HYDRA-IN-A-BOX

USER INTERVIEWS & FOCUS GROUPS

Executive Summary

As part of the Hydra-in-a-Box discovery process, we conducted user interviews and held focus group sessions to form a better understanding of the digital collections and digital asset management system needs of the library, cultural heritage, and memory institution community.

Method

In August 2015 we developed a set of interview questions and supporting documents (consent form, a checklist and set of guidelines for interviewers) to use for one-on-one participant interviews. We also made minor modifications of those documents for focus group interviews. Interviewees were primarily selected from respondents to the user survey who indicated willingness to participate in an interview. Interviews were conducted both in-person (at the 2015 Digital Library Forum conference held in Vancouver) and remotely using the BlueJeans video conference system. Interviews averaged 45 to 60 minutes and were audio-recorded.

A total of four focus groups were conducted. Two of these took place at the 2015 HydraConnect conference in Minneapolis, MN, and included volunteers who were attending the conference, largely consisting of people from universities using Hydra. Another focus group was conducted at the University of Minnesota and consisted of representatives from Minneapolis area library, cultural heritage, and memory institutions. The final focus group was conducted at a meeting of the Mid-Atlantic Fedora Users Group in Philadelphia.

The audio-recordings of both one-on-one interviews and the focus groups were transcribed and imported into the Dedoose content analysis tool. Members of the Hydra-in-a-Box Design and Requirements Specification (DRS) team developed a coding scheme in Dedoose (dedoose.com) and then processed the interview transcripts by applying codes to potentially relevant passages in the transcripts. Each transcript was coded by both a primary and a secondary coder. A total of 93 codes were applied to 1258 passages from the 49 distinct interviews. The DRS team analyzed this complete set of coded transcripts to produce the detailed findings below.

Participants

A total of 23 people were interviewed, representing 21 different institutions. In addition, there were 34 participants in the focus groups, representing 27 different institutions. There was a slight overlap in interviews and focus group participation, so in total we talked to 55 people, representing 46 different institutions. When soliciting potential interviewees, we tried to obtain roughly equal numbers of small, medium, and large institutions, where size was determined by institution student enrollment and digital library staff size.

The distribution of interviews by institution size and staff size is shown in Table 1.

Size	Institution[1]	Staff[2]
Small	12	15
Medium	11	17
Large	20	13

Table 1. Number of interviews by institution and staff size

[1] Institution size based on student enrollment (small: < 7,000; medium: 7,000 - 15,000; large: > 15,000).

[2] Staff size based on number of FTE devoted to digital collection creation and technical management (small: 0-2; medium: 3-5; large: > 5).

The distribution of interviews by institution type is shown in Figure 1.

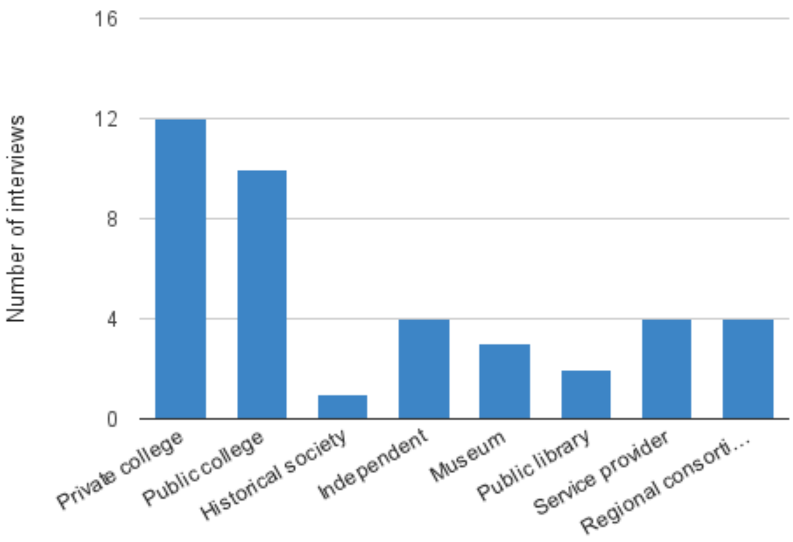

Figure 1. Number of interviews by institution type

Summary of Findings

The next several pages contain a summary of the findings from the user interviews, broken into ten sections based on our interview questions. For complete findings in each of these ten sections, see the "Detailed Findings" section that follows this summary.

Current repository systems

Interviewees mentioned many challenges and pain points associated with their current repository systems. Customization is a challenge because "out of the box" solutions tend to offer little that is easy to customize, but third-party customization can be costly, and local customization difficult to maintain without good documentation. Commercial software is seen as inadequate with poor support (and costly), but open source solutions require development and operational support that many institutions cannot commit resources to.

A big pain point for institutions is lack of desired features. Limitations in features and functionality within a given solution drive institutions to turn to additional solutions to fill the gaps. Interviewees mention a desire for bulk operations for ingest and metadata, preservation features, more robust metadata transformation, support for wider diversity of content types, and administrative features (among others). Customization and configuration is also a concern. Institutions would like the flexibility to have configurable workflows and access policies, add facets for browsing local terms, have custom metadata templates by content type, and tailor front-ends to particular content types and themes.

Preservation practices across institutions cover a spectrum, from doing nothing to backing up to third-party cloud storage (Glacier, DuraCloud) to integrating with a third-party preservation service (Archivematica, ArchivesDirect, Preservica, Rosetta). Interviewees told us they would rather outsource preservation infrastructure and services than develop their own; have a better handle on preservation metadata; and preserve versions (or at least track/log changes), particularly of metadata and scholarly content. They'd like preservation to be automatic and invisible to those users of the system who are not responsible for preservation. And they'd also like the preservation system to be independent from the asset management repository, so that components of the layer can evolve and be replaced without impacting parts in other layers.

Staffing is a big challenge. In-house staffing levels for repository development and technical support tend to be very low, and hiring for technical skills (particularly Ruby) is difficult. Using contract/vendor services to support local deployment and customization can help small institutions, but these services tend to be very costly and recurring, and if the work is not executed or documented well, it is difficult-to-impossible to maintain and sustain over time. Many institutions are dependent on central IT for infrastructure and tech support, which adds complexity and hinders flexibility and responsiveness. Regional consortia/service providers are a good option for those without local or central IT staff, with the tradeoff of limiting customizability, which tends to make customers complain. Hosted solutions are another alternative, but interviewees mention poor responsiveness to feature requests and technical support needs, and a suspicion that small institutions don't get timely assistance.

Decision-making

When asked about how their institution makes decisions about their repository solutions, interviewees discussed a wide range of factors. Cost is important, both in terms of the monetary cost of licensing, hosting, and/or outsourcing and the cost of staff time. Available features (support for their resource types, flexibility, ease of use) and the larger ecosystem are factors, with interviewees mentioning knowing how the repository software could be integrated with other institutional software and technologies, including locally developed projects, as important. Another decision-making factor was migration, where institutions are concerned with the difficulty of migrating their content and data to a new system and whether any data will be lost in the transition. Interviewees were on both sides of the fence in terms of open source versus proprietary solutions, most often based on their experience and comfort with open source software efforts.

Interview discussions around hosted solutions revealed both positive and negative experiences. While hosted solutions were seen as beneficial for institutions that lack the staff or technical infrastructure to maintain their own solutions and freed up local staff to focus on collection management, interviewees complained about limited customization options and the ability to add needed features. Important to some institutions is the ability to maintain separate settings, controlled vocabularies, branding, and UI's when using a hosted solution, especially when they are part of a consortium. Having help with migrating to the new system is also a concern.

Some specific questions and concerns about making decisions around Hydra-in-a-Box included wondering about support and sustainability, such as who would own the application component and whether Hydra-in-a-Box would offer support for both framework and application level issues. The vagueness of what exactly Hydra-in-a-Box will be arose more than once, and there were questions about whether the "out-of-the-box" solution would be robust enough to meet the needs of adopting institutions. Migration was also mentioned as a concern about Hydra-in-a-Box, as interviewees wondered whether the project would facilitate migration from other repository solutions.

Migration

Many interviewees indicated that their institution is interested in migrating to a new repository system. A common frustration is having to use multiple repository systems to manage all of the institution's assets (e.g., separate IR and digital collection systems), which create headaches in locating and remediating objects, duplication, different approaches to metadata and cataloging, etc. Whether or not they are using multiple systems, many institutions would like to migrate to a new system because they are frustrated with their current system. A desire to migrate from Bepress/Digital Commons, CONTENTdm, DSpace was expressed by multiple institutions; migrating content from an Omeka site was also mentioned.

