HARVARD UNIVERSITY
Information Technology

HARVARD
LIBRARY

Digital Repository System (DRS) Futures Summary Report on Focus Group Sessions

Table of Contents

Principles, Planning, and Organization

The Digital Repository System (DRS) Futures team offered a series of focus groups as part of the stakeholder engagement effort. Community members were invited to have facilitated discussions about their workflow, vision for the future, and current pain points for the long-term preservation of digital content.

The DRS Futures team sought to engage underrepresented communities as well as already engaged participants during the focus groups. Discussions were organized around specific topics but were all designed to provide participants an opportunity to discuss their thoughts on the functions of the core DRS preservation system, dealing with concerns such as content submission, metadata description, curatorial and administrative management, preservation outcomes, and reporting. Focus group discussions primarily addressed the means of ensuring the ongoing archival integrity, authenticity, accessibility, and usability of Harvard's digital collections. Additional topics that came up during discussion were recorded to be shared with other Library groups and initiatives as appropriate.

The DRS Futures Team's objective was to elicit stakeholder needs through both pre-conceived prompts as well as unprompted, unanticipated comments. This can be most productively achieved through clear, consistent expectations for how to participate in discussion in a respectful, balanced way. By using a lightweight version of the "stack method," we aimed to create a more democratic discussion and amplify under-represented perspectives, in alignment with the DRS Futures Principles of Collaboration.

Participants were reminded that the Futures project is concerned with revitalizing core digital preservation infrastructure. As shown in Figure 1 below, the DRS sits at the center of a larger – and ever-evolving – ecosystem of independent but interoperable systems and services providing complementary stewardship functions.

Since these services have been carefully designed to work together seamlessly, it is easy to assume that they are all part of a single system. However, this is not the case. In particular, the various Library delivery services for images, streaming media, and page-turned objects are external to the DRS proper. Therefore, delivery services changes are currently out of scope of the DRS Futures Project (beyond the

necessary changes needed to integrate them with the new preservation repository).

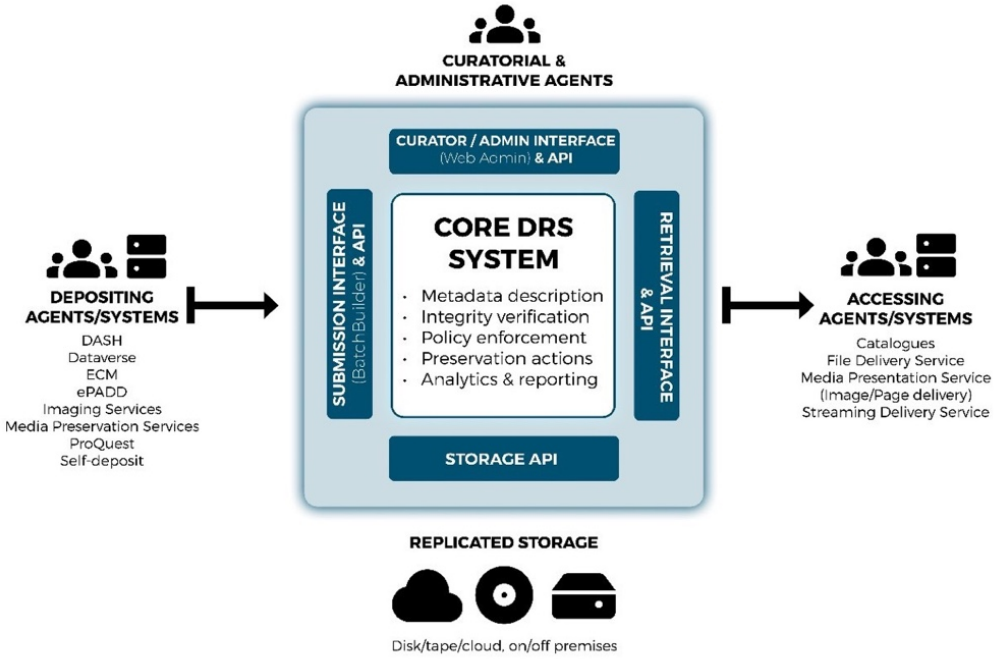

Figure 1 – The DRS within the wider stewardship ecosystem

As stated in the DRS *Policy Guide*, any digital content that supports the research, teaching, and learning mission of the University as well as its smooth administrative operation is eligible for stewardship in the DRS. Participants were invited to discuss all their anticipated content needs, regardless of whether those content types are currently supported in the DRS.

The DRS Futures team selected the focus group topics by prioritizing the issues that stakeholders discussed at the open meeting and office hours. The topics were designed to parallel the topics of the survey. The DRS Futures team offered Access, Content Management, Curators and Administrators, Deposit, and Reporting focus groups 2-3 times each with both morning and afternoon sessions to make the times as inclusive as possible. The Research Data focus group was added after we reviewed the participants and noticed that the research data community was underrepresented.

Focus Group Participation

There was strong interest in the focus groups, with 76 registrations and 56 participants. Approximately 41% of participants attended more than one focus group. Representatives from approximately 39 departments attended the six focus groups.

Access Focus Group Summary

Highly granular access to preserved digital content in the new repository will be valuable. While this is a problem that will largely be the purview of the delivery system, the repository will need to support metadata on the restrictions for access. Repository users would value a system that allows for both simple and complex access policies. A matrix of access possibilities in which curators could choose from a variety of access options would be an attractive possibility. The system should support embargoes and the automatic reassignment of content from restricted to free access categories at the end of the embargo period. Curators would also like the delivery system to support time-limited access to content users.

Content Management Focus Group Summary

The DRS was designed as a repository for final versions of digital content. Content curators find themselves with digital content that requires updates, changes, and amendments. While there is a tension between what file or object modifications are appropriate in a repository system and which should be made using external programs. Administrative views for content in the new repository would be highly valuable including: thumbnail views especially for visual content, a dashboard with statistics, The new repository should be interoperable with a variety of search, discovery, and access systems. Interoperability and the ability to sync with metadata from other systems is of high importance.

