DCAT/Committer Meeting Notes - DSpace Futures (Nov 2012)

This page details the "DSpace Futures" discussions that took place between DCAT members & DSpace Committers on November 20 and 28, 2012. These two meetings were held to provide DCAT & Committers an early summary of the feedback that came out of DuraSpace's "DSpace Futures" discussions with DSpace sponsors & service providers. A public report is also being drafted by DuraSpace to report the results of all these meetings back to the entire DSpace Community.

- First Meeting - November 20, 2012
  - Attendees
  - Discussion Notes
- Second Meeting - November 28, 2012
  - Attendees
  - Discussion Notes

First Meeting - November 20, 2012

Attendees

Amy Lana - University of Missouri

Bram Luyten - @mire

Ciarán Walsh - Enovation Solutions, Ltd

Elena Feinstein - University of North Carolina at Chapel Hill

Elin Stangeland - Cambridge University Library

Jim Ottaviani - University of Michigan

Jonathan Markow - DuraSpace

Maureen Walsh - The Ohio State University

Sarah Shreeves - University of Illinois at Urbana-Champaign

Stuart Lewis - University of Edinburgh

Tim Donohue - DuraSpace

Valorie Hollister - DuraSpace

Sarah Potvin - Texas A&M

Discussion Notes

Intro - Valorie Hollister

- Purpose of call is to bring people up to speed on future of DSpace discussions
- Three calls with Sponsors -- several of you were participants
- Will share a summary on the feedback we received so far
- Jonathan Markow will first start with an overview of recent thoughts around DSpace Futures & Fedora Futures
- Val will then talk about DuraSpace sponsor feedback
- Will then open up the floor for feedback from all of you, discussion, questions/comments

Overview of DSpace and Fedora Futures - Jonathan Markow

- began talking with DSpace and Fedora communities about future of the platforms - because of feedback we've gotten from users - would like improvements to now 10 year old software
- had some discussion at OR - generated some ideas for how to get more energy into the open source development process
- because current pool of volunteer committers only have time to make incremental improvements in software, may want to consider another way of getting more ambitious/bigger projects done
- Fedora community has identified a need to make major changes to Fedora and interested institutions have pooled together resources to spin up a project and get it going
- sense is that the DSpace project has more diversity of needs among users - many users are satisfied with DSpace, but others who want to push the envelope around research data mgmt, digital collections - willing to trade easier out of the box version for more complex one
- just completed a series of calls with DSpace sponsor

DuraSpace Sponsor feedback so far - Valorie Hollister

- What Sponsors LIKE about DSpace:
  - meets needs for an institutional repository
  - allows for open access to content
  - useful workflows
  - lots of excitement about features coming in 3.0 and features available in 1.8
- What are some of the CHALLENGES Sponsors face
  - digital preservation - some hooks in DSpace for digital preservation, but functionality to serve as a preservation tool has a ways to go
  - front end / tools for Admins / UI improvements - customization needs to be easier and more flexible/modern (i.e., customize by collection) and more accessible for Admins
    - UI is currently difficult to customize
  - multimedia needs - multimedia streaming / downloading video/images/audio files
  - digital asset mgmt strategy - several institutions are in the middle of a review of their overall strategy, particularly as it relates to research data management
  - data management - many institutions mentioned the need to store data and present it in different ways, also the ability to get data in and out of DSpace easily
  - author identity / author profile - need to allow content to be showcased in different forms (i.e. BibApp, VIVO) - don't want to re-create metadata or content
  - backend improvements - a few institutions mentioned the need for a unified backend
  - open/closed content - a few institutions mentioned a need for more flexibility in dealing with open / closed content in order to deal with copyright issues
- What are Sponsors INTERESTED in:
  - DSpace/Hydra - a lot of curiosity and interest in the potential of creating a DSpace Hydra head, Univ of Hull has a IR Hydra head, several institutions would like to have one solution for all content (student/faculty, digital collections, data, etc.), some concern about re-creating the DSpace workflows - critical that DSpace maintain its core functionality
  - DSpace/Islandora - a few institutions interested in Islandora - which is a Drupal frontend with a Fedora backend, some believe it might be easier / more accessible to customize Drupal than Hydra and that Drupal would provide more flexibiilty for displaying digital assets
  - DSpace needs to have a "repository abstraction layer" - e.g. REST module (GSoC project) - give the ability to use any UIs (i.e. Drupal, Ruby on Rails based, etc) or more easily "hook up" other things to Dspace

