

DSpace 2013 Vision and Roadmap Meeting

•

DSpace 3-5 Year Vision and High Level Roadmap Meeting - May 9 & 10, 2013

- Logistics
- Attendees
- Preparatory Reading
- Contributions from Others
- Agenda
- Notes
 - General Discussion
 - Institutional Visions / Use Cases for DSpace
 - Repository Use Cases for next 3-5 years
 - Brainstorming Vision
 - Brainstorming Exercise: What Use Cases should DSpace meet for the next 3-5 years?
 - Basic Vision Consensus
 - Questions we need to answer as a Community
 - Next Steps

DSpace 3-5 Year Vision and High Level Roadmap Meeting - May 9 & 10, 2013

Logistics

- Meeting Location: UIC Campus, [Richard J Daley Library](#), Room 1-470 (Floor 1, Room 470)
 - Walking directions from [Crowne Plaza Hotel](#) to Richard J Daley Library (801 S Morgan St): <http://goo.gl/3uZOX>
 - Daley Library is located mid-way between Harrison and Taylor streets. It has a large west entrance on Morgan Street.
 - You can also take the #8 Halsted bus (runs every 10mins or so) for \$2.25 (they accept dollar bills & quarters). See CTA schedule: http://www.transitchicago.com/riding_cta/busroute.aspx?RouteId=167
 - If you are coming direct from an airport, the "UIC Halsted" stop on the CTA Blue Line train is the best option. From there, you can follow the walking directions and continue south on Halsted (~5 minute walk)
 - From O'Hare airport, just take the Blue Line directly to the UIC Halsted stop (No transferring). From there, you can follow the walking directions and continue south on Halsted (~5 minute walk)
 - From Midway airport, take the Orange Line to downtown Chicago. Transfer (for free) to the Blue Line at the "Clark & Lake" stop (descend two flights of stairs from elevated tracks to subway). Take the Blue line towards Forest Park. Get off at the UIC Halsted stop. From there, you can follow the walking directions and continue south on Halsted (~5 minute walk)
- Meals
 - Lunch & Dinner on Thurs will be provided for all attendees
 - Breakfast on Fri will also be provided

Attendees

- Jonathan Markow, DuraSpace
- Tim Donohue, DuraSpace
- Allan Bell, U of British Columbia
- Sandy De Groote, University of Illinois at Chicago
- Amy Lana, University of Missouri
- Jim Ottaviani, University of Michigan
- Sarah Potvin, Texas A&M
- Monica Rivero, Rice University
- Robert Sandusky, University of Illinois at Chicago
- Sarah Shreeves, University of Illinois (at Urbana-Champaign)
- Tito Sierra, MIT
- Maureen Walsh, Ohio State University

Preparatory Reading

- Summary from DuraSpace Summit discussion (March 2013 in Baltimore)
 - <http://duraspace.org/looking-future-dspace-and-fedora>
 - More Details from "DSpace Breakout" session: <http://bit.ly/YcEri5>
- Summary of "DSpace Futures" virtual discussions (Fall 2012)
 - <http://www.duraspace.org/dspace-futures>

Contributions from Others

- Leonie Hayes (University of Auckland) was unable to attend, but provided slides summarizing their institution's viewpoint: [DSpaceVision2013UniversityofAucklandcontribution.pdf](#)

- Elin Strangeland (Oslo and Akershus University College of Applied Sciences), also provided a statement: [DSpace vision workshop input 20130510.rtf](https://www.dspace.org/2013/05/10/rf)

Agenda

Day 1: Thursday, May 9, 2013, 12:00PM – 5:00PM

1. Lunch (12-1 PM)
2. Introductions
3. Expected Outcomes.
 - a. What do we hope to achieve by the end of these planning sessions?
 - b. What happens next?
4. Sidebar.
 - a. Diversity in the DSpace community
5. Vision and Product Placement.
 - a. What is unique about DSpace?
 - b. What important niche does it fill for you?
 - c. What about it provides value to your institution?
 - d. What is your vision for DSpace over the next five years?
6. Pain Points.
 - a. What has been most frustrating about the use of DSpace at your institution?
 - b. What characteristics of DSpace stand in the way of fulfilling your vision for the product?
7. Brainstorm: Use Cases and Associated Features.
 - a. What Use Cases are important for your institution over the next five years?
 - b. What are the associated features that need to be supported?
 - c. What kind of content needs to be supported?