Curators and Administrators Focus Group Summary

Departments are currently adapting their workflows to the requirements of the DRS but would love a flexible system that would better connect with other databases and reduce duplication of work. Depositing is challenging and requires users to learn the system over and over again. The DRS is well-respected. Users would like access and restrictions to be more granular. Data continues to grow in both size and scale.

Deposit Focus Group Summary

Curators have little control over the materials they need to preserve and would value a system as flexible as possible in regard to file format, file structure, folder structure, and interoperability with other systems. High priority is placed on simple, intuitive deposit that allows departments to perform their own deposit, specify metadata fields, remember departmental preferences for fields, simplify reporting and allow bulk editing features such as modification of files, deletion of files, and association with specific metadata.

Reporting Focus Group Summary

The deposit load email is being used to provide a receipt of deposit, to identify when digital content is ready for subsequent tasks, to be scraped and loaded into spreadsheets for processing tracking, and other uses. Reporting that minimizes human data manipulation would be most welcome. Curators would like the new repository system to offer CSV or other outputs that can be ported or synced with HART or other data visualization programs. They do not see that as the job of the DRS but would like the DRS to support that work by offering interoperability with other systems or exports of data that can be easily ingested into other systems. Although usage data is largely the purview of the delivery system, curators are nevertheless interested in repository usage data – how often certain digital objects are used by curatorial staff. It is important to curatorial users that reporting data be current, however, there is not a common consensus on what that means.

Research Data Focus Group Summary

The relationship between the research data lifecycle and digital preservation was explored with special attention on issues of integrity and trust. The relationship between research data and software for use as well as the challenge of preserving user interactions with humanities research data were talked through. Integrity, transparency, and reusability were top priorities in this focus group.

Comparison of Stakeholder Input

It is helpful to consider the topics surfaced from the focus group discussions in connection to the topics that were identified from other stakeholder engagement opportunities.

Focus Groups	Survey	All Harvard Open Meeting	Executive Meeting
Self-service and ease of use	Easy deposit/uploading	Easy to use	Improved usability
Interoperability with other systems including metadata syncing and system prioritization	Interoperability with other systems	Seamlessly integrated with other Harvard systems	Easier System Management
Support all formats and support complex file/folder structures	Increasing the content supported by the repository	Able to support all digital formats	Extending the User Community
Bulk Editing and deleting, “shelf-prep” space for processing	Bulk/Batch features	Scalable and flexible	

Table 1. Comparison of Feedback

Stakeholders consistently communicated that their top priorities for the new repository are ease of use, improved usability, strong interoperability and integration with other systems, and support for bulk and batch processing features. Ease of use and improved usability is a constant message from all stakeholders. In addition to highlighting similar priorities for the future of the Harvard repository, the focus groups surfaced additional topics including a medium-term storage space for manipulation, processing, and editing of content before preservation.

Appendix I. DRS Futures Access Focus Group Report

Scope

The access focus group provided DRS (Digital Repository System) stakeholders with an opportunity to discuss their needs and concerns about curatorial access to data in the DRS. This focus group topic was intended to provide DRS stakeholders with an opportunity to discuss their needs and concerns around DRS access policies. Participants were invited to register and self-selected their preferred topics, dates, and times. Specific questions for the focus group were:

1. What kinds of flexible access policies are needed for batches, individual objects and files uploaded to the preservation repository?
2. Currently, three access levels are available for deliverable content: public, restricted, or no access. Which other more nuanced policies are required, possibly involving user group-level policies?
3. What kinds of time-based access policies are needed to facilitate content sharing and access?
4. What kinds of resources need access from specific environments or locations (such as reading rooms for example)?

The curators and administrators focus group was offered three times: on Monday, April 24, 2023, 10:00 AM to 11:00 AM, Wednesday, April 26, 2023, from 4:00 PM to 5:00 PM and on Tuesday, April 11, 2023, from 11:00 AM to 12:00 PM.

Summary of Discussion

Highly granular access to preserved digital content in the new repository will be valuable. While this is a problem that will largely be the purview of the delivery system, the repository will need to support metadata on the restrictions for access. Repository users would value a system that allows for both simple and complex access policies. A matrix of access possibilities in which curators could choose from a variety of access options would be an attractive possibility. The system should support embargoes and the automatic reassignment of content from restricted to free access categories at the end of the embargo period. Curators would also like the delivery system to support time-limited access to content users.

Highlights of the Conversation

What kinds of flexible access policies are needed for batches, individual objects and files uploaded to the preservation repository?

- We need both simple and complex access policies such as:
    - Simple: Open access out on the world wide web
    - Complex: Open for approved researchers only
        - Requires acknowledgement of copyright restrictions
        - Open only to users with a Harvard ID
        - Open for units or departments
        - It would be fine to require an initiation to request access to restricted materials
        - One person or one class access to one item
        - Access limited to a single computer, for example in a reading room
        - Vetting system to determine if the requesting party is considered responsible in terms of security and privacy.

        - Approved research for tribal content may require seasonal access, gender restrictions, or other non-typical solutions.
- Requests for access for physical access to items take between 25-28 hours. While it would be great to offer access to digital items sooner than that, there is a concern that immediate access to requested data might lead to constant requests. There are instances where a human needs to vet the digital content for use before providing access. That needs to be built into the process too.
- Acknowledgement of violent or disturbing visual content before viewing would be helpful. We don't want to be gatekeepers to history, but viewers should be notified before they download something violent, racist, or offensive. The National Archives in Australia has examples for how to do this. Although Harvard has a harmful language message now it doesn't really work for visual content.
- A system that communicates to users when they have the right to access scholarly materials in a friendly way. And that makes it clear that content exists even when access to it is restricted.
    - Example: This digital item exists, based on your login, you cannot access it as it is not publicly accessible. Right now, it is unclear if a digital item exists, needs to be digitized, etc. The user has to refer to an internal Harvard Library spreadsheet if it's not on SDS or is restricted.
- The current system for agreeing to terms of access is cumbersome. We field the request through email and then manually share the content with the user after they agree to the terms. A system in which there was a click-through statement where people agree to terms, and then they can have access to the item would be great. Some departments need a record of this agreement, but others do not.