Discussion/feedback from DCAT/Committers - ALL

- Bram: any comment/feedback on RDF/linked data? No, it didn't come up in the most recent discussions
  - Amy: we would be interested, but not at top of our list - preservation, streaming would come first
  - Elin also mentioned an interest but not at top of list
- Stuart: preservation important - DSpace has "hooks", but not not hanging anything off the hooks in DSpace.
  - e.g. there is a bitstream registry, but don't really use it, no reporting mechanism, need to extend that support
  - Also would be nice to hook into DROID / JHOVE / PRONOM
- Elin: preservation important - need to be to do things like identify files at risk, have the ability to export them out to something like DuraCloud for example to run migration jobs, and then easily import it back – linked back to original metadata
- Stuart: need to ask what we see the direction of DSpace? preservation, great interface, streaming, RDF imagining - is it possible to do everything? or should it be something modular?
  - Worries about the extent at which we seem to ask DSpace to do everything. Modularity may be the key.
- Jim: functionality doesn't have to be native to DSpace - like APIs - digital objects can be stored and preserved - but just clicking on a button to download might not be enough - need it to preserve and provide access to whatever people want to put in - I am selling DSpace as a place to put data at our institution
  - Dspace should "preserve and provide access to the stuff that people want to put in there"
- Bram: have sponsors mentioned interest in leveraging repositories as source for out-metrics (altmetrics?) and more advance statistics? Not specifically.
- Statistics discussion
  - Amy: looking forward to elastic search based statistics in 3.0
  - Tim: elastic search is just a different backend to stats (vs. SOLR), not additional reporting
  - Stuart: aggregating stats might help store more data, but we need to work out what are the reqmts for statistics
  - Sarah: interested in Google Analytics - stats is a pain point for us
  - Maureen: stats are important - using elastic search because of scalability issues, but doesn't solve all of what we need
  - Elana: stats are important - our needs are different than normal IR - but helps to justify our existence
  - Elin: stats are a basic need, so folks may forget to bring it up. Cambridge hasn't been getting statistics we need - looking for Google Analytics plugin to help
  - Bram: Google Analytics vs. existing stats in DSpace - tough tracking downloads - could be opportunities to registered w/Google Analytics, need to think about how stats could be used in the future
    - Points out that Google Analytics cannot show stats just for a specific Collection/Community – it's more a tool for site-wide stats
    - DSpace 3.0 adds internal stats for search queries & workflow events
    - Maybe there's a way to use both Google Analytics & internal stats?
- Maureen: scalability is a challenge - we moved to elastic search because of scalability - being able to handle the scale and have functions operate properly - basic performance issues, java error codes - not sure what the problem is - whether it is scale or something else - repository has 50,000 items and 100,000 files
- Stuart: do we need to look at general architecture changes when it comes to scalability - do we split end user functionality / admin functionality - a lot of checks in DSpace - is there a scalability advantage in splitting them?
  - We tend to do everything in one interface for Dspace. We may see scalability advantages if we split up the access (user) interface from the administrative interface.
- Tim: mentions the talk of a "repository abstraction layer" (REST API) - does that allow us to do what Stuart is talking about - create a read-only interface and a separate admin interface to manage content - could you use REST to plug DSpace into Hydra (similar to how REST is used to plug Fedora into Hydra)
- Stuart: makes more sense to build Hydra on top of DSpace - or Islandora on top of DSpace - keep DSpace workflows, but allow for new interfaces on top of DSpace
  - This is more of "Hydra on DSpace" rather than "DSpace on Hydra"
- Tim: this idea (Hydra on DSpace) could let you tailor interfaces towards diff types of content - sounds plausible, not sure if it is possible - step 1 would be to make sure DSpace has a repository abstraction layer (REST API)
- Tim: we actually have a few REST API "in the works". There was the original that was built as a GSofC project - had some scalability problems and access issue (everything open) - several project forks in Github:

    - Original GSoC REST project (not very scalable): https://github.com/DSpace/dspace-rest
    - Hedtek enhanced REST API: https://github.com/hedtek/dspace-rest (more info at: http://hedtek.com/2011/dspace-rest-api-testing/)
    - Wijiti enhanced REST API: https://github.com/wijiti/dspace-rest-api (more info at: https://jspace.atlassian.net/wiki/display/DSPACEAPI/DSpace+REST+API+Home)
    - Currently we just haven't had the developer/committer resources to do a thorough review of the two enhanced REST APIs. Would be nice to review each and either "merge them" or adopt one as the official REST API & work to keep it maintained, etc.

