likely in the *Preview* release (late Jan / early Feb) and which will wait for *Beta* (April). This info can be found in the first column of our dev planning spreadsheet
https://docs.google.com/spreadsheets/d/18brPF7cZy_UKyj97Ta44UJg5Z8OwJGi7PLoPJVz-g3g/edit#gid=0
Other than that, feel free to drop in our DSpace 7 meeting tomorrow (15UTC) if you want to get more development updates
On the DSpace 6 side of things, there's no updates to note (all effort/concentration is on DSpace 7 at this time, though folks are always welcome to submit/review/merge PRs to prep for the next 6.x release)
So, that skips us past the first two topics, unless there's any questions?

DSpaceSlackBot (IRC) APP [9:04 AM]
CorneliusL[SULB] has quit the IRC channel

James Creel [9:04 AM]
Hey all, jumping in now. But I'll have to leave at the half-hour.

Pascal Becker [9:05 AM]
Do we already have concrete plans for DSpace 6.4? (edited)

Tim Donohue [9:06 AM]
@pbecker: nope, as noted on the agenda, I'm sure it will happen at some point. No concrete plans at this time though. My best guess at this point is it may not be until 7.0 dev starts ramping down, or until there's a major reason to release 6.4, whichever comes first
(Or more simply put, it won't happen until someone has a chance to concentrate on making it happen)

Terry Brady [9:08 AM]
I am working on an upgrade to 6.x. There are a handful of 6.4 changes that I am having to integrate into my code in order to release successfully.

Tim Donohue [9:09 AM]
If any of those are *unmerged*, @terrywbrady, then it'd obviously be good to report back and get those fixes /changes merged into 6.x (at a minimum)

Mark Wood [9:09 AM]
One is always free to argue that we *have* a major reason (or several) to release 6.4.

Pascal Becker [9:09 AM]
I found two small bugs regarding DOIs yesterday. I will file tickets an PRs as soon as I can. Would be happy to see those getting into 6.4 whenever that's going to be released.

Terry Brady [9:10 AM]
Will do. Fortunately, the things I need have been merged. I will re-review that list to confirm.

Tim Donohue [9:11 AM]
Yes, and as @mwood notes, if there is a major reason that I'm not aware of to push out a 6.4 soon, then it'd be good to talk that through. I think we'd need to try to avoid taking anyone off of major 7.0 efforts though (where possible), but it's always possible to do a quick release (just release what we've merged now)
(Not saying we need to talk that through now...but, just realize you can always bring reasons to these meetings & we can dig in deeper to analyze whether we feel 6.4 is warranted)

Terry Brady [9:12 AM]
I think it is wise to wait for more to accumulate.

Tim Donohue [9:13 AM]
Ok, moving along then. I wanted to touch base quickly first on Solr Server upgrades (for DSpace 7):

https://wiki.duraspace.org/display/DSPACE/Upgrading+Solr+Server+for+DSpace
Any updates to share here, @mwood?

Mark Wood [9:13 AM]
I'm still trying to understand why that one test fails.

Tim Donohue [9:14 AM]
ugh, that's definitely not good news (that one test is so so stubborn). :disappointed:
Any decent leads yet? Or is it mostly a mystery?

Terry Brady [9:14 AM]
Here is the state of my work on the issue:

https://github.com/DSpace-Labs/DSpace-Docker-Images/pull/65/files

Mark Wood [9:15 AM]
Mystery. The query in that test should match only one document, but it gets back three.
I added a method to dump the entire content of the core when asked. It shows exactly the records I would expect, and no others.

Tim Donohue [9:17 AM]
Are you able to load that core elsewhere and test the exact query against it (to see if it returns 3 as well when run manually)?

Terry Brady [9:17 AM]
What is the query?

Mark Wood [9:17 AM]
I haven't been able to make a useful snapshot of the core. After the test exits, the core is empty.
Hunting for the query...

Tim Donohue [9:19 AM]
This is the failing test: https://github.com/DSpace/DSpace/blob/master/dspace-spring-rest/src/test/java/org/dspace/app/rest/DiscoveryRestControllerIT.java#L2683
And the query is here: https://github.com/DSpace/DSpace/blob/master/dspace-spring-rest/src/test/java/org/dspace/app/rest/DiscoveryRestControllerIT.java#L2735-L2736
It's trying to search by dc.date.issued. Earlier in that method, it adds 3 Items with different dates, and tries to find just *one* of them.
But, instead the result is always *3* (presumably matching all 3 items, which is odd)

Terry Brady [9:22 AM]
Thanks for sharing that. It does not look like any of the dynamic fields would be introducing something unexpected.

Tim Donohue [9:22 AM]
I can add those notes to the PR for reference to others. It's a total pain as (besides this single date query) the Solr Upgrade PR is "ready to go". (edited)

Mark Wood [9:24 AM]
The dump shows exactly three documents that even *contain* that field, and only one of them has the requested value.