Dinner out – 7:00 PM

Day 2: Friday, May 10, 2013, 8:30AM – 12:30PM

1. Light breakfast (8:30AM – 9:00AM)
2. Prep work on Vision Statement / High Level Roadmap
3. Prioritize Use Cases
4. Plan Next Steps
 - a. Volunteer assignments

Notes

General Discussion

- Do we need two platforms? DSpace & Fedora
 - Need to see if the 3-5 yr "visions" overlap for the platforms. Think of as a venn diagram - may be a lot of overlap or little
 - Would be important to the University Librarians - need a message as to why give to one or both. Show that we've analyzed whether merging platforms is worthwhile
- Types of DSpace institutions
 - Institutions who are essentially happy with DSpace as out-of-the-box IR
 - Institutions who are stretching the boundaries of DSpace
 - Faculty wanting something easier to use, "flashier". Even building their own tools, using other (non-preservation) system
 - "In between" - like the simplicity but want "flashier" interface, similar
- Is there a common vision for DSpace? (even amongst our small group)
 - In many ways it has morphed from it's initial use case that it was built for
 - Should it be a generic digital repository, or concentrate on solving just IR / preservation repository?

Institutional Visions / Use Cases for DSpace

(Anonymized, by request)

- Institution #1 - Lots of integration points & access - less about preservation
 - DSpace is free, relatively robust. Large User community.
 - End user deposit. published & unpublished content
 - Managing diverse research output (ORCID). Data with access controls. Digital Collection & Mgmt)
 - Research info mgmt systems. Needs good integration points
 - Integrates into a different digital preservation.
 - Streaming server, stats module were added as they went
 - "Killer App" = E-Theses. Harder stuff is images/video.
- Institution #2 - Started small and simple, constantly expanding
 - Initial decision was it is open source. Philosophy to support OS
 - Capturing university output
 - Getting to streaming servers
 - "In between group" - like ease of use. Small library - students could be used to do input
 - Migration of some content from ContentDM to DSpace. Having requests to extend DSpace to add some ContentDM features