Currently, three access levels are available for deliverable content: public, restricted, or no access. Which other more nuanced policies are required, possibly involving user group-level policies?

- It would be great to have a matrix of access possibilities in which curators could mix-and-match the following:
    - Challenge (challenge scenarios may require profiles)
    - Individual credentials group credentials
    - Individual computer credentials
    - Read-only (no copying, no downloading)
    - Duration restrictions (a definition of what limited time access)
    - Permission authentication/vetting to streamline granting limited access to content that must be restricted or suppressed.
- Donors often also have conditions of use, and the system might need to track the user acknowledgement of the condition of use form. It is unclear if an actual signature is required or simply an acknowledgement. It will depend on the specific collection. Commonly, signed terms of use documents go through a data use agreement process. Alma-D might be a good example of how to manage a variety of access restrictions. There could be instances where providing access to collection donors for help with descriptions or review before a soft launch would be really helpful.
- Each collection might have individual conditions as well as basics like copyright. It would be useful to have several common default access scenarios defined as well as the option to customize for specific restrictions.
- Consider the results from DLF Born-Digital Access Working Group "visioning remote access" sessions when thinking about the remote access piece.

- The workflow to get content from the DRS to systems for class use is onerous. The new repository should have an easier way to integrate with Canvas, RTL, and other systems that provide content to students.
- Geo-blocking would be helpful. Copyright restrictions are not geographically uniform – we need to restrict access to items in certain regions but not others.

What kinds of time-based access policies are needed to facilitate content sharing and access?

- Time-based access can apply to the duration of access that should be granted to various categories of researcher, but it can also mean restrictions to access based on embargoes or copyright restrictions for a certain time period.

Embargos:

- For currently restricted content that will become open after a period of time, it would be valuable to input that date when the content is being ingested so the system could automatically recategorize the content as accessible at the appropriate time.
  - Examples: University records 50-80 years
  - Donor restrictions
  - Research data commonly have 1–5-year restrictions
- For high-demand university records, like student records, a streamlined process for staff to allow access – approve access requests without having to download and email copies would be valuable

Limited Access:

- Aviary has some useful time-based features that might be worth considering. There are 24-hour content restrictions for some content, we can offer up to a week of access for remote researchers and can even create links with specific permissions.
- Temporary links could be person-specific or more broadly available but only for a limited time period.

What kinds of resources need access from specific environments or locations (such as reading rooms for example)?

- Virtual reading rooms would be useful. Sometimes there are limited situations that we need to respond to. For example, during the pandemic, we needed to provide limited access to copyrighted materials like commercial recordings.
- We also need to consider that some items in the licensed, controlled digital lending model (one copy, one user) might be in the repository. Perhaps the delivery system should be handling access for these items though. Automating this is important though, it is challenging to manage manually.

Participating Departments:

Loeb Music Library
Gutman Library
Woodberry Poetry Room
Academic programs
Discovery/Access
Museum of Comparative Zoology
Houghton Library
Harvard Graduate School of Design
Harvard Business School Baker Library)
Harvard University Archives
Harvard Kennedy School Library & Research Services
Schlesinger Library
Harvard University Herbaria
HUIT
Widener Library Judaica Division

Appendix II. DRS Futures Content Management Focus Group Report

Scope

The managing content focus group provided DRS stakeholders with an opportunity to discuss their needs and concerns around how to manage content in the DRS. Participants were invited to register and self-selected their preferred topics, dates, and times. Specific questions for the content management focus group were:

1. What are your requirements for altering content (updating metadata, adding files, deleting files, flipping images, etc.) that is already deposited in the repository?
2. What sort of administrative views of your content would you like to have?
3. What are your requirements for content retrieval at the aggregate, individual object, and individual file level? What about primary content versus accompanying metadata?
4. What enhancements are needed for more effective access policy and enforcement? What is the proper unit of granularity for designating access rules regarding content, users, and time?

The managing content focus group was offered three times: on Tuesday, March 21, 2023 from 4:00 PM to 5:00 PM, Wednesday, March 22, 2023 from 11:00 AM to 12:00 PM, and Tuesday, April 4, 2023 from 4:00pm to 5:00pm.

Summary of Discussion

The DRS was designed as a repository for final versions of digital content. Content curators find themselves with digital content that requires updates, changes, and amendments. While there is a tension between what file or object modifications are appropriate in a repository system and which should be made using external programs. Administrative views for content in the new repository would be highly valuable including: thumbnail views especially for visual content, a dashboard with statistics, The new repository should be interoperable with a variety of search, discovery, and access systems. Interoperability and the ability to sync with metadata from other systems is of high importance.

Highlights of the Conversation:

What are your requirements for altering content (updating metadata, adding files, deleting files, flipping images, etc.) that is already deposited in the repository?

- Rotating images – right now it involves downloading the image, deleting it from the repository, rotating it on the desktop, and re-uploading it back to the repository.
- Enhancing objects with supplementary files – for example adding OCR to large batches of page-turned objects.
- Updating of the digital object labels and the ability to connect to other metadata systems to sync updates (including ArchiveSpace, Alma, etc.). Currently all metadata changes must be updated by hand.
- Perhaps the old model of data in preservation repository being at rest is no longer completely true. Perhaps data is no longer fully at rest, even when in a repository. Updates are becoming more frequently required due to the increasing capability of tools. A system needs to allow for the evolution of content over time while maintaining the authenticity and integrity of the original content.
- Version history with a rich vocabulary of relationships with a strong midterm management layer – to keep content kind of safe, allow appropriate curatorial changes (add OCR, etc.), before the content is ready for a permanent repository system.

