Bridge2Hyku: IMLS Funded Project

Digital Collections Survey Report

By Bridge2Hyku Project Team

Todd Crocken, Santi Thompson, Anne Washington, Andrew Weidner, Annie Wu
University of Houston

Contributors

Kristi Palmer (Indiana University-Purdue University Indianapolis)
Dean Seeman (University of Victoria)
Elliot Williams (University of Miami)

07/05/2018

TABLE OF CONTENTS

Bridge2Hyku: IMLS Funded Project

Digital Collections Survey Report

- INTRODUCTION
- SURVEY RESULTS
  - Institutional Characteristics
  - Digital Collections
  - Repositories
  - Metadata
  - Workflow
  - Migration
- Survey Implications
  - Implications for Migration Tools
  - Implications for Migration Strategy
- CONCLUSION
- APPENDIX: Digital Collections Survey Questions
  - Institutional Characteristics
  - Digital Collections
  - Repositories
  - Metadata
  - Workflow
  - Migration

Note: this project was made possible in part by the Institute of Museum and Library Services National Forum Grant LG-70-17-0217-17. The views, findings, conclusions or recommendations expressed in this report do not necessarily represent those of the Institute of Museum and Library Services

INTRODUCTION

The University of Houston (UH) Libraries, in partnership and consultation with Indiana University at Bloomington (IUB), Indiana University-Purdue University Indianapolis (IUPUI), University of Victoria (UVic), University of Miami (UM) and primary community stakeholders including Stanford University, DuraSpace, and the Digital Public Library of America (DPLA) was awarded an IMLS National Leadership Grant (LG-70-17-0217-17) to support the creation of the Bridge2Hyku (B2H) Toolkit for sustainable data migration from CONTENTdm to Hyku. Hyku (formerly called Hydra-in-a-Box) is an open source digital repository developed by Stanford University, Duraspace and the Digital Public Library of America (DPLA). The B2H Toolkit aims to help institutions understand their digital library ecosystems and provide software and guidance for successful migration to Hyku. This two year grant project is divided into three phases. In phase one, the team will identify metadata and system requirements needed for crosswalking data from CONTENTdm to Hyku. Phase two will be dedicated to the B2H Toolkit development, documentation, and the creation of a B2H website. In phase three, the team will assess, improve and promote the developed toolkit.

To fulfill the task of the phase one migration needs assessment, the B2H Migration Strategists and Hyku Metadata Advisors created a survey to collect data from partner institutions related to digital collections including metadata schema, production workflows, stakeholder considerations, and work/file types. This report intends to convey the survey results and implications for digital collections migration planning and implementation. The B2H Partners completed this survey in February 2018.[1]

SURVEY RESULTS

This survey collects data on institutional characteristics, digital collections, repositories, metadata, workflow, and migration. Responses from the survey will be used to assess digital collections migration needs and considerations from B2H project partner institutions. The data collected will also be beneficial for other institutions in their digital collections migration.

Institutional Characteristics

[1] This survey and report stands on the shoulders of previous surveys of the digital library community. Hannah Frost, Gary Geisler, and Mark Matienzo (Stanford University) articulate what features were prioritized when building Hyku in the Hydra-in-a-Box Survey on Digital Repositories (June 2016, https://purl.stanford.edu/jk292fy8802). Erin Tripp's (DuraSpace) Reframing Open Source Repository Upgrades survey (October 2017, https://osf.io/kbhmx/) provides useful background information on how the community currently sees the migration process.

This part of survey aims to assess the overall institutional resources and staff/librarians devoted to digital collections management for our partner institutions. The number of FTE staff/librarians for our four partners range from 80 to 162. The number of FTE staff/librarians supporting digital collection management range from 5 to 10. The number of FTE IT staff/librarians devoted to digital collections management range from 1 to 4. The annual library budget for our four partner institutions range from $10 million to $27 million.

Out of an average of 124.75 FTE, 7.625 FTE support digital collections management, roughly 6%. When asked how many FTE of local IT staff time is devoted to digital collection management, the average drops to 2%. Multiple units are involved in digital collection management in all four partner institutions such as digital strategies, digitization services, metadata/cataloging services, web services, Special Collections, library systems, and digital scholarship.