Terry Brady [9:25 AM]
Regarding the server side...

In the PR I shared above, I am able to demonstrate how to use the solr api to build repos and set the schemas.
The schema work in that PR is incomplete. I migrated the easy fields. For the trickier fields, I have currently set them to something generic like "text" or "string". The README will convey the per-field status.

Tim Donohue [9:26 AM]
@mwood: I can try and lend some more eyes on this Solr issue once I finish up my REST API v7 Integration Test fixes (I've figured out that the way we run ITs is not ideal, as we reload all beans over and over for every test class. I nearly have it fixed up, hopefully a PR later today)

Terry Brady [9:26 AM]
Is is possible that the date string is being treated at 3 keywords that are being OR-ed together?

Mark Wood [9:28 AM]
Hm. The field is defined as string. "query" : "dc.date.issued:2010-02-13" is returned in the response.

Tim Donohue [9:28 AM]

possibly. I'm not sure we know what could be going on. But, it seems likely that either the date is being *stored* unexpectedly in Solr, or the query itself is an incorrect syntax in Solr v7 (so it returns everything instead of one thing).

James Creel [9:29 AM]
got to jet, gang
see y'all

Terry Brady [9:29 AM]
@jcreel256 have a good week

Tim Donohue [9:30 AM]
It might be interesting to see if we can get a Solr v7 setup, index a few documents with "dc.date.issued" fields, and try and run this query manually. I know that's a lot of work, but it seems like we need to *see* what is going on here.

Mark Wood [9:31 AM]
I agree: it would be good to just play with some documents in Solr 7. I may just make up some documents and try that. I have Solr 7.2.1 running (and I could upgrade to 7.3 without much trouble).
I'll do that.

Tim Donohue [9:33 AM]
@mwood: FWIW, just after I typed that, it reminded me also that we've "played with" query escaping in the past...most recently here: https://github.com/DSpace/DSpace/pull/2206 (merged in late Oct). Not at all sure this is related, but I wonder if either the periods in "dc.date.issued" are problematic or the colon (:).... but, it's possible this is unrelated if other fields work fine.
In any case, @mow
In any case, @mwood your approach sounds good.
It doesn't sound like we can solve this today, but we can keep digging & reporting back on #dev and on the PR itself

Mark Wood [9:34 AM]
Will do.

Tim Donohue [9:35 AM]
Moving along then.. @terrywbrady I know you wanted to briefly discuss Docker goals??   I also know that @pbecker would like to talk about Deletion of EPersons (for 7.x & GDPR compliance). Is there a particular order we want to discuss these in?

Pascal Becker [9:36 AM]
I have the next meeting starting at the top of the hour.
If there is a meeting next week, it will be a late one so hard for me to join.
@terrywbrady may I start? I'll try to keep it short. (edited)

Terry Brady [9:37 AM]
Go for it. Please reserve some time for me.

Pascal Becker [9:37 AM]
DSPR#2229 got reviewed and I handled all comments. Thanks again to @terrywbrady and @mwood for looking into the PR.
It is basically at +2 and could be merged.

Tim Donohue [9:37 AM]
https://github.com/DSpace/DSpace/pull/2229

Pascal Becker [9:38 AM]
I wanted to bring this up here, because I expect that we may run into some NPEs. I expect that we missed some places in the code that would need further changes.

Terry Brady [9:38 AM]
Fyi, the OR2019 submission date is in early Jan. Will we want to coordinate any presentations? If so, we might need to start that discussion.

Pascal Becker [9:39 AM]
Basically everywhere were an EPerson is referenced we might get null back instead.
I think that is the way to go and we discussed this here before.
It will be discussed in DSpace 7 meeting tomorrow as well (I won't make it there, sorry).
Here and now I wanted to ask again if there are any objections against this path or if we can proceed and merge this PR?
Further work would than needed to be done on the new REST-API and the Angular, and we (TU Berlin) are currently not able to do this work, so I hope for help of others. (edited)

Mark Wood [9:40 AM]
The sooner we merge it, the sooner we have a chance to discover any problems that were missed.

Tim Donohue [9:40 AM]
I have no objections & I agree with @mwood (was about to type the same thing)

Terry Brady [9:41 AM]
My approval is in place on the pr

Tim Donohue [9:41 AM]
I suspect the DSpace 7 team will also not have objections, but we'll talk with them tomorrow -- as they'll need to be aware that it's possible things may "break" in REST or Angular after merger (but, hopefully, nothing does)

Pascal Becker [9:42 AM]
If they are positive too, this could be merged as of tomorrow.

Tim Donohue [9:42 AM]
agreed. I can report back on the PR regarding the discussion (and merge if all are favorable)

Pascal Becker [9:42 AM]
great, thank you!