Most institutions seem to not have migration experience yet, but are concerned about the challenges involved in a migration effort, or whether it is really feasible at all. Mentioned more than once was the worry of how much effort would be needed to do metadata enhancement and remediation as part of the migration process. Institutions want migration to be easy and risk-less, in terms of losing data, and ideally they'd like support for doing metadata quality assurance and remediation before (or as part of) migration. Those wanting to migrate to a hosted service expect that the service will provide a path to migrate their data.

Storage

For those institutions serving **small-to-medium sized communities** with very few staff dedicated to digital collection management, numbers of objects tends to range in the tens of thousands and storage size is relatively modest (fewer than 10 TB). For **medium-sized institutions with more resources** devoted to collection creation and management, the number of objects heads into the hundreds of thousands and storage size grows to low-tens of TBs. For those **larger institutions serving large communities** who devote, relatively speaking, the most resources to digital collection management, the numbers of objects can reach 1+ million and the number of TBs approaches mid-to-high tens of TBs.

Growth projections are challenging to make with certainty, and most institutions don't really make projections. However, numerous interviewees mention growth of audio/video content and the associated storage concerns in the same breath. The location of file storage can matter, as some institutions specifically seek repository solutions that include geo-diversity of redundantly-stored content, while others must comply with government mandates that stipulate that resources are not stored outside of particular geographic region, such as a national border.

Managing large files in the context of a hosted service introduces challenges, such as hosted service costs, the need of some institutions to maintain high resolution copies locally for ready access when fulfilling patron requests, and the need for mechanisms to support large file transfer outside of browsers and http, which interviewees told us are not reliable for such transfers.

Collection management

A common challenge for institutions is being able to easily identify or inventory all of the digital assets they have. Institutions often employ multiple repository systems to manage their assets and being able to inventory and report statistics across systems is difficult. In addition, existing systems don't enable administrators to perform all of the functions on items managed in the systems that they would like to perform. Interviewees said they wanted collection management features such as:

- Better reporting capabilities
- Ability to have many-to-many relationships between collections and items
- Manage all digital assets from a single system
- Have a range of integrated administrative tools
- Customization capability for appearance and branding

Content types

Institutions tend to manage a wide variety of content types across the spectrum of library and archival collection materials as well as academic and research-based publications and data. Images represent the content type most frequently managed by interviewees, followed by audio/visual, articles, theses/dissertations, archival collections, text documents, data sets, and newspapers. Content formats such as email, software, disk images, GIS content and web archives do not represent a significant amount of the content types that institutions are now managing in repositories. Audio/visual and newspapers were mentioned most often as challenging content types.

In terms of future needs, audio/visual, archival collections, data sets, and articles were the content types mentioned most often. In addition, "compound documents" (oral history audio with text transcript, publication with data, multimedia works) are important to institutions now and are likely to be more so in the future. The increasing amount of audio and video content poses unique challenges. Institutions are struggling to manage it due to storage constraints and inadequate delivery options, and preservation is of particular concern, because of the large file sizes and corresponding costs to store and manage. Managing content from archival collections is also a challenge for many institutions, because commonly used repository systems, such as DSpace and CONTENTdm, don't support content and metadata that is hierarchical in nature very well. Newspapers are a special category of content: due to large page size and their serial nature, common repository solutions don't handle them well but specialized applications do.

Workflows

Interviewees said that the repository system needs to support multiple user roles with different privileges, and permissions or roles might need to be assigned at the collection level. The system should be able to handle multiple back-end users simultaneously while also supporting front-end search and use. Interviewees want to be able to both track steps in a workflow and capture nuanced curatorial information or feedback to the metadata creator in notes. They'd also like this information stored in the repository system. Approval may be needed at different steps, not simply to "publish."

Workflows often differ based on type of object digitized and interviewees expressed a desire for configurable workflows based on project/user/item types, and for a repository to support multiple interfaces for upload of objects. Interviewees would also like to configure workflows for self-submission, where metadata requirements, authentication, and interoperability with other systems (e.g., faculty portfolio software) might be relevant. Validation of self-submitted items based on local configuration (e.g., controlled vocabularies) and other variables (completeness) was mentioned.

A range of metadata creation workflows are currently used by institutions and interviewees spoke of the need to support multiple types of metadata creation at once. Workflows for bulk ingest, single item creation, and self-submission might all be needed in the same system. Interviewees want metadata creation and workflow forms that are intuitive and easy to use and include features such as auto-complete from controlled vocabularies and pre-population with information from other systems (i.e., MARC records). Approvers of records want to be able to see queues of records for approval, give feedback to creators of records, and approve them in batches. Interviewees said that technical metadata should be created automatically upon ingest but some also expressed a desire to be able to see and edit that technical metadata.

Metadata

Interviewees told us that metadata in repositories doesn't always originate there. It can be migrated from other systems, especially MARC catalogs and the easy transfer of data between these systems is crucial. In addition, some of the metadata retrieved from other systems will be partial and will need to be further enhanced before it is published in the repository. Interviewees told us that they want to be able to validate metadata fields against existing standard controlled vocabularies and authorities, and they also want to be able to define their own local vocabularies. They'd like to implement a controlled vocabulary either across the entire repository or for a single collection would like to see URIs for terms when possible.

When discussing types of metadata, multiple descriptive metadata schemes were mentioned, mostly MODS, DC and QDC. One interviewee mentioned VRA Core and GIS metadata. Several mentioned a desire for the system to be metadata agnostic. Many interviewees mentioned technical and structural metadata, but only one mentioned a specific schema (PBCore for video). Support for mapping to simple Dublin Core was mentioned in connection to Omeka, while support for mapping to the DPLA MAP was mentioned in relation to aggregation.

Institutions also need to be able to export metadata, in multiple ways. They may want to map the metadata to another standard for export, including for publication in a feed for others to aggregate. Multiple interviewees also mentioned wanting to be able to export metadata as a delimited file for use with OpenRefine for analysis or editing. Metadata exports should be in bulk. While it may be useful in some workflows to download individual files, downloads of entire sets of metadata are more important.

Discovery/UI

When asked about discovery and UI features they'd like to see, many interviewees said an ideal repository system would support multiple discovery layers and emphasized the need for powerful index and keyword search. SEO is an important component of discoverability for many interviewees. Interviewees wanted interfaces that are optimized for search engines. Many interviewees want to create and present collections along with contextual and descriptive information specific to that collection.

It is important to institutions that they can customize the theming and branding of the end-user interfaces of the system. Sub-collections within a single repository may need distinct branding. Some interviewees stated that it is important to customize aspects of the discovery interface, such as which metadata fields to display in a search result, which to display on an item show page, and which facets are shown to the user. Interfaces should be tailored appropriately for different content types. For example, readers with zoom, rotate, and pagination features should be used for text- and image-based works. Finding aids should include content lists and links to digitized content. Some interviewees want the ability to incorporate custom features, such as digital humanities tools, annotation tools, etc. And several interviewees expressed a desire to collect some form of user contributions, including comments, annotations, translations, and OCR corrections.

Although we asked interviewees about social and personalization features, overall there was not a lot of discussion on this topic. Social media is a part of many institutions' engagement strategies, however, and interviewees would like to control which social media buttons are available to their users and customize how they work. Many interviewees indicated that their institutions are currently using, or plans to use, exhibitions of their digital objects to engage broader audiences. They'd like the ability to build exhibits around repository collections, and to integrate exhibitions into repository search results. Potential creators of exhibits include librarians, archivists, curators, academics, students, and collaborations between the aforementioned groups.

Usage analytics reporting

Many institutions consider reporting features to be a must-have in a repository system. Usage analytics can serve as a communication tool for repository managers/librarians, who commonly need to report on usage to various stakeholders including management and administration, institutional repository (IR) contributors, content providers, and academic departments. Many institutions currently use Google Analytics as at least a base level of reporting, but find it of limited usefulness. Some institutions rely on reporting features built-in to their current systems, but find they don't provide data for all the things the institution is interested in and when built-in reporting breaks, it is difficult to understand why. On the other hand, more powerful commercial analytics products are seen as too expensive for some institutions.