What sort of administrative views of your content would you like to have?

- Thumbnail views would help us sort visual files and do quicky quality checks.
- Downloading search results into other output forms (e.g. CSV).
- Fuzzy search functionality with dynamic ability to search for strings across multiple fields at the same time at different DRS hierarchical levels.
- Persistent result sets from which administrators can create derivative result sets.
- GUI dashboard with statistics on file formats, version, at-risk statistics, recently deposited info, etc.
- Different administrative view options – basic vs. advanced view with the ability to do command line query searches or offering an API search.
- Pre-deposit view would be useful but has to be done in such a way that there isn't a backlog and curators aren't duplicating work. This could help reduce the need for altering content.
- The ability to set up different views – such as a pre-deposit view, a data view without load reports, a search view for all objects or files for search, an export view so I can customize what data I need.
- The DRS1 interface was efficient for searching batches of things that were not deposited together in the same name range.
- The ability to share DRS data outside of the DRS system – such as using the DRS data in HART: Harvard Analytics and Reporting Tool. The ability to create reports with data from the DRS, Alma, HART, and other systems is key.
- Having a clear "system of record" when things are described in multiple systems is important. You can't manage your collection without getting your metadata in the system.

What are your requirements for content retrieval at the aggregate, individual object, and individual file level? What about primary content versus accompanying metadata?

- The ability to assign materials to a collection aggregation would be great so that we aren't using billing codes for that purpose or relying on batch, object, or file naming conventions. This would allow easy retrieval of all materials from a collection.
- The ability to search by collection title would also allow someone to create collections that run across projects/billing codes.
- Most of the things we deposit to the DRS are also going to other systems – it would be great to develop the structural metadata once and be able to upload it to other systems or to the DRS.
- The ability to download large quantities of files or objects – bulk downloading.
- Visual search results such as thumbnails would help identify the quality of images for reference files or publication queries.
- Deleting items from the search results without actually deleting the files from the DRS and leaving empty digital objects.
- Additional downloading options with finer granularity would be helpful – to have the flexibility to only request certain parts of an object for instance.
- Increasing the speed of downloads would be helpful.

What enhancements are needed for more effective access policy and enforcement? What is the proper unit of granularity for designating access rules regarding content, users, and time?

- A more sophisticated system of assigning permissions and access with more granular access options (time-bounded, site-specific, content-limited, only users registered for specific courses, limited permissions for student employees, etc.).

- The ability to schedule and keep track of release dates or access embargos – It would be great to have it all at the file level but automated with a reminder notice so that everything would just become available at the appropriate point without specific intervention.
- Targeted access with an option to evaluate for specific requests and specific users.
- Instead of offering security through obscurity where the link is public but not published, it would be better and more secure to offer targeted access.
- Increased granularity in terms of permissions and ownership -- sometimes things get misattributed to one department when they belong to another. It would be good to be able to look at or change ownership to specific data. It would be good to at least be able to see what others are depositing and owning in the repository.

Note

A concern with automating data loads from anywhere is that we might give the DRS the Alma ID but there might be better description in ArchiveSpace – the system of record from which the metadata should come isn't clear at this point. If we are updating things, system to system updates automatically. It will be imperative to make sure that there is a system of record from which to draw so that the most current and detailed metadata information is updated appropriately.

Systems integrating data with the DRS include:

Aeon
Archive-It
ArchivesSpace
ePADD+
Harvard Geospatial Library (HGL)
HOLLIS
JSTOR Forum
LibraryCloud
Museum DAM
TMS
Macaw

The ability to indicate a preferred system from which to draw metadata would be really nice.

Participating Departments:

Harvard University Archives
Gutman Library, Harvard Graduate School of Education
Botany Libraries
Fine Arts Library
Loeb Music Library
Gutman Library, Harvard Graduate School of Education
Harvard Business Publishing)
Ernst Mayr Library, Museum of Comparative Zoology
Fieldwork Archives at Dumbarton Oakes
Property Information Resource Center
Harvard Library Imaging Services
Recorded Sound & Media
Baker Library
Houghton Library
Harvard University Archives
Harvard Art Museums, Digital Resource Department

Appendix III. DRS Futures Curators and Administrators Focus Group Report

Scope

The curators and administrators focus group provided DRS stakeholders with an opportunity to discuss their needs and concerns regarding the DRS. This focus group topic was intended for those that rely on content getting into the DRS and for those who are interested in exploring the DRS but do not currently use the service. Participants were invited to register and self-selected their preferred topics, dates, and times. Specific questions for the focus group were:

1. How well does the DRS system support your departmental workflows? Are there needs that you have for your data that are aided or impeded by the current DRS? What would you envision to help solve these problems?
2. Where does digital preservation fit into your collection development strategy?
3. How does Harvard's digital preservation service help or hinder your fulfillment of donor agreements, grants, etc.?

The curators and administrators focus group was offered two times: on Thursday, March 30, 2023 from 3:00 PM to 4:00 PM and on Tuesday, April 11, 2023 from 11:00 AM to 12:00 PM.

Summary of Discussion

Departments are currently adapting their workflows to the requirements of the DRS but would love a flexible system that allowed them to better connect with other databases and reduce duplication of work. Depositing is challenging and requires users to learn the system over and over again. The DRS is well-respected but users would like access and restrictions to be more granular. Data continues to grown in size and scale.

Highlights of the Conversation

How well does the DRS system support your departmental workflows? Are there needs that you have for your data that are aided or impeded by the current DRS? What would you envision to help solve these problems?