- Jonathan: managed projects idea - in the Fedora community institutions interested in getting the work done have started a project - agree on set of features that are critical to improving Fedora now, project participants have committed in-kind donations of dev time (mostly more time from current Fedora committers) as well as financial support of tech lead and project manager
- Jonathan: for DSpace we hear a lot of diff interests and ideas - if more ambitious ideas are going to take place, it will have to similar to Fedora, would like to hear from anyone who has specific projects that you would like to help with

Next Steps - Valorie, Jonathan, Tim

- Have one more DCAT/Committer call next Wednesday
- will summarize / make transparent what the feedback has been from all mtgs / discussions - likely posted onto the wiki and distributed via the mailing lists
- not sure what projects will come out of discussion - need to determine who has common interests
- want to include everyone who's interested in those projects
- DSpace community has to get the momentum - need more developer time / resources to work on projects - either new developers or more time from existing developers
    - Jonathan points out that on the Fedora side, most of the new Fedora project development is actually being done by existing committers (their institutions were willing to commit extra time to do this extra work on Fedora)
- DCAT's role will be similar to other development in DSpace - provide feedback on features, make recommendations based on different institutional perspectives
- Committers will help project answer questions and review code submissions

Second Meeting - November 28, 2012

Attendees

Ivan Masár (aka helix84)

Keith Gilbertson - Virgina Tech

Leonie Hayes & others - University of Auckland

Mark Diggory - @mire

Mark Wood - Indiana University-Purdue University Indianapolis

Iryna Kuchma - eIFL.net

Valorie Hollister - DuraSpace

Jonathan Markow - DuraSpace

Tim Donohue - DuraSpace

(A few others may have joined mid-call?)

Discussion Notes

Intro - Tim Donohue

- Purpose of call is to bring people up to speed on Future of DSpace discussions
- Three calls with Sponsors & one with DCAT/Committers. Discussion was general, e.g.
    - what do you like about Dspace
    - what do you need more of in DSpace
    - what other sorts of IR/digital preservation challenges do you face / Other software you use
- Will share a summary on the feedback we received so far
- Jonathan Markow will first start with an overview of recent thoughts around DSpace Futures & Fedora Futures
- Tim will then talk about DuraSpace sponsor feedback & summarize last DCAT/Committer discussion
- Will then open up the floor for feedback from all of you, discussion, questions/comments

Overview of DSpace and Fedora Futures - Jonathan Markow

- Discusses how things have evolved with Fedora's "Fedora futures" project -- key stakeholders wanting more rapid development
- Some talks on DSpace side -- DuraSpace has been contacted about DSpace on Fedora / DSpace on Hydra / DSpace on Islandora
- At OR 2012 conference in Edinburgh
    - talked about approaches other open source projects have used to bring forward platforms
    - e.g. managed projects
        - small changes happen all the time / episodic changes - This is how DSpace & Fedora currently develop
        - larger commitments/changes sometimes need more time/effort/funding (This is where "managed projects" come in)

            - managed projects - institutions donate additional time & money for developers to achieve some concrete community goals / larger project
                - Entire managed project is then a collaboration among many institutions/stakeholders, and managed by a "steering committee" of sorts
- For example, Fedora stakeholders came back and asked to do something "significant" with Fedora. It's 10 years old & needs some work
    - Fedora stakeholders/institutions decided to begin a grassroots "special project" to do a significant amount of work on Fedora
    - They have established their own steering committed & their project will be kicked off in December (at CNI)
- So, we're now beginning similar discussions with DSpace stakeholders - Sponsor calls, DSpace Committers & DCAT
    - DuraSpace is not expecting 100% agreement on one DSpace project to bring forward
    - But, hopes that we can find areas of need & perhaps get some managed projects or similar off the ground