When asked what reporting capabilities they're looking for, most institutions expect basic metrics (by object, by collection, by month, etc.) such as download counts, view counts, and geolocations of users to be available in a reporting feature. Going beyond those basic metrics was important to several interviewees, who suggested that knowing how an object is used (e.g., if and where it was cited, which tweet mentioned it, a blog post referencing using an object, how it was referred to in social media) would be useful, as would enabling depositors to optionally display download/view metrics for their items. A feature of Digital Commons where contributors have an option to receive regular emails with statistics about their deposits was given as a positive example. Some institutions are also interested in assessing user experience and are looking for ways to do that beyond simple count metrics. For example, being able to see the kind of queries or facets that are used by end-users, or seeing how long it takes a self-depositor to complete a deposit.

Interoperability

When asked about needs for interoperability with other systems, interviewees mentioned a wide range of possibilities. Specific use cases that came up included reading/writing data from the repository and index through an API; leveraging a variety of metadata standards, controlled vocabularies, and digital proxies; migrating data in and out of the system, and crosswalk data between different schema; harvesting data using an API; and determining whether records have changes through an API.

Specific types of systems interviewees mentioned it being useful for the repository system to interoperate with included preservation systems; IIIF servers; RSS feeds, exhibits (Omeka, Spotlight); library catalogs; external metadata streams, such as Elsevier Pure and Wikimedia; external indexing services, such as WorldCat and academia.edu; systems for managing conservation and loan information; and systems for users to request copies of digital assets and permission/license for re-use.

Detailed Findings

This section contains our complete interview summaries, broken into the areas below, based on our interview questions.

Current repository systems
Decision-making
Migration
Storage
Collection management
Content types
Workflows
Metadata
Discovery/UI
Usage analytics reporting
Interoperability

Current repository systems

What do people say (positive or negative) about the repository products they are currently using?

CONTENTdm

Positive

- Metadata interface and tools are very “understandable”
- Ingest process is easy for non-specialists
- Strong performance under load by multiple concurrent users

Negative

- High cost for inferior product and vendor support
- Desktop client is difficult and frustrating to use
- Inability to scale for very large collections
- Lack of support for large files and diverse content types
- Difficult to search across collections
- Lack of flexibility and customization
- Exported metadata is “unusable”

DSpace

Positive

- Role management works well
- Facilitates both deposit of access-restricted content and user request for access to restricted content
- Supports hierarchical content
- Provides preservation support

Negative

- Generally perceived as “old”, something to move away from
- Accumulation of “derelict code” and a “spaghetti mess” under the hood
- Is more understandable by librarians than faculty and student depositors
- Statistics/analytics are “terrible” and frail
- Lack of customizability

Bepress / Digital Commons

Positive

- Wide adoption in some distinct scholarly communities, e.g., US law schools
- Metadata template customization
- Strong analytics

Negative

- Dependent on tech support to make even basic changes
- Proprietary implementation of Qualified DC metadata makes field and mapping changes problematic
- Limited support for content access restrictions

Islandora

Positive

- From a service provider point of view, has real potential to grow a market
- Drupal component is attractive, as Drupal is familiar to many
- Open and welcoming community

Negative

- Code development has been driven more by singular contributors and a single vendor than a community
- Complexity of system makes development challenging
- Vendor support is costly

Fedora

Positive

- Strong community facing common issues, providing shared solutions

Negative

- Slow to ingest, slow to index to Solr

Homegrown

Positive

- Full control over specifying/adding/removing features and the UX
- Not tied to a vendor's service or road map
- Can customize display specific to content type
- Custom integration with other institutional services and web platforms, e.g. Drupal

Negative

- Migration out can be painful

Regardless of any specific solution in use, what are the general "challenges" or "pain points"?

Customization

- "Out of the box" solutions tend to offer little that is easy to customize.
- Third-party customization can be costly

- Any local customization is difficult to maintain if not documented well
- If a product solution grows old without keeping up with latest functionality and look-and-feel, the impetus to implement local customizations grows stronger.

Operational support

Open source solution

- Implementing locally an open source solution requires development and operational support that many institutions cannot commit resources to
- Having to engage in a larger community to keep up is perceived as a burden
- Concern for downturn in adoption, support, maintenance by the core committers

Commercial solution

- Inadequate, untimely support

Lack of desired features

- Limitations in features and functionality within a given solution drive institutions to turn to additional solutions to fill the gaps. With multiple repositories to manage, institutions face added complexity and costs for staffing and infrastructure, increases confusion among researchers and tends to lead to repository silos.
- Bulk operations for ingest and metadata
  - Especially as part of quality assurance steps
- Flexibility
  - Configurable workflows
  - Configurable controlled access
- Preservation features and service integration; access persistence
- Robust metadata transformation, export, harvesting; support for linked data
- Support for diversity of content types, especially audio and video
- Search: federated; relevancy tuning
- Deposit for non-specialists
- Administrative features

Performance

- Processing collection content at scale
- Handling large files

Cost

- High licensing costs of CONTENTdm
- Limited features of Digital Commons/Bepress despite high licensing costs

What types of customization and configuration needs do people mention?

- Adding facets for browsing local terms
- Social media buttons - placing on the page
- Custom metadata templates by content type
- Workflows

- Front-ends to tailor UX to particular content types and themes (sometimes with elements required by larger institution)

What are people currently doing for preservation, and what would they like to do?

Preservation practices across institutions cover a spectrum, including:

- Doing nothing
- Backing up to local drives and servers
- Backing up to third-party cloud storage (Glacier, DuraCloud)
- Developing and operating a local preservation repository
- Depositing with community-based services (DPLA hub service provider, AP Trust, DPN, Chronopolis; note: no mention of LOCKSS)
- Integrating (local or hosted) asset management system with third-party preservation service (Archivematica, ArchivesDirect, Preservica, Rosetta)

People would like to:

- Outsource preservation infrastructure and services rather than develop their own
- Have a better handle on preservation metadata, ie., the metadata that accompanies content in a preservation context (such as technical metadata, rights metadata, digitization process history, curatorial history).
    - Often preservation metadata is generated and stored in parts of the system that are invisible to content managers and/or is intermingled with descriptive metadata.
    - People are aware of related metadata standards (such as PREMIS), but implementation is not rigorous or, more commonly, non-existent.
- Have a preservation layer support the repository that librarians, curators, etc. don't have to think out. Preservation should be automatic and invisible to those users of the system who are not responsible for preservation.
- Ensure that the preservation system is independent from the asset management repository, so that components of the layer can evolve and be replaced without impacting parts in other layers
- Preserve versions (or at least track/log changes), particularly of metadata and particularly for scholarly content
    - And make clear to a user of content which version they are using

What issues related to staffing affect working with repository systems?

In-house staffing levels for repository development and technical support tend to be very low

- Sustaining a repository and associated services is at greater risk when there is only one developer on staff in the event that the sole staff member leaves the institution before being replaced, without knowledge transfer and training, etc.
- Technical skills can be hard for a library to afford given limited staffing budget
    - Ruby skills can be especially difficult to hire; without them, it greatly limits an institution's ability to adopt and adapt Hydra
- Challenging to balance both software development and on-demand user support

Contract-based development and support is a mixed bag

- On the one hand, vendor/contractor services that support local deployment and customizations enable small institutions to implement and operate a repository that meets their specific needs
- On the other hand, these services tend to be very costly and recurring, and if the work is not executed or documented well, it is difficult-to-impossible to maintain and sustain over time

Dependence on central IT for infrastructure and tech support adds complexity, hinders flexibility and ability to be responsive to repository user needs as they arise

- Inflexible campus-level mandates can narrow options for selecting a repository solution and developing associated services
- Centrally-managed support of a repository service may be given lower priority over other “mission critical” campus administrative systems
    - Tasks are assigned to more junior IT staff in positions where there is often higher turnover, resulting in inconsistent and/or inexperienced support
- Requires additional, ongoing work on the part of a library to justify need, communicate requirements, work out service levels, etc.

Regional consortia / service providers are a good option

- Achieves efficiency at scale
- Tradeoff is limited customizability, which tends to make customers complain

What are peoples’ experience with tech support?

Hosted solutions (third-party commercial service provider)

- Poor responsiveness to feature requests and technical support needs
- Suspicion that small institutions, without sway or influence, don’t get timely assistance
- With proprietary systems, many changes must be carried out by the service provider

Provided by central IT (outside of organization, within same institution)

- Reliance on central IT to implement changes can lead to delays

On-staff (Internal to organization)

- Most ideal scenario

Decision-making

How do people make decisions about their repository solutions?