- Feeding publications (university publishing) direct to DSpace
 - Getting data in and out easily
 - Went with Islandora for a Digital Library solution (better "Digital library" product than DSpace)
 - Question has come up whether to use Islandora instead of DSpace for some content
 - Possibility: Using DSpace as a true "preservation repository" and feeding content to Islandora (or similar)
 - Positives outweigh the negatives at this point. But, how many systems can they really support amongst digital library / IR services?
- Comment: "DSpace with a lot of 'hooks' on it" - could solve a lot of use cases with good integration points. But, shifts focus of spending staff time integrating and supporting a larger suite of software.
- Institution #3
 - At time of adoption (early on), unique & filled a necessary role. Capturing the scholarship in a repository (initial needs came from library community)
 - Main concerns are performance issues / scalability
 - Handling preservation mgmt in DSpace
 - Continued modularization of DSpace - lots of things people want, but do we keep adding into DSpace.
 - DSpace not a swiss army knife.
 - Lack of flexibility for non-text formats
 - Handle issues - cannot move content around easily as you cannot "split" a Handle prefix
- Institution #4
 - Important that it is OS and successful with textual formats. Good submission workflows.
 - Built up a lot of local expertise with DSpace
 - DSpace as sole Digital content mgmt system
 - Lots of user demand for images & data. DSpace not designed for these materials
 - Need for stronger preservation support.
 - More complex metadata
 - Moving in a more modular direction. Want DSpace to fit well into that ecosystem (modular instead of "stand alone")
 - Not the staffing to support Fedora. DSpace is "perfect fit" in that it's turnkey, etc.
- Institution #5
 - DSpace provides Persistent long term access. Easily findable items
 - Want a system that can meet multiple types of needs. Not enough staff to support many systems
 - DSpace is part of preservation strategy (and DuraCloud and other tools)
 - Need for stronger preservation support
 - Need to better support special collection
 - Journal article metadata becoming more critical. As is data
 - Want it to also "work well with" streaming server solutions (for video / audio). Better integration
- Institution #6
 - At the time, it was the "out-of-the-box" product
 - Place to put documents for easy access.
 - Using both for archival materials and scholarly content
 - Future to make it look "partitioned" to search types of content separately
 - Integration with things like Symplectic Elements and/or VIVO. Pull in metadata from external sources (easier deposit)
 - Data becoming more critical. Both open access data, and data only for local community
 - Managing research data (long tail data...small data)
 - Hard to get stuff **out** of DSpace once it is in there (e.g. move it elsewhere). Handle issues (cannot split up handle prefix)
 - Willing to run different systems for different purposes. But, limited staff – so needs to be easier integrations. Simplicity important
- Institution #7
 - Role: mature IR platform, but has not evolved to solve all the various other use cases beyond narrow IR institutions
 - Imagine DSpace as an IR "backbone". Enforces various use cases for IR needs. But interoperate with other tools/services that can solve other use cases
 - Interoperability with Tools: e.g. DSpace more friendly with existing tools that solve preservation problems / dissemination, etc.
 - Interoperability with Services: Large user community, which could be leveraged to build an 'ecosystem' of services which are "DSpace-aware".
 - Framework for modules/plugins, which would allow institutions/service providers to integrate other services into DSpace. Could be supported by DuraSpace
 - Don't want to build more & more functionality on top of a monolith. Want to create an "adapter" to plugin to other services & tools.
 - Some examples: Discovery & Access
 - E.g. specialized interfaces for searching across ETDs. Perhaps ways to link that up to printing ETDs.
 - E.g. distributed digital preservation
 - Why use DSpace instead of something else
 - sunk costs - costs to switch
 - not a lot of digital content solutions that meet the base IR needs
- Institution #8
 - Twin goals of DSpace: preservation & access to research & scholarship
 - content has to be related to research / scholarship. Other types of content go in other systems
 - Worked well for that purpose. Works well with textual docs.
 - Now getting some images / research data sets. Small sized to medium sized data sets, DSpace works well
 - Limitations in terms of preservation side of things
 - investing in Fedora as a preservation platform (for all content, not just IR)
 - DSpace will be more of an ingest/access system. Preservation will be in a separate platform "underneath"
 - Need to move content easily in/out of DSpace because of that
 - Increasing value. Use ability to delegate control of Collections & Communities to departments to do their own training/submissions. Easy for people to pick up and use in that way
 - Have a large amount of "sunk costs". Would like to see platform/community move
 - DSpace should continue to provide base IR functionality. But, expand to handle more complex environment (e.g. relationships between sets of items).
 - DSpace should either improve with Preservation or have easier hooks to other preservation tools/services
 - Easier hooks into research profiler system or similar
- Institution #9
 - DSpace is about Preservation, visibility & access to your work
 - Dspace great at end user deposit, creation of collections.

- Do virtually zero vetting of what goes into DSpace. Trust faculty to make this decision
- "Directors cut" - multiple things under one handle
- Good that you can put anything in it. Can be a preservation problem
- Preservation tools could be improved.
- Like open source nature.
- Want to look at handling small or large data sets in DSpace
 - hard to get stuff "out" (especially large data sets)
- Concerns about the monolithic nature of code. Need: "set of legos" instead