DuraSpace Sponsor feedback so far - Tim Donohue

- What Sponsors LIKE about DSpace:
    - meets needs for an institutional repository
    - quick to get running & allows for open access to content
    - useful workflows
    - lots of excitement about features coming in 3.0 and features available in 1.8
- What are some of the CHALLENGES Sponsors face in DSpace:
    - digital preservation - some hooks in DSpace for digital preservation, but functionality to serve as a preservation tool has a ways to go
    - DSpace Admin UI tools is lacking
    - DSpace UI (XMLUI especially) is currently difficult to customize
    - multimedia needs - multimedia streaming / downloading video/images/audio files
    - open/closed content flexibility
        - a few institutions mentioned a need for more flexibility in dealing with open / closed (embargo) content in order to deal with copyright issues
- General institutional CHALLENGES (not necessarily DSpace specific...but related)
    - digital asset management strategy
        - several institutions are in the middle of a review of their overall strategy, particularly as it relates to research data management
    - data management
        - many institutions mentioned the need to store data and present it in different ways, also the ability to get data in and out of DSpace easily
    - author identity / author profile
        - need to allow content to be showcased in different forms (i.e. BibApp, VIVO) - don't want to re-create metadata or content
        - Would like ways to "hook" this into DSpace
    - backend improvements (unified)
        - a few institutions mentioned the need for a unified backend between Fedora & DSpace
        - this was especially important to institutions using both DSpace and Fedora (or Hydra/Islandora/another Fedora frontend)
- What are Sponsors INTERESTED in:
    - DSpace/Hydra - a lot of curiosity and interest in the potential of creating a DSpace Hydra head, Univ of Hull has a IR Hydra head, several institutions would like to have one solution for all content (student/faculty, digital collections, data, etc.), some concern about re-creating the DSpace workflows - critical that DSpace maintain its core functionality
    - DSpace/Islandora - a few institutions interested in Islandora - which is a Drupal frontend with a Fedora backend, some believe it might be easier / more accessible to customize Drupal than Hydra and that Drupal would provide more flexibiilty for displaying digital assets
    - DSpace needs to have a "repository abstraction layer" - e.g. REST module (GSoC project) - give the ability to use any UIs (i.e. Drupal, Ruby on Rails based, etc) or more easily "hook up" other things to Dspace
- DCAT & COMMITTER FEEDBACK (from first meeting on Nov 20)
    - Brought up RDF / Linked Data / Semantic web. Some interest expressed, but not a top priority yet
    - Preservation tools -- hang more stuff off the DSpace "hooks" (DROID / JHOVE / PRONOM / Migration jobs)
    - Concerns about "asking DSpace to do everything" - it cannot
        - Where do we start to draw the lines around what is "out-of-the-box" & what is third-party plugins. How do we become more modular.
        - Better APIs? (e.g. REST API)
    - Discussions about Statistics -- all feel it's such a basic need. Interest in enhancing Google Analytics connections & Stats in general.
    - Discussions on Scalability. The more we ask DSpace to do, the harder it becomes to make it "scalable".
        - Brainstorms of splitting up Admin UI versus View/Browse UI (the latter can be made more scalable by caching, etc)
    - REST API / Repository "Abstraction" Layer
        - Is there a way to add "Hydra" on DSpace via the REST API (Hydra communicates with Fedora via REST as Well).
        - Would allow you to use Hydra more as a view interface...tailored to specific content types even.

Discussion/feedback from DCAT/Committers - ALL

- Iryna
    - 2 hot topics in our community
        - research data mgmt and sharing - - need different types of metadata
        - keep hearing feedback from repo mgrs to have better author profiles
- Mark Wood
    - interested in archiving research data
    - need help finding collaborators - DSpace users doing similar things or would like to - but never get around to it because it is a little larger than they can take on on their own
        - Tim: notes we would like to have this be an outcome of these discussions - trying to find common solutions to common problems

- Ivan
    - +1 for author metadata / authority / author profiles
        - would like to do author profiles, but DSpace doesn't support author metadata well. There's no place to store this metadata in DSpace right now (for recent developments, see the dspace-cris work by CILEA)
        - This seems to be related to the Metadata For All brainstorms/initiative (which suggest to support metadata to **all objects** in DSpace)
    - need other types of metadata - need to store metadata for Journals, Communities and Collections. Currently you cannot translate Community/Collection names cause of lack of metadata support on them.