Cost

- What is the monetary cost of licensing, hosting, and/or outsourcing? How does the cost scale as collections grow?
- How much staff time is needed? Will staff need to be re-trained? Will new staff need to be hired?
- How long will it take to get the repository up and running?

Features

- Does the repository support the type of content we have?
- Is there an appropriate balance between flexibility and ease of use?
- Are there demos of the solutions that give a clear sense of how they work?
- What data management features are included?

Data and migration

- How difficult will it be to migrate existing data into the new repository? How much data will be lost?
- Once in the new repository, will the data be standardized and migrate-able?
- Can the repository support the size of our collection?
- Where will the data be stored? Will it be secure?

Sustainability

- How robust are the technologies and their supporting communities?
- Will the software continue to be supported and updated?
- How complex is the system to maintain?
- Is there a viable exit strategy?

Ecosystem

- Can the repository be integrated with other institutional software and technologies, including locally developed projects?
- Does the solution align with preservation needs?
- Do collaborating institutions use this repository?

Culture and values

- Is the repository software proprietary or open source? Some institutions have preferences for one or the other.
- Do stakeholders have experience or comfort with the repository and its associated technologies and communities?

What concerns or questions are there with the Hydra-in-a-Box project?

Sustainability

- Who will own the application component? Who will manage development? Is there strong leadership?
- What are Hydra-in-a-Box's long-term priorities?
- Will developers at adopting institutions be able to keep up with the pace of community changes to the code base?
- Will the Hydra-in-a-Box community offer support with both framework and application level issues?

Product design

- What are the demonstrated benefits and drawbacks of RDF and the PCDM data model and their associated technologies?
- Will the "out-of-the-box" solution be robust enough to meet the needs of adopting institutions, or will a (potentially expensive) hosted service be required?

DPLA

- What is the connection between Hydra-in-a-Box and DPLA, and how will this association benefit Hydra-in-a-Box users?

Impact on global communities

- Will Hydra-in-a-Box create a mass migration from other repository solutions? If so, what happens to those global communities who have invested in other solutions, and may not have the resources to migrate?

What are people's experiences with hosted solutions?

Positive experiences

- Hosts take care of "technical headaches," freeing staff to focus on collection management.
- Hosted solutions are beneficial for organizations that lack the staff or technical infrastructure to maintain their own solutions.
- Costs are explicit and easy to anticipate.
- Hosting services are accountable to client wishes.

Negative experiences

- Clients have a limited ability to make changes. They may have to wait a long time for hosts to make requested changes.
- Customizations options may be severely lacking. Hosts may not have the capacity to develop a feature that the client wants.
- Hosted solutions may not integrate with other software in our ecosystem.
- Privacy and access restrictions may not be robust enough to comply with laws or institutional policies.
- Tools for reporting usage statistics may be inadequate.
- Hosted solutions do not create the same sense of institutional pride as a locally supported solution.

Special considerations for decision-making

- Hosted services may be more or less expensive than locally maintained repositories.
- Are there effective tools for migrating data into the hosted solution?
- Is there a viable exit strategy if the hosting service disappears?
- Does the hosted solution enable clients to control and manage preservation data?
- It can take considerable time to migrate from a local to a hosted solution, even if the repository system is the same.
- Some institutions have an established preference for or against hosted solutions.

Recommended services

- Provide a flexible package of services, including infrastructure, applications, and/or operating system.
- Offer training.
- Provide good documentation of all available features, and effectively communicate with clients about new features.
- Provide enough customization options to suit client needs, keeping in mind that too many customizations can complicate future upgrades.
- Do thorough requirements gathering with clients.

- Migration is one of the biggest, most challenging components. Help clients migrate content, or hand that service to another trusted provider.
- Allow clients to retain their URLs.
- Allow a hosted backend and a locally managed front-end.

Multi-tenancy experiences

- Individual collections/institutions within a consortium need the ability to maintain separate settings, information, search results, controlled vocabularies, branding, and UI's.
- When using a hosted solution with a team, coordinating changes can be challenging.
- Some repositories do not support the ability to make changes across multiple collections, making data management very difficult.
- Multi-tenancy hosting can be very cost-effective because customers share platforms and basic infrastructure.

Migration

Why do institutions want to migrate from their current system(s)?

- Many institutions are frustrated by having to use multiple repository systems to manage all of their assets (e.g., separate IR and digital collection systems). The use of multiple systems can create headaches in locating and remediating objects, duplication, different approaches to metadata and cataloging, etc. For these institutions, the idea of moving all of their digital assets to a single system is appealing:

  > *"Clearly, I think that the tools we are using right now are coming to the end of their life or desperately need to be brought into the new world. [...] bringing things all together into one neat package would be great."* - Librarian, Large public university
  >
  > *"I think we are also looking to simplify, because right now we have many different ways of storing and managing digital objects. Being able to standardize that would be really great."* - Librarian, Large public university

- Whether or not they are using multiple systems, many institutions would like to migrate to a new system because they are frustrated with their current system.

What systems have institutions migrated from, or would like to migrate from?

- Bepress/Digital Commons, CONTENTdm, DSpace were each mentioned by multiple institutions. Wanting to migrate content from an Omeka site was also mentioned.

What are institutions' experiences with migration?

- Although one institution told us that exporting item metadata from CONTENTdm went very well for them, CONTENTdm was more often mentioned as being cumbersome to export from.
- More institutions seem to not have migration experience yet, but are concerned about the challenges involved in a migration effort, or whether it is really feasible at all.

- Mentioned more than once was the worry of how much effort would be needed to do metadata enhancement and remediation as part of the migration process.

What features would help with migration?

- Migration needs to be easy and riskless, in terms of losing data:

> *"It needs to [...] be super easy to migrate the data and not worry about losing any of the data or how it's configured in the database."* - Assistant Director, service provider

- It would be helpful to have support for doing metadata quality assurance and remediation before (or as part of) migration.
- Those wanting to migrate to a hosted service expect that the service will provide a path to migrate their data.

Storage

What is the size of digital collections and objects that are stored?

For those **institutions serving small-to-medium sized communities** with very few staff dedicated to digital collection management, numbers of objects tends to range in the tens of thousands and storage size is relatively modest (fewer than 10 TB):

> *"We recently got CONTENTdm, which is where we will be migrating [20,000 objects already online] in addition to about 20,000 more objects that are digitized and ready, but not currently available to the public."* - A county public library

> *"We have 150,000 images maybe in CONTENTdm and some oral history audio files."* - A small, private college

For **medium-sized institutions** with more resources devoted to collection creation and management, the number of objects heads into the hundreds of thousands and storage size grows to low-tens of TBs:

> *"Right now we are taking about 20 TB of disc, maybe 30,000 items and 350,000 preservation binaries."* - A private university

> *"We have close to ... 12 to 15 TB at this point of archival digital material whether that's coming born digital or we've digitized."* - An independent research archives

> *"We have somewhere between 300,000 and 400,000 preservation copy images plus all the different access manifestations, so probably pushing a million photographic images."* - A state historical society and DPLA contributor

> *"We are going to be migrating away from DigiTool soon. That resource has 75 digital collections, which include over 27,000 images and it's constantly growing. We have added about 12,027 new items last year to that collection."* - An independent research library

> *"Our preservation server is actually not as large as you would think right now. We have about 4-4.5 TB of data being backed up right now through our preservation server and Glacier storage."* - A regional consortium and DPLA hub

> *“[For one initiative] we have 173 contributing organizations ... who have contributed a total of just over 45,000 things, but then that translates into 240,000 digital [file] assets because we have lots of compound document things, like oral histories and transcripts, books, booklets and all.”* - Another regional consortium and DPLA hub

> *“[One customer has] a 3 TB DSpace. The sheer volume of that was a little bit challenging. Theirs was so out of the box because they had it hosted with their IT infrastructure that aside from the sheer size of it, it really wasn't that hard for us to [manage].”* - A state-based hosted service provider

For those **larger institutions serving large communities** who devote, relatively speaking, the most resources to digital collection management, the numbers of objects can reach 1+ million and the number of TBs approaches mid-to-high tens of TBs:

> *“For university archives content we have maybe about 60,000 or 70,000 objects that were digitized from the physical collection. … being photographic to textual documents. For the institutional repositories -- we’re running DSpace currently -- we have about 45,000 individual records with about 50,000 [files]”* - A large public university

> *“It depends on how you count objects in CONTENTdm. … All the OCR lives at the article-level so we have around 20 million newspaper items on one CONTENTdm server. Our other server is a little bit under 2 million for digital collections.”* - Another large public university

What are some of the special issues around content storage and collection size facing interviewees?