- There is a general consensus that the DRS requires departments to conform to its requirements rather than the DRS flexibly adapting to the departmental workflow requirements.
- Departments feel there are too many constraints on: filetypes, how materials are packaged, and on user-friendliness.
- Departments need the ability to evaluate and assess in a self-service kind of way – for example how many files are in the DRS and how much storage they use as it impacts collection and budget strategies.
- The current DRS does well with previously described analog digitization with a defined content model – that's really fantastic. Stuff that hasn't been previously described or described in systems that are not HOLLIS it doesn't handle as well. Also born-digital stuff isn't handled as well. For example, image collections have had Exif data stripped out, impoverishing the collection. The new repository should capture the richest possible context possible, including production, collection, etc.

- There is a strong desire for systems to communicate well in order to sync metadata and other input information. There must be strong interoperability with other systems both within Harvard and outside of Harvard.
- Data keeps getting bigger and bigger – we need the new repository system to have a user-friendly workflow, but we also need access to storage at scale for both interim files and long-term preservation.
- The current DRS workflow for PDFs is challenging. Deposit in general is challenging if you don't do it on a regular basis. Downloading items and getting them out of the DRS is challenging.
- Restrictions to data aren't always communicated from the DRS to the discovery system which is challenging for staff. We find ourselves doing the exporting and sending as well as maintaining redundant storage on local drives.
- The new repository should be database agnostic – we might need to put an identifier and database name that can connect materials to multiple databases. When designing specs for how databases should communicate with the new repository system think they should be agnostic and open to a variety of databases – not too specific. We need to avoid putting orphans into the DRS, but we shouldn't be too prescriptive about the specific database. Some data may need to be private.
- Fine-grained permissions and restrictions would be helpful as well as an ability to communicate with the discovery system with viewer warnings for harmful content or violent depictions. Many items going into the DRS have 50–80-year restrictions to access. We do not have the capacity to go in and manually manage each of those files. We need the new repository system to make the switch automatically at the right time, from having access restrictions to being open for public access. Perhaps an automated notification would be useful.
- Currently it is a pain point to have to update items file by file. The ability to do updates at scale or volume, basically bulk updates would be highly useful.

Where does digital preservation fit into your collection development strategy?

- Digital Preservation is becoming our collection strategy – the digital objects themselves are becoming the collection not just a representation of the physical items.
- Our digital content must be connected to multiple systems – preservation but also DAMS, and various metadata databases, the DRS may not be the solution to all these needs, but it has to work with the solutions.
- We need secure storage for digital content before we are ready for preservation. The idea of shelf prep is missing right now, we need shelf prep for digital assets.
- Deposits to DRS can be tricky – especially for smaller batches where there have been changes to the spreadsheets. It feels like depositing is a new process every time and that we don't have routine workflows for it.

How does Harvard's digital preservation service help or hinder your fulfillment of donor agreements, grants, etc.?

- Donors to Harvard understand that the DRS is secure. We manage donor expectations by showing the delivery system functionalities. The demos are very powerful. Donors know that there is a team dedicated to the DRS and they find that reassuring. They like that DRS is central to Harvard and not local to the department. It can be a significant factor in collections acquisition.

- It isn't clear how data generated from grants is valued. Grants cover some storage costs for hard drives and the like but how do we determine what should go to DRS? And who pays for the storage costs after the grant is over?
- It can be challenging to know how to manage deposits that take place over time – for instance if the create updates their records and sends a new version. It would be great for the repository system to support dynamic versioning over time.
- Objects don't really go into the DRS in their final form, they are dynamic, and the new repository system needs to accommodate that.
- It would be great to be able to have more granular access.

Participating Departments

Houghton Library
Harvard University Archives
Botany Libraries Senior Archivist
Recorded Sound & Media
Baker Library Harvard Business School
Fine Arts Library
Harvard Divinity School Library
Museum of Comparative Zoology
Dumbarton Oakes
Harvard Museum of the Ancient Near East
Harvard Graduate School of Design Frances Loeb Library
Gutman Library
Widener Library
Harvard University Graduate School of Design

Appendix IV. DRS Futures Deposit Focus Group Report

Scope

The deposit focus group provided Digital Repository System (DRS) stakeholders with an opportunity to discuss their needs and concerns about curatorial access to data in the DRS. This focus group topic was intended to provide DRS stakeholders with an opportunity to discuss the needs and desires they have around the content deposit process in the new repository system. Participants were invited to register and self-selected their preferred topics, dates, and times. Specific questions for the focus group were:

1. What is the range of materials that you would like to submit to the DRS? How much discretion do you have regarding the specific format and structure of these materials?
2. What interface enhancements would help streamline support for single and bulk deposits?
3. How important are automated deposit workflows from external feeder systems (e.g., Dataverse, DASH, ECM, etc.) and the DRS?
4. What are the specific, repetitive tasks that you find yourself doing as part of your deposit workflow that you would like to automate?

The curators and administrators focus group was offered two times on Monday, March 20, 2023, 11:00 AM to 12:00 PM, and Thursday, March 23, 2023, from 4:00 PM to 5:00 PM.

Summary of Discussion

Curators have little control over the materials they need to preserve and would value a system as flexible as possible in regard to file format, file structure, folder structure, and interoperability with other systems. High priority is placed on simple, intuitive deposit that allows departments to perform their own deposit, specify metadata fields, remember departmental preferences for fields, simplify reporting and allow bulk editing features such as modification of files, deletion of files, and association with specific metadata.

Highlights of the Conversation

What is the range of materials that you would like to submit to the DRS? How much discretion do you have regarding the specific format and structure of these materials?