Pain Points / Frustrations

- Poor end user experience
- Customizations are "hard". Plugging things in. Code modifications (monolithic)
 - Hard to maintain once you make customizations. Upgrades become more painful.
- Current Content Model - especially difficulty with relationships
 - no metadata per bitstream (e.g. preservation or admin metadata)
 - different types of files all related, but requiring their own unique metadata
 - no hierarchical metadata
 - no relationships between items
 - Needs a more flexible content model in general (hierarchical content model)
 - for preservation use cases, you might want to organize in one way. for access, perhaps another way
 - Communities, Collections & Items hierarchy do not work for all use cases
 - inflexibility of this model causes you to have to work around it or "hack it"
- No native support for complex metadata
 - Research data metadata is hard
- Lack of training possibilities
 - Lack of user documentation for DSpace
- Cost of ownership. Making installation/configuration/upgrades easier
- DSpace primary UI technology based on aging technology (Cocoon)
- Ease of use of getting data in/out of DSpace (metadata, actual content, etc.)
 - Getting data out in a form that is "useful" to researchers (for data mining, etc)
 - Also statistics lost if you move data out and back in
- Scaling concerns.
 - Concurrency issues (tuning for large scale concurrent access)
 - Scaling issues related to Collection size
- Getting content in/out
 - Delivery of large files out of DSpace
 - Also getting large files into DSpace
 - Improved support for Bulk Uploads into DSpace (not to have to send to your programmer)
- Governance & getting things (fixes / features) into the codebase. Not enough developer resources.
- Model to share common tools into a "commons" that are "DSpace aware". Lack of a framework to share these tools & manage.

Repository Use Cases for next 3-5 years

- Large research data sets / large files / big images/videos
- Need for streaming video / audio service
- Integrated publishing system
 - publish journal articles
- Current Research Information System (CRIS) (BePress does that...why doesn't DSpace)
 - Faculty Research Pages
 - e.g. Hong Kong's work with DSpace-CRIS
- Preservation Management
- Newspapers, Serials, Complex Objects in general (or interoperability with an external system to handle)
- Interoperate in general with external tools & services
 - Interoperability at any level of the DSpace hierarchy (Items/Collections/Communities) to other services
- Archival vs. Access Copies - distinguishing different file types (for different use cases)
 - Storing master images (archival copies) - tag it in a particular way for preservation services
 - But, display a lower resolution copy (access copy)
 - Almost better relationships between files (and allowing metadata on individual files)
- Building different access "views" of objects (based on the type of content or audience or similar)
 - Possibly enabling different functionality per type of content (e.g. image viewers, document page-turners, ETD search/view, geospatial data)
 - Not necessarily a different **interface**, but a different "visualization" of content.
 - Image Server / Page Turner / Geospatial / Media Player
- More ease of branding. Not having everything be "DSpace-wide"
 - More customization abilities / theming at Community/Collection levels.
 - Making this process easier. Provide a set of templates / base themes. Manage this from the UI or similar
- Version control
 - In the control of the end user. So end users can choose when to version/update their content
- Mediated & Author self-deposit
 - Mediated = approval workflows, batch loading
- Metadata Editing
 - Batch tool that is Admin UI-based
- Self-service configuration (manage configuration from the Admin UI)
 - Ingest forms
 - Controlled vocabularies, etc.
 - More admin tools made available to UI
- More Metadata Schemas

- PREMIS
- Geospatial
- Tools that automate extraction of technical metadata (e.g. duration of videos, other admin/preservation metadata)
- Granular Access Controls
 - Limiting access to new Item deposits as needed
 - Better communicate what is open access and what is restricted access
- Identity Management
 - Author IDs
 - Object IDs - Not just Handles (also DOIs or other identifiers)
 - Authoritative Handling of Identification
- Statistical Reporting
 - Usage statistics (filtering out spiders/bots by user-agents)
 - Analysis of repository content
- Search Engine Optimization (SEO)
 - support different use cases
 - need to constantly keep on top of it