Growth projections are challenging to make with certainty

- Most don’t really do it
- One regional consortium reported 1-2 TB growth each year
- Often large collection projects come up without much advanced planning and proceed without a robust storage solution in place, so files are stored on external hard drives, adding to collection management challenges

AV content is a hot topic when it comes to storage

- Numerous interviewees mention growth of audio/video content and the associated storage concerns in the same breath:

> *“What's funny is when we tell them how much we have just of video right now, all these vendors, they kind of do that blink and pause and go, ‘That's a lot.’ It's several terabytes. I don't know if we're at 20 terabytes, but it's getting there. It's a lot, and we're going to be adding more. We have tons of images as well.”* - A medium-sized museum

> *“We have a lot of audio and video that we deal with. A subset of that 12 to 15 terabytes is also sort of the stuff that we've digitized whether it's audio and video in the collections or photos.. We generate our own born digital and video and audio content, which is several terabytes worth of that subset.”* - A small independent archive

> *"I manage about 3,000 oral history interviews. Those are audio files and transcripts mostly. In the future there will be a lot more other audiovisual stuff that we're just working on digitizing now. So we try to build our infrastructure for that."* - A state historical society

Location of file storage can matter

- On the one hand, some institutions specifically seek repository solutions that include geo-diversity of redundantly-stored content
- On the other hand, some institutions must comply with government mandates requiring that academic products, including theses, dissertations, and research data, are not stored outside of particular geographic region, such as a national border

Tension arising from the promise of an affordable hosted service

- Low-cost cloud storage is attractive for those with content to manage but insufficient infrastructure, yet the added hosted service fees can make the promise out of reach

Managing large files in the context of hosted service introduces challenges

- For content made available through a hosted service, some institutions must maintain high resolution copies locally for ready access when fulfilling patron requests, because the high-resolution copies are not readily accessible via the service provider. This can result in version control / file tracking problems.
- Institutions responsible for managing research data need mechanisms to support large file transfer (both repository deposit and dissemination) outside of browsers and http, which are not reliable for such transfers

Collection management

What are current pain points related to collection management?

- A common challenge for institutions is being able to easily identify or inventory all of the digital assets they have. Institutions often employ multiple repository systems to manage their assets and being able to inventory and report statistics across systems is difficult.
- Existing systems don't enable administrators to perform all of the functions on items managed in the systems that they would like to perform.

What collection management features do institutions like in their current repository systems?

- Preservica: batch editing, batch ingest; lot of administrative tools
- DSpace: enables you to create hierarchical collections
- EmbARK: collections manager that does a good job

What collection management features are institutions looking for?

Better reporting capabilities

- Aggregators want to be able to easily show providers statistics related to the content they provided and how it relates to all the content held by the aggregator.

- It is often a requirement that repository managers submit quarterly or year-end reports to their management or institution administration, and an easy way to generate repository statistics would be helpful.
- More generally, being able to easily track the repository and collections in terms of objects, file types, and sizes would help with day-to-day operations.
- Quality assurance reports (for metadata) were mentioned because it is common to use a discovery interface as a QA interface and that can be problematic (it is better to look at the metadata before it goes into the Solr index).

Many-to-many relationship between items and collections

- Multiple interviewees expressed desire to be able to put an item into multiple collections.
- The lack of ability to put an item into multiple collections is seen as a notable limitation of existing systems (e.g., CONTENTdm).

Managing all digital assets in a single system

- Strong desire to be able to manage all institution assets in a single system:

> *"Having some way of at least knowing the digital assets that we have would be fantastic, so maybe something like Hydra would be a way to aggregate that across all our different systems."*
>
> *"It'd be great if there were really one product that could do all of those things, do the preservation and inventory management aspect of it as well as the enterprise aspect that they like to call it."*

- No good solution now for tracking and describing content for an institution with archives and a museum, and the worry is they will add another sub-optimal tool to the ones they are already using.

Integrated tools

- Appeal in having integrated administrative tools:
  - Batch editing, batch ingest; schedule, review, update file; watermarking; licensing and payment; rights and protection information; making clips from larger content items (video).
- Would be nice if the system could also be a physical collections manager
- Ability to do accessioning for special collections, an inventory at the beginning stage of scanning.
- Be able to append an appraisal or collection development policy that was in effect at the time of ingestion (which would be inherited down to file, folder, or item level).

Appearance and branding customization

- For institutions hosting content from other institutions, the ability to easily visually customize and brand different collections would be very useful.
- Even within a single institution's repository, the ability to group collections and provide a custom visual identity to them would be useful:

> *"The [interviewee's university] has a lot of things dealing with ski archives and multiple collections that deal with skiing as a topic. It would be cool to have a portal, like, 'If you care about skiing, go here.'"*

Content types

Which content types need to be supported? Which are the most important to interviewees?

Institutions tend to manage a wide variety of content types across the spectrum of library & archival collection materials as well as academic and research-based publications and data.

Images represent the content type most frequently managed by interviewees (17 excerpts), followed by:

- Audio/visual (12)
- Articles (9)
- Theses/Dissertation (9)
- Archival collection (8)
- Text Documents (5)
- Data sets; Newspapers (3)
- Web archives; Geo; Manuscripts; (2)
- Oral histories; Email; Books (1)

Content types most frequently co-occurring with Images are:

- Audio/visual (17)
- Text Documents (10)
- Theses/Dissertations (9)
- Archival Collection (7)
- Newspapers (5)
- Manuscripts (4)

Top content types that present "Challenges" in "Current Repository System" from the Hydra-in-a-Box user survey are:

- Audio/visual (8)
- Newspapers (4)

Top content types representing strong Future Needs from the Hydra-in-a-Box user survey are:

- Audio/visual (10)
- Archival collection (6)
- Data sets (5)
- Articles (4)

Conclusions drawn from interview excerpts:

- "Compound documents" -- oral history audio with text transcript, publication with data, multimedia works -- are significant, especially to the institutions with more established digital library programs, and reflect the emerging hybrid nature of content being generated and collected for preservation and re-use today
- Institutions are collecting and creating an increasing amount of audio and video content, and struggling to manage it due to storage constraints and inadequate delivery options within common repository solutions. Preservation of this content is of particular concern, because of the large file sizes and corresponding costs to store and manage. Oral history collections and event recordings, as well as AV within mixed archival collections are very common.

- Managing content from archival collections is a challenge for many institutions, because commonly used repository systems, such as DSpace and CONTENTdm, don't support content and metadata that is hierarchical in nature very well.
- Content generated by students and faculty is a high priority for academic institutions, especially theses, dissertations, research data and articles. Most interviewees are not managing significant amount of research data, but it is a growing trend and there is awareness of the special concerns regarding open access mandates and scale of "big data"
- Newspapers, both historic and more contemporary, are a special category of content: due to large page size and their serial nature, common repository solutions don't handle them well but, specialized applications do tend to work well. Demand for access by genealogists is high.
- Content formats such as email, software, disk images, GIS content and web archives do not represent a significant amount of the content types that institutions are managing in repositories.

Workflows

How do workflows depend on user roles, approvals, and authentication?

- Users require multiple user roles with different privileges
- Approval may be needed at different steps, not simply to "publish"
- As an example, one user outlined four roles:
  - Student uploads files and matches to an OCLC id to retrieve an ILS bib record
  - Digitization manager does technical work like implementing OCR and technical metadata
  - A specialist librarian created descriptive metadata
  - Metadata librarian approves final record and publishes
- The need for actual approvals in the above scenario may be mitigated by the implementation of tools to track workflows rather than actual approvals
- Sometimes reviewers or approvers need to be outside the library: faculty or museum partners are examples
- Another user articulated need for "save for later" "submit for approval" and "publish"
- In the case of self-submission or faculty submission, the use of proxies for uploading is needed (see section on self-submission above)
- Interoperability with campus authentication systems is necessary and the ability to pull identifying data from these systems for pre-population of records is desired.
- Some permissions or roles need to be granted for particular collections and not for others (i.e. a user can upload to one collection but not another)

How does system performance affect workflow support?

The system should be able to handle multiple back-end users simultaneously while also supporting front-end search and use. Typical workflow scenarios might include several users creating and editing metadata, while others are uploading objects and still others are generating reports or using the front end.