    - Also interested in Statistics – but that's a separate topic
- Mark Diggory
    - Seconds the idea of "Metadata for All"
    - We need to rework DSpace data model - allowing everything to have metadata - communities, collections and to allow for better translations, author profiles, local/external authority files
- Mark Wood
    - Revamping the "internals" of DSpace is just coming slowly. It's a lot of work.
- Ivan & Mark Diggory
    - Need ways to support structured Metadata in general. Currently DSpace only supports 'flat' metadata. No MODS/METS, etc.
    - Mark points out some benefits from Hydra in terms of metadata how it handles/manages/maintains METS.
    - We may want to look at DSpace model & perhaps see if we can model things that Hydra has done inside DSpace
    - But, don't throw the "baby out with the bath" – may not need to ditch all of the DSpace data model
- Leonie
    - Likes the idea of setting some "boundaries" around DSpace. We shouldn't ask it to do everything. Just do the things that DSpace does well. Let it have plugins, but not everything needs to be in "core DSpace"
    - Modeling of "loose" articles in DSpace is awkward. DSpace doesn't allow you to store journal article metadata easily
- General discussion about the need to do more collaborative sharing.
    - Schools do same work independently of one another. We should find a way to "institutionalize" a process for collaborating. Once a collaboration is established, developers shouldn't take it offline. They should continue to discuss on the developer list.
    - Ivan: We should identify shared requirements/ideas on the wiki. This step is missing
        - Mark Wood - agreed. Wiki is a good area for that. GitHub also good for coding collaboration & discussions
    - Someone else said we should publicize some successful models for collaborations that have taken place.
    - Start with one or two common cases -- to "demonstrate" this as a work in progress
        - We've heard these common issues today.
- Leonie
    - Newbies and repository managers have complained to her that they don't feel comfortable bringing things up in discussion with developers. They sometimes feel dismissed, don't feel their remarks are welcome.
        - Some newer folks on lists have expressed concerns about possible "negativity" on mailing lists. Mentions that one person felt their question/brainstorm was "shot down" on list – immediate response was highly technical and seemed somewhat negative in nature. Individual said they wouldn't want to post to list again.
        - Mark Wood mentions that a "technical response" is not necessarily a negative one... sometimes it's actually meant to be a sign of respect.. but maybe it didn't come across that way
    - Perhaps we need a way for newbies & repository managers to more easily provide feedback/brainstorms. Is DCAT the correct route?
- Mark Diggory
    - Sometimes questions go unanswered on the lists too. We need to find ways to ensure this doesn't happen
    - We used to have much stronger "regional" DSpace Groups...but those have mostly "faded away". Would there be any use to trying to bring these back?
- Someone else from Auckland (works with Leonie, but missed the name)
    - We need more statistical reporting. Things like detailed totals (e.g., counts of open access items in the repository).
        - Tim replied that we should pull together lists of reports that people would like to see.
    - Would be nice to have graphs automatically generated & ability to generate reports from DSpace
    - Mentions they are mostly using Google Analytics
        - Ivan mentions that João Melo (of Lyncode) is also working on Solr-based Content Analytics + DSpace. They may want to get in touch with him & give him some examples of reports they'd like to see
    - Maybe we should ask DCAT to help us generate a list of statistical reports/charts that would be useful? The Developers need a list to work from...otherwise we may not know what is most important.
- Mark Wood
    - It would be nice if DSpace itself could "plug in" to a larger structure so we could use it as a component (e.g., let other services provide statistics for DSpace). It need not do all the statistical stuff internally.
- Tim et al.: Some discussion of the desirability of having a "repository abstraction layer". This could also be a way to "hook in" external statistics stuff
    - Ivan mentions that one could even think of the OAI 2.0 Server (rewritten OAI-PMH server now in DSpace 3.0) as a form of "repository abstraction layer". You can actually query it rather easily (as it's based on Solr). May be ways to use it more.

Next Steps - Tim, Jonathan

- This meeting concludes our formal discussions around "DSpace Futures".
- DuraSpace will summarize / make transparent what the feedback has been from all mtgs / discussions - likely posted onto the wiki and distributed via the mailing lists
- not sure yet what projects will come out of discussion - need to find institutions/stakeholders who want to work towards common goals.
- We hope the discussions continue, and also hope that DSpace institutions will want to start up one or more projects (or managed projects) around of the common issues expressed during these discussions.