Tim Donohue [9:43 AM]
Thank you as well, @pbecker for creating this important change

Mark Wood [9:43 AM]
The last two (unreviewed) commits look well to me.

Pascal Becker [9:43 AM]
That was it already. I think we can move on, unless someone raises questions about this.

Tim Donohue [9:44 AM]
Sounds good. Let's move along then. @terrywbrady you also brought up OR2019 (above). I have that on the DSpace 7 team agenda for tomorrow, but was there more we wanted to discuss here?
I'll admit, my tentative plans are similar to OR2018 -- training/workshop on DSpace 7 and a DSpace 7 update /release announcement (fingers crossed) talk

Terry Brady [9:45 AM]
What presentations will be coordinated by the project? Will DSpace 7 take up the whole pre-conference day again?

Pascal Becker [9:46 AM]
I'm thinking about giving a talk about national User Groups, their relation to their international community, why user needs local user groups and how to build those.

Terry Brady [9:46 AM]
I could see some usefulness in a DSpace Docker workshop if there is not a competing presentation.

Tim Donohue [9:46 AM]
@pbecker: that'd be wonderful. I bet Mic would help you with that, if you needed help (though he doesn't return from his leave until after the OR2019 deadline)

Pascal Becker [9:47 AM]
@tdonohue I was thinking if I should give that talk together with Mic.
@terrywbrady If you need a second pair of hands for a docker workshop, I would volunteer. (edited)

Terry Brady [9:48 AM]
Thanks @pbecker. I had you in mind.

Tim Donohue [9:48 AM]
@terrywbrady: a DSpace Docker Workshop would be great. I don't have an exact plan for what to do for DSpace 7, but we could talk more tomorrow in the DSpace 7 meeting. It's possible we could just do a general "Customizing DSpace 7" workshop (oriented more towards UI)

Pascal Becker [9:48 AM]
:slightly_smiling_face:
@tdonohue I would hope for a "How to update to DSpace 7" workshop.

Terry Brady [9:49 AM]
@tdonohue that sounds good. @pbecker, I will start a conversation with you about a workshop.

Pascal Becker [9:49 AM]
Maybe even a "How to migrate to DSpace 7" workshop. :wink:

Terry Brady [9:50 AM]
Thanks for the OR chat. i did want to move to https://wiki.duraspace.org/display/~terrywbrady/DSpace+Docker+and+Cloud+Deployment+Goals

Mark Wood [9:50 AM]
"Customizing DSpace 7" sounds like a good idea. The Angular team have been building quite a lot of UI code -- it would be good to have some guidance on how to find what you want to change.

Tim Donohue [9:50 AM]
@pbecker: right, the biggest part of that is customizing DSpace 7 UI though. So, I think we're talking about similar ideas, but your wording around "migration/upgrade" may be a better one
We can loop back on OR2019 again next week. But it's good we got a few ideas out there -- maybe next week we can narrow things down a bit more & start to finalize the main talks we want to collaborate on

Pascal Becker [9:51 AM]
@tdonohue happy to hear that. I had in mind that there are changes in the configuration too.

Mark Wood [9:52 AM]
Yes, the two top questions about DSpace 7 (after "when?") are likely to be "what does it have for me?" and "how do I get there from here?"

Tim Donohue [9:52 AM]
@terrywbrady: did you want to talk about Docker now quickly (and we can table additional OR2019 discussions for next week)

Terry Brady [9:52 AM]
This document was my attempt to explain to my work on DSpace + Docker to my colleagues. I realized that it would also be a good way to check in with you all to see if these goals make sense.

I recommend chatting about each of the numbered goals in the document and then looking at the comments.
(thanks @tdonohue... let's move through the doc)
The "Goal 1" is essentially the work that we have been doing this year. The text explains the reasoning for that work. Are there any surprises or reactions to that text?
I'll move to the next one. Goal 2 - create some standard AIP files to be shared
What do you all think of this goal? What do you think of the proposed distribution options?

Tim Donohue [9:57 AM]
I like the goal in #2, as it is something we've talked about for some time (essentially having a "test corpus" of sorts both for demos & even for testing/testathons, etc).   Where to store them & how/where to gather them is the big question for me

Mark Wood [9:58 AM]
And if in a for-pay option, who pays?

Terry Brady [9:58 AM]
It is amusing that we work on a repository platform... and here we are discussing sharing assets.

Tim Donohue [9:58 AM]
It'd have to be run by DSpace Steering and funded out of the DSpace budget....unless some institution stepped forward to host these on the project's behalf.

Terry Brady [9:59 AM]
If we had a consensus on a particular mechanism, I would be happy to put a proposal together to describe what we need.

Mark Wood [9:59 AM]
We need to gather the corpus and see how much storage we'd need.