What types of overall digital object management are needed?

- Users would like to have access to all of their digital objects from one application: high resolution master or preservation copy, and access copy. Existing systems sometimes either do not facilitate the storage of high-resolution/master copies or an institution has implemented a separate preservation system and the two are not tied together by anything other than metadata/identifiers
- Several workflows specifically mentioned either digitizing objects or uploading digitized objects before the creation of descriptive metadata
- Workflows often differ based on type of object digitized. This is to accommodate the creation of object specific structures and data such as OCR text, TEI, newspaper structural metadata, and audio/video data
  - Text: metadata from bib records; images are uploaded and then OCR or TEI is done. Sometimes by staff, other times automated; structure like page sequences are sometimes done manually
  - Newspapers: one user mentioned using a product called Paperboy specifically for the coordination of metadata and images of newspaper pages
  - A/V: video was digitized by a vendor and a large amount of non-automated technical data (such as the type found in PBCore) was created in a spreadsheet and had to be uploaded along with objects

What types of overall digital collection management are needed?

- Users want to be able to manage certain aspects across collections, such as the implementation of controlled vocabulary or the ability to edit metadata
- It should still be possible however to manage the same aspects collection by collection
- Typically when a collection in an archival setting is created there will be a lot of initial activity to create records and then a slow down. After the initial burst, activity will be sporadic for many collections.

What kinds of project management tools would be useful or desired in the repository admin interface?

- Users want to be able to both track steps in a workflow, and capture some nuanced curatorial information in notes. Information like whether or not metadata is complete or rights are undetermined.
- Others mentioned being able to give feedback through notes to the creator of metadata if remediation is necessary
- All would like records of these interactions stored in the system
  - Some users actually used online agile development tools to track these kinds of things

Do users prefer a web-based tool for management or a client?

- A web client would enable users to share access to various tools and interfaces with external partners
- If an interface is web-based it should still be fully featured. Users were not happy with systems that required some kinds of work to be done in a client and others to be done in a web-based tool

How does digital preservation figure into workflows?

- It would be ideal if preservation were not carried out in a different system so that the digital object creation/upload process was a seamless one with preservation

- Only one user had a relatively robust preservation workflow which involved preserving born-digital material. The workflow included pre-ingest activities like
  - Quarantining data
  - Virus scanning
  - Scanning for PII
  - Characterization such as creating technical metadata for future emulation, deriving checksums and other fixity features, hardware/software environment etc.
- This user also was functioning as an archives, which meant that the users submitting material expected to be able to continue to add/update and access the data. It was documentation of work, not cultural heritage artifacts
- Their workflow also involved a lot of media-specific practices based on dealing with a variety of carrier media (floppy disks, VHS cassettes, hard drives, etc.)
- They were interested in making certain pieces of their preservation metadata public, such as assessments, quarantining, and appraisal decisions

Metadata

What types of metadata creation workflows are prevalent?

- Creation of metadata in a spreadsheet to be uploaded to a repository
  - Spreadsheets are both loved and hated. When combined with tools like OpenRefine powerful metadata management and validation can occur. Interviewees mentioned being able to control values for particular fields, match values to URIs or vocabularies, to have multiple users share a spreadsheet and to create more than one metadata format from a single source (i.e. generate both MARC and MODS, for example). Spreadsheets are disliked by some for their awkwardness and unwieldiness, but overall reactions were neutral to positive.
- Creation of metadata in bulk, but uploaded to a repository via batches of XML files
  - This metadata may or may not be created in spreadsheets, but its ultimate format is XML records. Metadata may be created by a vendor and transferred via BagIt bags, scripted from another XML standard like EAD, or created through some other input like a webform (possibly even one at a time)
- Creation of metadata for objects one-at-a-time by staff
  - Many of the same issues of quality control, validation, and authorization are relevant to singe item workflows as well as spreadsheet workflows, but the mechanics are different. Many variations and aspects of these workflows will be explored throughout this analysis.
- Creation of metadata by users/submitters
  - Mostly these workflows involved single-item submissions, but some involved multi-work submissions (i.e. a submitter filling out a spreadsheet of metadata and submitting for ingest or review). Aspects of this workflow will also be explored in other questions of this analysis.
- Many users spoke of the need to support multiple types of metadata creation at once. Workflows for bulk ingest, single item creation, and self-submission might all be needed in the same system.

When a user creates metadata in spreadsheets, what does their workflow look like?

- Workflows may include an intermediary step where spreadsheet is transformed to XML

- QA may or may not happen in the spreadsheet. May not happen until available in repository
- Spreadsheet may contain any kind of metadata: descriptive, technical, preservation, structure
- Spreadsheet may be the result of vendor digitization (i.e. not developed in-house)
- Multiple users may have contributed to the same spreadsheet over time
- Creators of data in spreadsheets may not be librarians or work for the repository. They could be contributors from other partnering organizations like museums, academic departments, etc.
- Spreadsheet cells can have controlled vocabularies to help with data quality and normalization
- Data in spreadsheet can be easily normalized or analyzed. Activities like "fill-down" "drop-down lists" "analyze" "play around" and "correct" were all mentioned in reference to OpenRefine in particular.
- Usage diverged on how metadata that originated from a spreadsheet was edited once it was ingested into a repository. Some were able to update spreadsheets and re-ingest, while others were limited by built-in editing tools in a software, which may or may not include bulk editing operations.
- At least one user mentioned specifically not wanting or needing a repository to support all of the metadata analysis and normalization tools that OpenRefine offers. Instead, having an external tool to analyze and normalize data, which specialized in these activities, offered an opportunity to audit the data outside of the repository. It was seen as not really a "core" activity of the repository.

When a staff user create metadata one item at a time through an ingest form, what does their workflow look like?

- Users expressed a desire for configurable workflows based on project/user/item types, and for a repository to support multiple interfaces for upload of objects
- Some workflows included importing or pre-populating records with data, specifically from MARC bib records via OCLC number, or through XSLT from EAD finding aids
- Controlled vocabulary lookup while entering data in text fields was mentioned numerous types
- Several users mentioned manually creating structural metadata for complex objects like books.
- TEI Lite was also mentioned as specific markup language created for book objects
- The need for features related to workflow management and approval was mentioned often. Features like "save for later", "submit for approval," "approval queues or notifications," "administrative notes/tracking of progress" were all noted.
- One user reported using a specialized tool for the creation of library metadata (Pageboy) which allows the creation of metadata on a page by page basis
- Another user reported the use of an open-source tool for metadata creation that was completely outside of any repository workflows

When an external user creates metadata, what does their workflows look like?

- Workflows for external self-submitters are rarely monolithic. An ideal system would support configuration of things like metadata requirements, authentication, interoperability with faculty portfolio software among others, and ability to do batch uploads.
- Some self-submission workflows need to be able to support uploading by proxy. Students and admins are often made proxies for departments, meaning they need to upload documents and then have them reviewed by departmental faculty before they available. Permission by the proxy may be needed for an individual faculty member or an entire collection/department.

- Multiple users required interoperability with external faculty scholarship tools such as portfolio software, ORCID ids, citation export, etc. These can have an impact on the metadata creation
- Validation of self-submitted items was desired. This validation would be based on local configuration and include variables like completeness, and use of required vocabularies (which would be enforced in the interface through things like drop-down menus for selection).
- Workflows that included reviewing records by both staff users and self-submitters (in the case of proxy uploading) before publication was mentioned. Suggestions in this area including allowing records that met validation requirements to be published without review, and for review to have bulk reviewing features.
- On the other hand, there was a desire for uploads by self-submitters to be published "right away" or at least to have access to a DOI immediately
- Workflows for ETD submission differ from those for other academic self-submissions. Details here were not evident in the transcripts analyzed, but presumably other tools like Sufia and Virio, as well as other transcript excerpts will shed light on this.

What types of forms/interfaces are desired by users for metadata work?