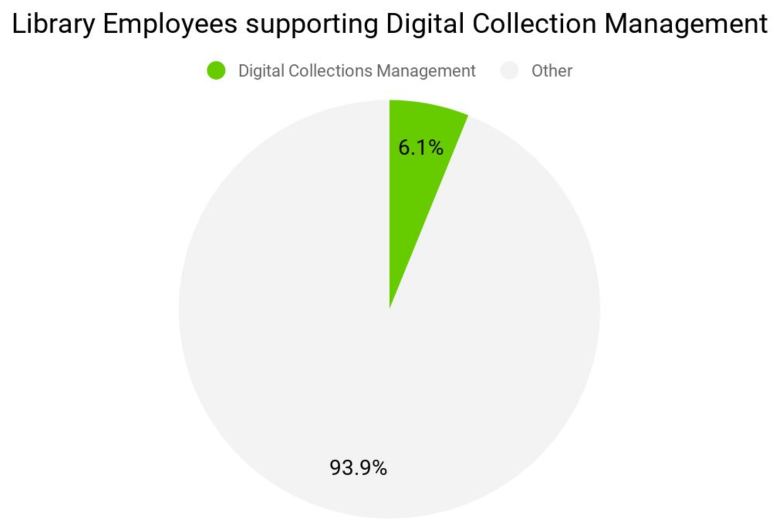

Figure 1: Percentage of employees supporting digital collection management

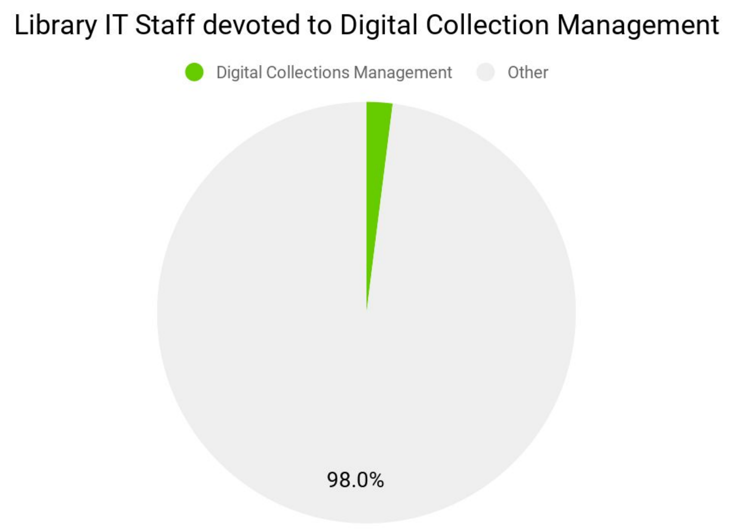

Figure 2: Percentage of Library IT staff devoted to Digital Collection Management

Digital Collections

The second section of the survey examines the size, type and format of each institution's digital collections. Table 1 illustrates the general digital collection characteristics reported by the B2H project partners. Figure 3 shows the work type and Figure 4 shows the file types reported by the B2H partners, with images (jpeg, tif, pdf) and audio/video (mp4, wav, mp3) as the primary digital object types.

Table 1: General digital collection characteristics

	# of Collections	# of Objects	Size in Terabytes
PI 1	135	379,279	3
PI 2	79	14,600	9
PI 3	61	46,149	1.2
PI 4	116	575,824	2.2
Average	97.75	253,963	3.85