- HUA and HBS have little discretion – administrative records from all offices are accepted. The repository needs to support the deposit of complex structures with unpredictable file types. These are currently largely document-like files but there is interest in preserving database files in the future.
- There are specific content types that are not currently supported by the DRS:
  - Virtual reality files
  - Embedded metadata in audio and video files that need to be preserved
  - AutoCAD, Revit, Navisworks file types
  - RAW files
  - 3-D models via photogrammetry
- There are deposit problems that a new repository system could help address:
  - Intuitive, self-service deposit environment.
  - Shelf-prep space for processing digital content before adding it to the repository
  - Support file path structures that provide contextual information
- There are deposit problems that need to be sorted that are departmental rather than about the preservation system. Examples:
  - Limited staffing
  - Lack of clarity around changing collection goals

- Institutional and department collection policy guidance
- Ability to evaluate the quality of digital content before deciding on preservation

What interface enhancements would help streamline support for single and bulk deposits?

- Metadata syncing – for example, you could search HOLLIS, Alma, or ArchivesSpace and drag a record into a box, you search our computer and drag a file into a different box and say "go." Not a match on a file idea – just a window that allows you to link the metadata from a Hollis search to add to the deposit record. The user searches and confirms the match even if they aren't a full match in the two different systems. In other words, the repository system could scrape the metadata results from different applications and then auto populate.
  - Benefits to this include
    - Cut down on transcription errors
    - Reduce duplication of work
    - Eliminate copy-paste workflow that slows down deposit
- Ability to assign metadata fields from file names and file structures would be valuable
- Bulk deletes, bulk file name change, bulk object name change would all support file reconciliation with older work. For example, if a large number of files need to be replaced because they have been rescanned or miscataloged. Frequently the file itself is being replaced with a higher quality scan.

How important are automated deposit workflows from external feeder systems (e.g., Dataverse, DASH, ECM, etc.) and the DRS?

- Deposit emails are a challenge in many workflows. Departments have to save the email, copy the text, and insert it into spreadsheets to track.
- Departments expressed a wish to select content from Internet Archive, ECM, Directory (IAMPROD), and others to be automatically transferred to DRS, however, it is unclear if this is necessary data to be included in the Harvard repository or if it is data that the departments need easy access to in a Harvard delivery system.
- Transfer files in current file/folder structure rather than in a content model
- GUI deposit that allows editing before preservation – move pages, associate files with other files, review metadata, etc.
- Customizable dashboard that remembers the fields you want to show vs. hide
- Self-deposit AV files
- Download individual files without having to download a zip directory of other files I don't want.
- Download files with file owner supplied name instead of the original file name
- **V**iew the data of filename, deposit ID, Object name, etc., all in one place into something that can be downloaded. Wants a list of the files in the order in which they appear in a PDS object.

What are the specific, repetitive tasks that you find yourself doing as part of your deposit workflow that you would like to automate?

- Deposit reports should be updated to a spreadsheet rather than come in email. We need to scrape data from the emails and manually update spreadsheets to track project deposits. An application programming interface (API) might be one way to offer this information in a more user-friendly way.
- Re-evaluating the content of the emails is worth considering too – they contain information that isn't part of the departmental post-deposit workflow and at the same time there is information that the department has created as part of the deposit process that isn't conveyed in the email message and the curators must look up in the DRS.

- Customized reporting would be highly desirable. Or sticky searches that remember the fields we searched on in the past would help a lot of current users. A customizable dashboard would be useful.
- Perhaps there could be user data fields selected during deposit so when we send new deposits with our user identifier the fields are the ones we have pre-selected.
- As content managers, we usually check the number of objects that have been deposited, sometimes there are missing items. That is frequently identified when we are linking the data to search and access systems. Sometimes there is a data and metadata mismatch – maybe 1% of the content seems to have that problem and ending human matching might resolve that issue.
- Some audio content is reviewed after deposit, but it is a time-consuming process.
- There is now direct interoperability between DRS and Dataverse and almost interoperability between DRS and DASH. Subsequent changes to content in Dataverse are propagated to the DRS as the records do change over time.
- Archival processing takes time, so it is frequent for work on records to take place on a preservation copy. If there were shelf-prep space in which we could safely store files until they were processed, that could help us avoid changes to the preservation copy.
- In some cases, deposits don't happen because we don't yet understand the structure of how we want to deposit the files. In those instances, shelf-prep storage space would also be a solution. Curators are open to either manipulating the files in the shelf-prep storage space or to having the shelf-prep storage be more static and downloading a local copy to make changes prior to depositing in the repository. In either case, it would be important for curators not to be working on the original files until they are confident that the edits should be permanent.

Participating Departments:

Harvard University Archives
Harvard Library Imaging Services
Botany Libraries
Property Information Resource Center
Harvard Graduate School of Education
Harvard Business School Baker Library
Houghton Library
Archives at Dumbarton Oaks
Harvard University Archives
Americas, Europe and Oceania Division
Houghton Library
Ernst Mayr Library, Museum of Comparative Zoology
Buddhist Digital Resource Center
Harvard Faculty of Arts and Sciences Fine Arts Library

Appendix V. DRS Futures Reporting Focus Group Report

Scope

The reporting focus group provided DRS stakeholders with an opportunity to discuss their needs and concerns regarding reporting in the DRS. This group is intended to focus on the information and reporting needs of repository users. Participants were invited to register and self-selected their preferred topics, dates, and times. Specific questions for the focus group were:

1. Do you generate repetitive reports? Are there scenarios in which automated reports would be valuable?
2. When compiling report data, how current do you need the data to be for the reports?
3. Preservation reporting: What information do you require in relation to preservation actions?
4. Content reporting: What information do you require in relation to content deposited into the DRS?
5. Usage reporting -- How can the DRS support the usage reporting you need to have generated by other access or data systems?

The curators and administrators focus group was offered two times on Wednesday, April 5, 2023, from 10:00 AM to 11:00 AM and Friday, April 14, 2023, from 2:00 PM to 3:00 PM.