Terry Brady [9:59 AM]
I suspect that the cost is trivial if we agree on a good mechanism.

Tim Donohue [9:59 AM]
@terrywbrady: amusing yes, but a part of this challenge is the knowledge (that we all have) that copyright is hard/complex. Finding a decent test corpus means finding (or creating) content that is copyright free

Pascal Becker [9:59 AM]
I'm sorry, I have to run to the next meeting. Good bye!

Terry Brady [10:00 AM]
I also think that if we had a place to place assets, it would be much easier to assemble the corpus.
IIRC, Tim's demo.dspace.org assets are about 1.5 G which is bigger than the size permitted for a typical github repo.

The GitHub LFS might be attractive to the developer community since it should play nicely with the code assets.

Tim Donohue [10:02 AM]
We do have to realize though that "assembling a corpus" will require some care / "double checking". If we let any content in (without ensuring a lack of copyright restrictions), we could risk the project being sued for copyright infringement.

Terry Brady [10:02 AM]
(In this doc I have shared Goals 3, 5, and 5 build on the first 2 goals).

Tim Donohue [10:03 AM]
So, I honestly feel the biggest challenge here is copyright...and ensuring we can find copyright free materials that we can legally redistribute in this fashion.

DSpaceSlackBot (IRC) APP [10:03 AM]
mhwood has quit the IRC channel

Terry Brady [10:03 AM]
How did you manage that for the demo site?

Tim Donohue [10:04 AM]
The demo site content started out self created (by me). It's "dummy content" (not at all realistic). I think Bram Luyten added to it with some content from a sibling of his.
So, we've gotten around it by keeping a tight control over what is in the demo site AIPs

Terry Brady [10:05 AM]
I would be happy to help with the technical side of this effort. I presume we could ask DCAT for some copyright expertise help.

Of the options I listed, do you have a preference for one of the distribution methods?

Tim Donohue [10:07 AM]
I'd likely prefer whichever is the least expensive for redistribution. For example, while S3 is low cost for storage, you are charged *per download*. So, I think S3 is potentially risky if this content ends up widely used (as costs will increase as downloads increase).
But, I don't know where the line is with S3...would need to do more calculations on whether we'd only hit issues at thousands or millions of downloads (and it may depend on the size of the test corpus too)
(Longer term) Another option we could consider is whether there'd be an easy way for DSpace to *harvest* test content from another location... this would require new code, but could make gather test data simple if you could pull it down from PubMed Central, or similar
is realizing we are well over time here...and I think others have dropped off, it may just be the two of us at this point

Terry Brady [10:10 AM]
Does this feel like a worthwhile goal to you? As I have spun up Docker instances of DSpace, I realize that I need data. Rather than solving the issue just for myself, I would like to find a way to solve it for the project.

Tim Donohue [10:11 AM]
I think the goal of getting test data loaded into Docker (or any test instance of DSpace) is very worthwhile. I'm just not sure there's a "silver bullet" that easily solves the problem.

Terry Brady [10:12 AM]
We may do well to just pick a solution and grow into it. If the costs become a challenge, then we reconsider the option.

Tim Donohue [10:13 AM]
And honestly, longer term, I wonder if we should consider building a way to pull in test data from another source (with a decent API), rather than attempting to store all the test data ourselves.

Terry Brady [10:13 AM]
I suspect that everything would be easier if we keep the assets relatively small.
We would likely still need AIP's to create a hierarchy and to set permissions.

Tim Donohue [10:15 AM]

Agreed, I just worry that costs that aren't "stable" are risky for the project (and Steering will see them as risky), if we don't carefully plan / monitor those costs. So, this is an area where S3 storage could be risky if the public can download from it freely.
Keeping the test corpus small does help though...and if we could find a place that doesn't charge per download, even better

Alexander Sulfrian [10:17 AM]
I think it should be possible to find a "mirror" (ftp server or something like that) at an institution, that would be hosting the testdata.

Tim Donohue [10:19 AM]
This has been a good discussion, and it's definitely something we need to revisit. I suspect we need to wrap up this meeting though (as we're now ~20 mins over)

Terry Brady [10:19 AM]
Thanks for the time on this. Could we keep this floating on the agenda?

Tim Donohue [10:20 AM]
Sure, I can keep it on the agenda...some topics may jump it (like OR2019 next week), but I'll keep it on there
Thanks all for the discussion, we'll consider the meeting adjourned. Talk to you all next week, if not sooner (and by the way, next week's meeting on Weds, Dec 19 will be the last of 2018)

Terry Brady [10:20 AM]
It seems like a very solvable problem. We have agreed there is a need. I have offered to help organize the effort.

Mark Wood [10:26 AM]
Sorry that I dropped off. I was reminded why I don't rebuild Chromium during the daytime: load average shot up over 100 and nothing would move.