- Users mentioned needing metadata creation and workflow forms that were intuitive and easy to use, especially those that need to be shared with external users in the cases where a partnership is in place between the repository and another organizational unit like a museum.
- "Feature-rich and streamlined"
- Features like auto-suggest and autocomplete with controlled vocabularies were desired.
- The pre-population of records with information from other systems (i.e. MARC records) in an interface was desired
- Workflows might involve multiple users working on the same record (i.e. one cataloger does description and another authority control), so it is important that more than one user can access records before they are published
- Approvers of records want to be able to see queues of records for approval, give feedback to creators of records, and approve them in batches
- Interfaces for creating metadata should be configurable for the type of material, user, or collection. For example, if a user is uploading audio material and the repository wants to record specific information from the PBCore schema, those elements should be available. If the user is a self-submitter from the Geography department and the collection being submitted to requires GIS coordinates, those elements should be available, etc. These configurations should be controlled by the repository admins.
- Appropriate technical metadata should be created automatically upon ingest depending on the type of content. However, some users expressed a desire to be able to see and edit that technical metadata.
- Mostly users want web-based forms and tools for dealing with metadata, but would like the option to directly access the native data (i.e. see the XML of records)
- One user mentioned the use of a third-party tool (SIMP) for metadata creation. This allowed them to standardize metadata creation across all their systems since metadata created there could be exported for different systems.

How do users want to edit metadata?

- Support is needed for both bulk and individual item metadata editing

- Almost all users mentioned the need for better bulk metadata editing
- Related to the need for bulk editing is a need for better tools to analyze and spot problems in metadata (as well as for various kinds of strategic planning -- related to reporting). These include powerful methods for searching, sorting, and refining sets of records.
- Many mentioned using OpenRefine for this type of work if they were able to export metadata.
- CONTENTdm offers some simple tools for bulk editing, but these were mostly unsatisfactory. Users can completely rewrite all instances of a field within a collection, but could not do cross-collection editing, or alternately, editing of only smaller subsets of data
- Some users reported being able to export metadata out to a spreadsheet, correct it there, and then re-upload, but not all were satisfied with this workflow
- Examples of editing/analysis that users would like to do in batch interfaces includes:
  - Normalizing values (i.e. finding close matches and making them the same)
  - Normalizing formatting of things like dates or punctuation
  - Reconciling with controlled vocabularies
  - Searching for values with typical errors and correcting
  - Wholesale replacement of values (i.e. updating a rights statement in all records)
  - Validating the records adherence to internal standards (i.e. which fields are required, which vocabularies are used)
- Bulk edits to metadata may need to be done before data is live. If it is uploaded in bulk, it may need to be edited/normalized/validated in bulk before it is published.
- One user expressed a desire to be able to edit not just descriptive metadata, but also technical or structural metadata directly. Or at the very least to see it.
- Users want to have some form of version control over metadata. To be able to see when it was created and by whom, possibly to store previous versions of it.
- Mostly users want web-based forms and tools for dealing with metadata, but would like the option to directly access the native data (i.e. edit the XML of records)

What kinds of support are desired for controlled vocabularies?

- Users want to be able to validate metadata fields against existing standard controlled vocabularies and authorities
- They also want to be able to define their own local vocabularies
- Support for searching and auto-completing terms from those vocabularies was mentioned
- It should be possible to implement a controlled vocabulary either across the entire repository or to a single collection
- Controlled vocabulary support should include URIs for terms when possible. Manual creation of URIs is not preferred.

Where do users get metadata from? How is metadata shared and re-used between systems?

- Metadata in repositories doesn't always originate there. It can be migrated from other systems, especially MARC catalogs. The easy transfer of data between these systems is crucial
- When a project is sent out to a vendor for scanning, records of technical metadata may need to be ingested along with objects

- Metadata may come from other systems in the form of spreadsheets, BagIt bags, batches of XML, OAI feeds, Z39.50 protocols, ILS bib records(?)
- Some of the metadata retrieved from other systems will be partial. It will need to be further enhanced before it is published in the repository
- Syncing metadata across various systems is a problem. If metadata for digital objects comes from finding aids or MARC records, for example, keeping updates current between all the sources is a problem
- Alternatively, more than one user mentioned creating metadata in a repository and exporting from that records for another system. For example a self-submitter may upload an ETD and a MARC record needs to be created from it.
- When data is coming from an external source, it should be possible to periodically "refresh" it

How do users want to get and use metadata outside of the system?

- Users need to be able to export metadata multiple ways
- If they are exporting metadata to a feed, such as an OAI PMH XML feed, they need to be able to configure how the metadata in the feed will appear
- They may want to map the metadata to another standard for export
- Multiple users mentioned being able to download metadata as a delimited file for use with OpenRefine for analysis or editing. One user expressed a desire specifically to be able to do various analysis outside of the repository as a "double-check" (i.e. the metadata is not appearing a specific way because of a repository setting or interface)
- Metadata exports should be in bulk. While it may be useful in some workflows to download individual files, downloads of entire sets of metadata are more important.

What kinds of metadata are needed/created/used?

- Multiple descriptive metadata schemes were mentioned, mostly MODS, DC and QDC. One user mentioned VRA Core and GIS metadata. Several mentioned a desire for the system to be metadata agnostic.
- Many users mentioned technical and structural metadata, but only one mentioned a specific schema (PBCore for video)
- Preservation metadata was mentioned several times, but not by specific schema. References were made to checksums and provenance elements.
- Users have difficulty dealing with metadata coming from museum-based projects because it does not align well with library standards
- Most users did not mention creating local elements directly, but implied the existence of such elements in discussion of data quality and mapping data for aggregation
- Support for mapping to simple Dublin core was mentioned in connection to Omeka.
- Support for mapping to the DPLA MAP was mentioned in relation to aggregation
- Specialized metadata for particular types of objects such as TEI markup, or OCR output was mentioned as part of metadata workflow and object structures
- Only one user (from an archive environment) mentioned the need to support hierarchical metadata
- Same user also mentioned needing to store the technical metadata necessary to drive future emulation of objects.
- One user mentioned needing to handle the complex relationships between records for serials.

What kinds of validation and quality control of data are desired?

- Configurable validation based on rules set by the user including
  - Completeness (i.e., a value is present in a specific field)
  - Correctness (i.e., matches a vocabulary)
  - Format (i.e., EDTF dates)
  - Schema-compliance
- Validation and quality analysis needs to happen before records are published
- Quality analysis and bulk editing share a lot of characteristics such as the ability to look for completeness, adherence to standards etc., as well as the ability to look at both the full data set as well as subsets
- Users should be able to set up validation so that if a record from a self-submitter validates it is automatically published. Otherwise it is held in a queue for review.

What is the intersection between metadata creation and digital preservation?

- Users did not have a clear articulation of needs around digital preservation. Most were aware of preservation concepts, but did not have a clear list of activities they would like to be able to perform
- When it was mentioned users reported either:
  - The storage of some very basic preservation metadata elements, which would be covered in the PREMIS schema
  - Duplication of objects in a separate digital preservation system (more on preservation will be covered in general workflow)

What kinds of metadata aggregation activities need to be supported?

- Not including large-scale aggregation activities, typical repository users need to be able to incorporate feeds of data using protocols like OAI PMH, Z39.50, an API with JSON-LD, (and potentially ResourceSync, though it was not mentioned)
- Users need to be able to publish feeds of data for others to aggregate
- CONTENTdm, Repox, and Drupal’s feed harvester for Islandora were both mentioned as sources or aggregators of OAI feeds
- Users need to be able to map their existing data to other standards for publication in a feed
- Users need to be able to configure what data and data collections will appear in their feeds
- One user would like to be able to store metadata crosswalks such that they can apply them to multiple data feeds (this was a user who was not using an external XSLT, but an internal mapping tool in their repository)

Discovery/UI

What discovery and UI features would people like to see?

Discovery environments

- Many interviewees want to support multiple discovery layers for their repositories.
- Within any given UI, interviewees strongly desire the ability search across multiple repositories, subsets of repositories, exhibits, and external data streams.

Customizability

- The ability to create custom theming and branding is essential. Sub-collections within a single repository may need distinct branding.
- Some interviewees want the ability to incorporate custom features, such as digital humanities tools, annotation tools, etc.
- Admins want to customize:
  - Which metadata fields display in search results.
  - Which metadata fields are displayed alongside digital objects based on content type.
  - Relevancy and weighting in search algorithms.
  - Which facets appear on search results pages, and how facets are grouped.

Search interfaces

- Interviewees emphasized the need for powerful index and keyword search. They reported that users demonstrate a preference for Google-like interfaces with emphasis on keyword search and streamlined display of search results.
- A couple of interviewees suggested a grid view for search results, showing large image thumbnails with minimal metadata.
- Facets are another important discovery tool.
- At least one interviewee wants an advanced search feature, which would allow users to build complex queries across specified repositories and APIs.
- One interviewee suggested showing results that match any (as opposed to all) query terms as a way of encouraging users not to abandon searches.
- One interviewee expressed the importance of facilitating discovery based on researchers.