Summary of Discussion

The deposit load email is being used to provide a receipt of deposit, to identify when digital content is ready for subsequent tasks, to be scraped and loaded into spreadsheets for processing tracking, and other uses. Reporting that minimizes human data manipulation would be most welcome. Curators would like the new repository system to offer CSV or other outputs that can be ported or synced with HART or other data visualization programs. They do not see that as the job of the DRS but would like the DRS to support that work by offering interoperability with other systems or exports of data that can be easily ingested into other systems. Although usage data is largely the purview of the delivery system, curators are nevertheless interested in repository usage data – how often certain digital objects are used by curatorial staff. It is important to curatorial users that reporting data be current, however, there is not a common consensus on what that means.

Highlights of the Conversation

Do you generate repetitive reports? Are there scenarios in which automated reports would be valuable?

- Simple weekly reports of what was deposited to the DRS which are aggregated manually via SQL database parsing
- Monthly content preservation reports
- While not part of the DRS, usage reporting continues to be highly valued
- The ability to upload a template of the desired fields in a specified order for a report would be great.
- Some reports have millions of rows that must come down in chunks or come down asynchronously. All information is still required even if it comes in phases.
- Ad hoc reporting is frequently used. Pivot tables and how to build them to answer different questions for different audiences would be helpful.
- Running reports on which collections are missing metadata (including links, finding aids, and other issues) would be helpful in identifying what content needs additional curatorial work.
- Running reports on file types to narrow down the results and be more directed in what we are searching for would be great.

- Reports on restricted digital content – what is restricted and in what categories of restriction.
- Various file types and the item counts for each.
- Clear documentation about the repository metadata available and regularly scheduled offerings of training would be a big help.

When compiling report data, how current do you need the data to be for the reports?

- Reporting has been used as a means to get information about recent activity (deposit failure, status of bulk updates, etc.). If there were other, reliable means for getting this data in a 24-48 hour timeframe, that would be fine.
- Answers varied from as current as possible, weekly, and monthly.

Preservation reporting: What information do you require in relation to preservation actions?

- Prompt notifications about specific preservation actions aren't necessary, however, users would like to have a clearly accessible information on which preservation actions (example: fixity failure detection, remediation, obsolete file formats, etc.) have been taken. Even if curators don't have an action in making the preservation decisions, reports on specific preservation actions would be helpful in communicating with stakeholders and donors.
- How roles assigned to the digital objects inform which preservation actions are taken would be good to know. Storage fees for content have implications for managing the content so it would be useful to know how content gets moved and changes get propagated through the system.

Content reporting: What information do you require in relation to content deposited into the DRS?

- A clear discussion of the DRS load reports and deposit emails needs to be held before final changes are made. Several departments have created scripts to scrape the load reports and will need to find a way to get the data they need in the future without that particular process. It would be great to offer the data that departments need in a different format so that scripts weren't needed.
- Batch load reports are helpful – we've caught discrepancies by comparing the batch load reports to what's in the system.
- The load reports often also are used by departments as a prompt to begin the next steps of post-deposit linking, communication with others about what's been deposited, and what's available to be cataloged. If there were an audit trail for digital objects, maybe we wouldn't need the emails to find out when digital content was deposited.
- Replacing the load report would be fine but we need:
    - Load report as trigger/notification for departments (especially those who use depositing agents) to know that they are free to move on to the next step of the process
    - Receipt/audit trail
    - Move metadata back and forth between systems
- Syncing of the cataloging systems and the DRS – deposit content and have it automatically linked to catalog records. We used to be able to populate the MODS with just the HOLLIS ID. But we still need to go into all the records and put in the net holdings – there's this partial integration. If metadata could be automated again, that would eliminate the need for scripts and spreadsheets. Manually linking creates backlogs.

Usage reporting-- How can the DRS support the usage reporting you need to have generated by other access or data systems?

- Usage reporting should include:
    - Users by role (undergrad, grad students, users with/without Harvard Key, ideally by geographic location (inside/outside Harvard or the US). This information helps in targeted funding requests.
    - Monitor usage restrictions – these dovetail with access controls
    - Baseline information – more sophisticated than just clicks or citations.
    - Downloads from the DRS would also be useful for usage reporting as staff download digital content for our users and we need to track that information.

Additional Discussion

Insight into how things have been viewed/accessed/downloaded would be helpful – even if it isn't for end-user access but for curatorial access.

Generally, outputs of csv files or something else that can be ported to a data visualization system would be more useful than pretty graphic outputs. Trends and preferences change, the DRS data should support the visualization work in other systems, not prioritize visualization itself. Connecting DRS to HART would be one way to address this.

Some common, canned reports would be useful to have on offer. Even more helpful would be sticky report columns – or report templates that are saved to user profiles.

Key word search data would be helpful.

A contact person for questions – everyone gets trained at the beginning of a new system. But things shift and priorities change. We need to have someone to ask when we have questions.

Participating Departments:

Harvard Library Imaging Services
Baker Library Special Collections and Archives
Gutman Library, Harvard Graduate School of Education
Dumbarton Oaks Image Collection & Fieldwork Archives
Schlesinger Library
Houghton Library
Buddhist Digital Resource Center
Harvard Library
Harvard University Archives

Appendix VI. DRS Futures Research Data Focus Group Report

Scope

The research data focus group provided Harvard Library Digital Repository Service (DRS) stakeholders with an opportunity to discuss their needs and concerns regarding research data in the DRS. Research data was not the main focus of the DRS in its first iteration. As a new repository system is developed, the DRS Futures team is interested in thoughtfully and intentionally supporting research data needs.
This group is intended to articulate the research data community needs for short-term and long-term data management and how they can/cannot align with preservation solutions on the information and reporting needs of repository users. Another goal of the group was to identify specific funding agencies and specific data management requirements that need to be met by a preservation system. Participants were invited to register.