Brainstorming Vision

- If silent majority likes simple, out-of-the-box...but others want extra functionality. Is this a reason to investigate more closely DSpace + Fedora integration
- If we want to preserve simple / out-of-the-box, do we need to concentrate more on the "core". Concentrate on making it modular (lots of hooks) for any "non-core" features / functionality.
 - harder to support a system that keeps adding more and more functionality (e.g. JSPUI & XMLUI)
- More concentrated "core" would improve sustainability of the product/project
 - more understandable, easier to maintain
- A lot depends on how the community would build extra "modules" / services
 - How to support these extra "modules" in a sustainable way
- "Freezing the Spec" at some point? "Effective core functionality" is whatever is in 3.x or 4.x or similar?
- Stepping back and re-thinking what is the "value" of DSpace. What does it do best?
 - e.g. a Content Model, a core set of services = make up the "core backbone" of what is DSpace
 - Stand up something simple with core services. Try and get others to migrate to this new platform and build for there.
 - Could "hosted DSpace" be a place to try this out and have customers help support extra module development
- Challenge: we don't have a vibrant ecosystem for enhancing the DSpace platform
 - System not setup to be able to "evolve" to address new use cases.
- Hydra as an example Fedora-based framework
 - Many Hydra developers need not know about how Hydra communicates with Fedora
 - The Fedora "complexity" is hidden from institutional Hydra developers (who mostly work in Ruby on Rails)
 - The connections between Hydra & Fedora are maintained centrally as the Hydra "core" (by the primary Hydra Committers)
- Whatever we choose. We should optimize for a "software as a service" use-case. Wonder if lots of institutions would gladly pay for a hosted solution elsewhere.
- Existing Community vs. Potential Community
 - Need to think about upgrade paths of existing community (obviously)
 - Also consider - are there a blossoming set of use cases (white house OA etc) which would be interested in a DSpace-like platform. Perhaps software as a service solution.
 - Don't "shed" too much of the existing community – but also want to expand potential community.
- Need a real "turn-key" IR solution. Both free Open Source, and a hosted solution.
- What was a traditional IR 8-10 years ago is quite different than today. Still interested in DSpace as a modern traditional IR
 - DSpace as an IR for the next 10 years. Not necessarily well suited for that now
- IR for the next 5 years
 - software that plays well in an ecosystem of services (easier to get content in & out of DSpace).
 - Solve the IR needs, not necessarily all general digital repository needs.
- Institutional Asset Management system v. "All in one" digital repository system
 - What if you have other services be "DSpace aware". External tools/services an "slurp" in content (based on types/collections) and provide other views/services (page turner system, etc.).

Brainstorming Exercise: What Use Cases should DSpace meet for the next 3-5 years?

- We took part in a brainstorming exercise around what common Institutional Repository Use Cases should be a part of "core" DSpace, and which could be handled by external systems/tools/add-ons.
- Essentially, we grouped Use Cases into three main categories:
 - **"DSpace Core Use Cases (next 3-5 years)"** : These are use cases we feel should be met by "out-of-the-box" DSpace.
 - **"Possible Extensions to DSpace Core"** : These are use cases which could be provided "out-of-the-box", or might be met by external tools/services (or DSpace "add-ons" / plugins)
 - **"NOT provided by DSpace Core"** : These are use cases which we feel should NOT be provided "out-of-the-box". They should either be handled by integrations to external services/systems, or they should be developed as a DSpace "add-on"/"plugin" which you can install in your DSpace instance.

DSpace Core Use Cases (for next 3-5 years)	Possible Extensions to DSpace Core (some may be external services or DSpace "add-ons")	NOT provided by DSpace Core (but possible services DSpace should integrate with)
---	---	---