Related materials

- Interviewees want to direct users to related materials, based on topic, collection, author, time period, etc.
- They also want to incorporate links to relevant external materials, such as related Wikipedia content.

Help features

- Several interviewees pointed out the necessity for meaningful and substantive feedback when searches return zero results, and for suggestions/corrections for misspelled search terms.
- At least one interviewee wants to incorporate discovery aids to teach users how to use the interface effectively for research.

Interacting with digital objects

- Interfaces should be tailored appropriately for different content types. For example, readers with zoom, rotate, and pagination features should be used for text- and image-based works. Finding aids should include content lists and links to digitized content.
- UIs should provide full-text search for relevant items, and access to transcriptions or translations.
- A few interviewees had specialized use cases for their content. One wanted users to see different versions of textual files and track changes between versions. Another wanted to allow users to make clips of audio or video files.

Collections

- Many interviewees want to create and present collections along with contextual and descriptive information specific to that collection.
- Search results should include collections as well as individual items.
- Users should see which collections individual items belong to.
- One interviewee expressed a desire to showcase collections on their homepage.

Re-use

- Permanent URLs to digital materials are essential for re-use. Institutions also need to retain permanent URLs when migrating from other repositories.
- Presenting users with clear rights and reuse information is also important.
- Interviewees suggested a couple of ways to give users access to high-quality copies of digital objects, including downloading individual objects and YouTube-style embed links.
- A couple of interviewees suggested an interface for users to request copies of digital assets or permission for re-use.

User contributions

- Several interviewees expressed a desire to collect some form of user contributions, including comments, annotations, translations, and OCR corrections.
- The system should notify admins when users submit contributions.

Metadata

- One interviewee wants to allow users to bulk-download library metadata.
- One interviewee wants to include hyperlinks in metadata fields such as descriptions and rights. They need to retain these hyperlinks when data is exported.

Access

- SEO is an important component of discoverability for many interviewees. Interfaces should be optimized for search engines.
- Audiences for UIs include K-12, undergraduate, graduate, and faculty.
- At least one institution needed international localization and language support.

What social and personalization features would people like to see?

Social media

- Social media is a part of many interviewees' engagement strategies.
- Users want to share items and collections on social media.
- Admins want to control which social media buttons are available to users and customize how they work.

Personal accounts

- It should be easy for users to create a personal account.
- Users want to curate personal collections of digital library materials.
- Users want to annotate items within their personal collections.

Special access

- Donors want special access UIs to view their own materials without restrictions.

Are people interested in exhibits, and if so, what role should they play?

Interest and current use

- Many interviewees are currently using, or plan to use exhibitions to engage broader audiences.
- Creators of exhibits include librarians, archivists, curators, academics, students, and collaborations between the aforementioned groups.
- Exhibit solutions must allow creators to work without developer intervention.
- One interviewee stressed that exhibits must support large collections of materials.
- Most interviewees use Omeka, and would need a migration path from Omeka to any new solution.

Integration with repository

- Interviewees want the ability to build exhibits around repository collections, and to integrate exhibitions into repository search results.
- Current solutions such as Omeka do not integrate well with repositories.

Additional desired features

- Ability to easily migrate data in and out of exhibits
- Transcription
- IIIF integration

Usage analytics reporting

How important are usage analytics reporting capabilities?

- Many institutions consider reporting features to be a must-have in a repository system.
- There are two distinct forms of reporting that are often discussed together:
  - Usage analytics, or reporting on objects and collections that are available to end-users.
  - Collection statistics, or reporting on the size and characteristics of objects and collections in the repository. Collection statistics are covered in the Collection Management section.

Why is usage analytics reporting an important feature?

- Analytics can serve as a communication tool for repository managers/librarians.
- It is common for repository managers to need (mandated or not) to report on usage analytics to various stakeholders:
  - *Management and administration*: It is often a requirement that repository managers submit quarterly or year-end reports to their management or institution administration, to demonstrate impact, that repository-related efforts are worth the time.
  - *IR contributors*: Providing faculty and researchers with analytics related to their deposits is very useful in encouraging those users to become depositors, and to maintain satisfaction with the IR. “Individual metrics for individual depositors are really important.”

  - *Content providers*: Aggregators often want to give their content providers analytics so the providers understand what content of theirs has been aggregated and how it is being used.
  - *Departments*: Grouping analytics by department would enable repository managers to communicate with individual departments about items and usage relevant to them. “For all papers from this department or for all content from X, what are my bulk stats?”

What usage analytics features do institutions currently use and like?

- Although current capabilities were mentioned as limited, most institutions expect basic metrics such as download counts, view counts, and geolocations of users to be available in a reporting feature.
- Bepress/Digital Commons has additional features that are seen as useful by several interviewees:
  - Contributors have an option to receive regular emails with statistics about their deposits. They use this information for their own reporting needs.
  - Contributors can see a map that shows in real time where people in the world are accessing their deposits.

What usage analytics reporting capabilities are institutions looking for?

- Basic, commonly used metrics are expected (by object, by collection, by month, etc.):
  - Number of downloads
  - Number of views
  - Who (geolocated) is accessing items and collections
- Going beyond those basic metrics was important to several interviewees:
  - Understanding how an object is used (e.g., if and where it was cited, which tweet mentioned it, a blog post referencing using an object, how it was referred to in social media).
- Enabling depositors to optionally display download/view metrics for their items.
- Some institutions are interested in assessing user experience and are looking for ways to do that beyond simple count metrics. For example:
  - What kind of queries are used to search for things? How do end-users use facets? What do end-users click on? How do they find things in the Blacklight interface?
  - How long does it take a depositor to complete a deposit? How often did they have to use the FAQ to complete the deposit? What are the frustration points?

What are current pain points related to usage analytics?

- Many institutions currently use Google Analytics as at least a base level of reporting, but find it of limited usefulness.
- Institutions rely on built-in reporting features from their current systems and/or write custom code to get statistics on things they don’t get from built-in features. This is a problem because:
  - Writing custom code is time-consuming and inefficient when we know many institutions are interested in the same reporting features
  - Built-in reporting features don’t provide data for all the things institutions are interested in
  - When built-in reporting breaks, it is difficult to understand why
- Commercial analytics products are too expensive for some.

- An Islandora user say they have “a terrible time giving any sort of statistics to our partners, just like simple things like page views.”

What concerns do institutions have related to usage analytics?

Privacy

- It was a concern that trying to understand more about who accesses repository objects raises privacy issues.

Limitations of common metrics in use today

- Metrics such as views and download counts don’t help understand whether a resource was useful to an end-user, or what their motivation was for downloading it.
- Even using Google Analytics data can be problematic. For example, every new Solr facet a user adds to a search creates a unique URL.

Interoperability

Where are there opportunities and needs for interoperability with other systems? What APIs are important?

Which systems should be interoperable?

- Repository backend, possibly multiple repositories
- Data management UIs, including those for institutional repositories
- Preservation layer
- Index/search
- Multiple discovery UIs
- Exhibits
- External metadata streams, such as Elsevier Pure and Wikimedia
- External indexing services, such as WorldCat and academia.edu
- Identity management systems
- Library catalogs
- Usage reporting statistics
- Systems for managing conservation and loan information
- Systems for users to request copies of digital assets and permission/license for re-use
- Notification system to inform users when there is new content relevant to their needs or interests
- Research data systems, such as Electronic Labs
- IIIF servers
- RSS feeds

Specific use cases for interoperability and APIs

- Read/write data from the repository and index through an API.
- Build front-end UIs, searches, and exhibits around subsets of repository data, possibly pulling from multiple repositories or external metadata streams.
- Leverage a variety of metadata standards, controlled vocabularies, and digital proxies.

- Easily migrate data in and out of the system, and crosswalk data between different schemata.
- Harvest data using APIs.
- Output frontend data as JSON-LD.
- Determine whether records have changes through an API.

Specific systems mentioned that would ideally be interoperable with the repository:

- academia.edu
- AP Trust
- Chronopolis
- Drupal
- Electronic Labs
- Elsevier Pure
- HathiTrust
- Internet Archive
- Mirador
- Omeka
- Open Collections
- Open Library Environment
- Preservica
- PubMed
- Spotlight
- Symplectic Elements
- VIVO
- Voyager
- Wikipedia
- WorldCat
- WordPress