Framing Questions:

1. Research data sets are often living documents – how do they relate to long-term preservation? When they are highly changeable and dependent files that aren't necessarily "finished."
2. Would outreach to the preservation community change the relationship between the research community and the preservation repository? How could preservation services support research data in addition to the data management needs?

Research Data Focus Group Discussion Questions:

1. What are the main paths for contributing data?
2. Are the anticipated contributed datasets in a final form or are they "living" objects that will be continually enhanced over their lifetime?
3. What levels of support are needed for peer review and publication processes?
4. Is repository certification (Data Seal of Approval (DSA), DIN 31644, ISO 16363 (Trusted Digital Repository) important?
5. What are your data access needs, once the dataset is in the preservation repository?
6. What are your reporting needs? For material arising from funded research, what information is needed to provide back to funders?

The research data focus group was offered on Tuesday, May 2, 2023 from 11:00 AM to 12:00 PM.

Summary of Discussion

The relationship between the research data lifecycle and digital preservation was explored with special attention on issues of integrity and trust. The relationship between research data and software for use as well as the challenge of preserving user interactions with humanities research data were talked through. Integrity, transparency, and reusablity were top priorities in this focus group.

Highlights of the Conversation

What are the main paths for contributing data?

- Many in the research data community don't have current paths for contributing data to the DRS. The conversation focused on the benefits for data preservation and the current paths for research data to be managed.
- The research data community is using Dataverse, but also some of the NIH identified repositories for data management and analysis. There is another group of data that is not slotted for specific data management funding that could benefit from preservation.

Are the anticipated contributed datasets in a final form or are they "living" objects that will be continually enhanced over their lifetime?

- Curated data goes into publications but there is also research data that is valuable but does not have any formality – there is uncertainty about what data should be preserved. Researchers frequently go back to older data, for example at Harvard Business School we have some researchers going back to the old [Hawthorn] research from the 20s/30s that was never published. It's hard to know early on in the research data lifecycle what should be preserved. For the DRS, the team supports the most expansive view of what's eligible for preservation, which is anything that supports the mission of the University or administrative operation.
- Digital preservation is a managed set of activities that ensure 4 archival qualities: integrity, authenticity, accessibility, and usability. We need to think of the dividing line in the research data lifecycle at which it is ready to be preserved and re-used in the future
- There is a question about whether data in the repository system needs to be from Harvard or not. The DRS currently supports any data that Harvard deems worthy and important to preserve for the long term, with the caveat that there are rights questions involved in preservation – we need to have the legal authority to make lots of copies and modify over time, re-represent in new format, need to be able to do that and also distribute it, as long as content comes with a plausible claim that we have rights to do this it may be eligible.
- Long-term storage is FOR reuse; therefore, research data doesn't need to be static in order to be included in the DRS.
- It would be helpful to consider the relationship between the DRS and computing environments. For example, a computational environment where people can bring in data for reuse, like S3 buckets. It would be great to just pull in read-only access to DRS so people can pull in the project, make a copy of it, do whatever work they need to do but they keep that level of access to this long-term data. Being able to work with arbitrary storage platforms on or off campus with the current DRS and something to consider for the future repository.
- Trust would be of value to the research computing community – highly trusted resources with guaranteed integrity of datasets of value.
- It would be great to have a graphic that illustrates to researchers the ecosystem and costs of preservation. In computer science, we study the memory hierarchy, i.e., data resides in registers, cache, RAM, disk, tape, with other new technologies forming new layers in between those basic types with new attractive properties based on the technology. Registers are fast and expensive, all the way to tape, which is slow and cheap. I think data in the research data management lifecycle can be viewed similarly in that data moves from local systems to shared systems to sharing repositories to preservation. Each stage has latency, cost, and properties that are attractive for that stage. I think it might be useful to create a graphic that shows the latencies/costs/properties for each stage of the research data lifecycle.

What levels of support are needed for peer review and publication processes?

No discussion.

Is repository certification (Data Seal of Approval (DSA), DIN 31644, ISO 16363 (Trusted Digital Repository) important?

- Trust is key. The actual certification or formal audit is not as important as adherence to those standards. We can trust if you have enough transparency and don't need a formal audit.

What are your data access needs, once the dataset is in the preservation repository?

- It is a challenge that research data is often tied to software – which complicates how the data can be used. An important question: is software part of long-term evidence, or does scientific evidence and data structures live separate from software? Guidance would be helpful. It is important that data isn't just dumped but it usable in the future.
- Humanities data is different than scientific data – it's important to also document the interactions with the data, how they presented the data the way that they did, they're not trying to create re-creatable datasets. Preserving these interactions is challenging and still needs to be solved.
- People also need guidance around how to evaluate which research data are worth preservation – it's an investment and it would be helpful to have some policies or guidelines about how to make those decisions.
- There are also additional needs around Indigenous data – we need to be careful of access restrictions there. Depending on the community and the data, there are access restrictions for seasons, gender, and other concerns.

What are your reporting needs? For material arising from funded research, what information is needed to provide back to funders?

- Internal funding has different requirements than external funding. For Harvard funded data, they want to know about fairness of the data, getting to reusability—they want to make sure data is made available in a useful fashion back to the disciplines.
- For all funders, internal Harvard and external funders, it is useful to be able to articulate the core characteristics of data reusability and access, particularly like the core seal of approval. It is a good way to articulate how the DRS conforms to desirable attributes that funders look for. It's helpful to have language for folks who want to make use of this pipeline.

Participating Departments:

Harvard Medical School
Office of the Provost
Harvard Business School
Harvard Graduate School of Design
Harvard Institute for Quantitative Social Science
Harvard Dataverse
Harvard Faculty of Arts and Sciences, Research Computing