• Create, Read, Update Delete (CRUD) on objects • Self deposit & mediated (approval workflow based) deposit of content • Access controls (Authentication & Authorization) ◦ Also includes Embargo-style access controls • Batch Deposit of content (from a UI) • Batch Download of content (from a UI) • Basic Search & Browse functionality • Basic Preservation functionality (e.g. Fixity checks) • Basic Statistics (or "hooks") • Default out-of-the-box User Interface ◦ Preferably some sort of template-driven UI framework • Standard Machine Interfaces (e.g. OAI, SWORD, REST API) • Persistent Identifier support • User Interface should be "SEO Friendly" • Structured Metadata ◦ Metadata should be at all levels of object hierarchy ◦ Hierarchical Metadata formats should be supported • Licensing support ◦ Both deposit license and Creative Commons licensing • Support for Derivatives (e.g. thumbnail images) • Large File Support for End Users ◦ End Users should be able to upload and download large files themselves • More Flexible Relationships ◦ Including aggregations of objects, complex objects • Community & Collection "like" hierarchy • Ability to easily "hook" into external tools & services ◦ e.g. Curation Tasks & more robust ways to integrate with other tools/services • Versioning of objects • Configuration Management in the UI • UI Template/Theme Management in the UI • Machine interfaces should be able to target content at any level (Community, Collection, Item) • Enhanced Content Model ◦ Community, Collection, Item "like" model ◦ Should also include Author objects (which hold metadata about authors /researchers in the system) • Administrative Metadata at all levels	• Richer Licensing support (individual CC licenses on individual files) • Support for Delivery of Media ◦ Doc Viewers ◦ Geospatial ◦ Streaming content • Alt-metrics (downloads, tweets, etc.) • Support for small scale research data sets ◦ Relationship back to publication (linked) ◦ Also may include software programs • Metadata extensibility ◦ Stronger support for channeling user contributed metadata ◦ Schema agnostic • Compliance with Open Access directives (of various countries) ◦ models to track with general worldwide OA directives ◦ when possible, methods to check compliance ◦ when possible, support for automated evaluation • Improved Statistics (could be external, e.g. Google Analytics) • Improved Support for External Identifiers (DOIs, Handles, etc.) • Customized / Flexible UI support ◦ Users should be able to change their Collection's "theme" or "template"	• Advanced Statistics engine ◦ instead should look towards integration with Google Analytics • Advanced Preservation Activities ◦ instead should provide integration with external preservation tools / services • Publishing System ◦ instead, should provide integration with external publishing systems • CRIS (Current Research Information System) ◦ instead, DSpace should integrate with CRIS systems, or offer a CRIS plugin.
--	--	---

Basic Vision Consensus

- Getting back to basics & getting the basics right. Focus on fundamentals
- Re-architecting DSpace to be "leaner", but more flexible
- Core functionality that can be "extended" or have "hooks" to other services
- Designed in such a way that it can be easily/quickly configured to integrate with new tools/services in a large "ecosystem"
 - Agility and flexibility is a goal
- Want to support low-cost, hosted solutions/deployments
 - Has the benefit of potentially broadening the potential user base

Questions we need to answer as a Community

- What are those core pieces & what is needed to make those pieces "better"?
- Are we going to continue going down the path of an Open Source project primarily implemented as a local "stack"? As opposed to a model with explicit support for hosted-services as a primary vehicle
 - E.g. Drupal & WordPress can be thrown up on an ISP quickly/easier
 - Allow for rapid & hosted deployment as a model
 - Are we shooting for a hosted deployment model?
 - Do we want to expand community in this way?
- What are the other communities that we want DSpace to "play well with"?

Next Steps

- Getting to a **vision document** - describe overarching vision & use cases (not technical implementation)
 - http://en.wikipedia.org/wiki/Vision_document
- How does Governance discussion fit in?
 - Do we need to wait on Governance till we can get closer to a technical implementation plan.
 - Is OR13 an opportunity to get "buy-in" on the Vision (at a high-level), before even getting to technical implementation plan.
- Draft a Vision document from our five bullets above & our lists of core versus non-core use cases.
 - Very rough draft begun at [DSpace 2013 Vision Document](#)
- Visioning before Governance
 - Need to get excited about vision to form Governance group.
- Getting "buy in" at OR13
 - Could we introduce this idea as part of the DuraSpace Plenary?
 - Have a broader discussion as part of the DSpace User Group Meeting (just after the Plenary). Some sort of Panel? Open Discussion? - Tim can talk to DSUG folks