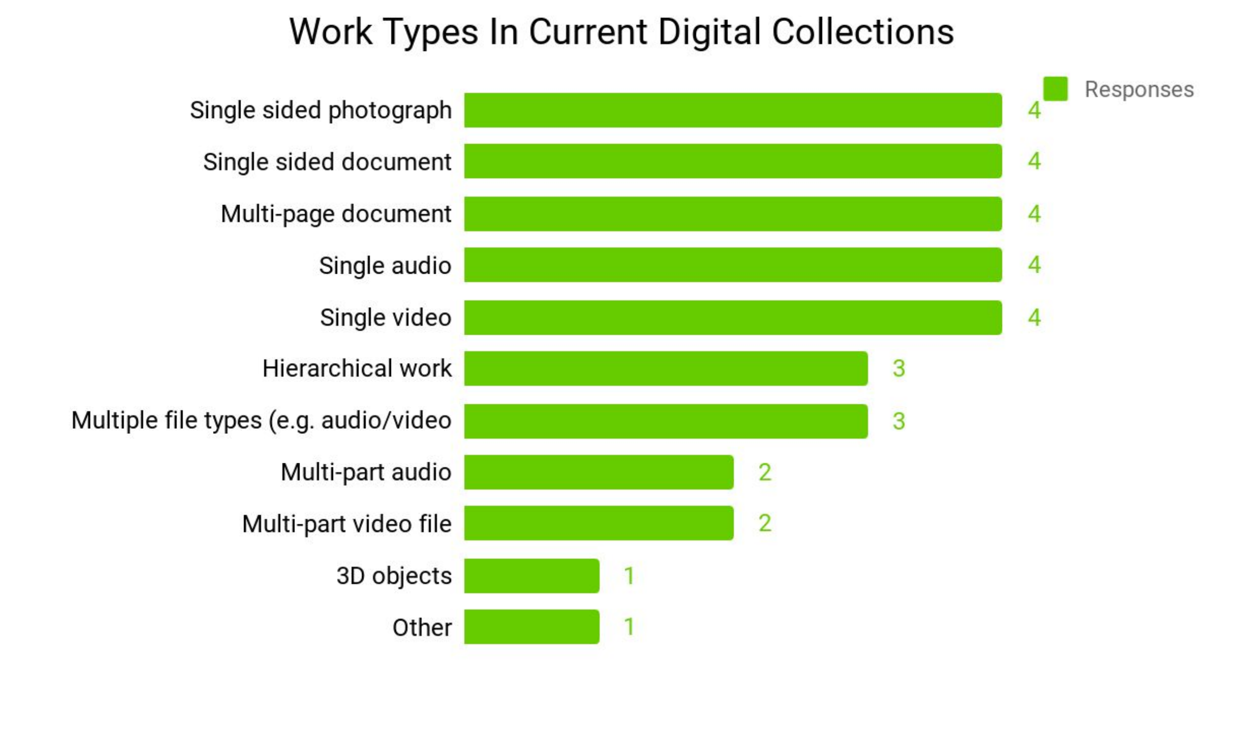

Figure 3: Work types present in digital collections

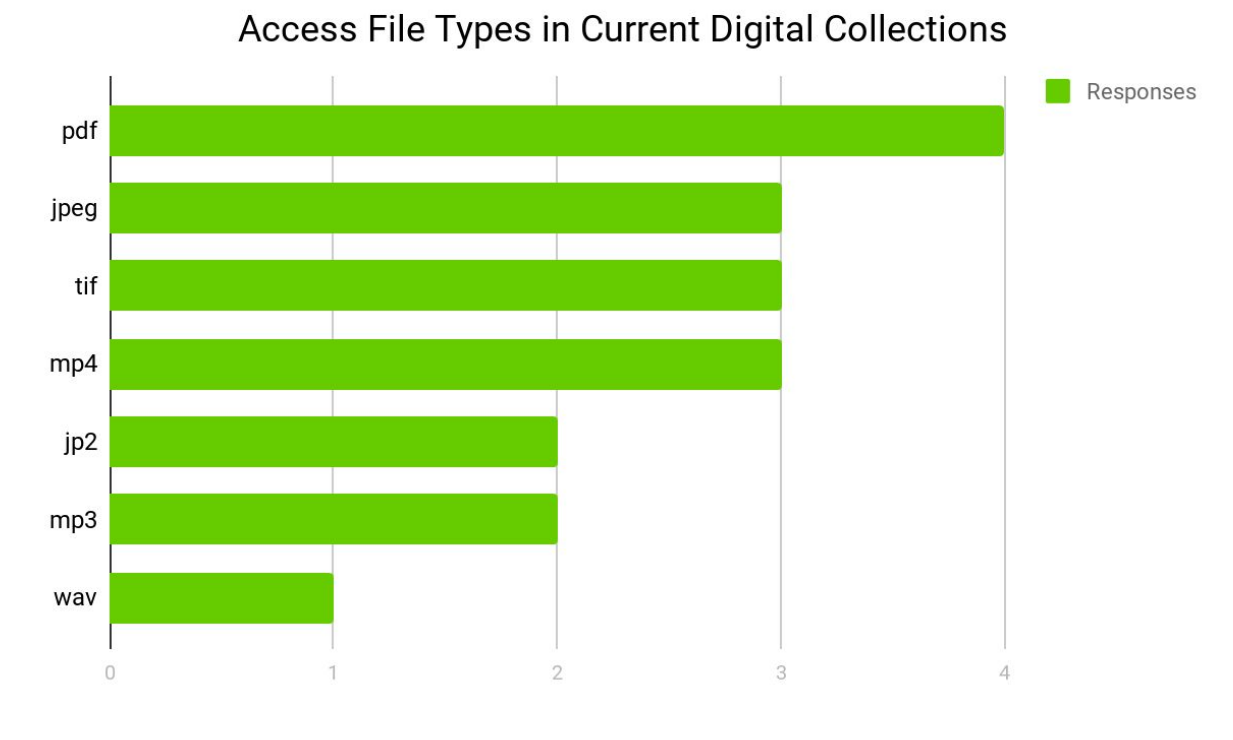

Figure 4: Access file types present in digital collections

This portion of the survey also included questions regarding digital collections stakeholders and audience. Special Collections, branch libraries, digitization units, metadata units, university archives, digital scholarship units, liaison librarians, and community partners are listed as primary stakeholders. The intended audience for all B2H partners' digital collections is broad ranging from students to faculty/instructors, library and university staff, university alumni and general public.

The last question in the Digital Collections section asks about the current administration of digital collections. The answers indicated a collaboration of stakeholders from different units such as digital systems, metadata, digital collections committee, and library systems.

Repositories

This section of survey explores the digital asset management system(s), dependencies and data model(s) currently used by our partner institutions. All four institutions are currently using CONTENTdm system. One institution also uses Kultura, another uses Archon and ArchiveSpace for finding aids. As shown in Table 2, all four partners indicated that they have dependencies for their digital collection systems.

Table 2: System dependencies for digital collections

Category	**Software/Platforms**
CONTENTdm customizations	Kaltura Media Player
Library transactions	Aeon request management system
Library catalog discovery	Primo discovery layer Summon discovery layer Digital object hyperlinks in MARC records
Digital exhibits	Blacklight Spotlight exhibit software Omeka
Aggregation services	Digital Public Library of America (DPLA) State-level aggregators
Other	Finding Aids Social Media

	Webpage links

Responses to the data model(s) question indicate that objects in B2H partner institutions digital collections vary from single image resources to two levels of hierarchy and to complex hierarchy with metadata at different levels.

Metadata

In the Metadata and Migration sections of the survey, questions were asked on metadata schema(s), controlled vocabularies used in current and future digital systems, metadata guidelines/standards and application profiles, data entry, and copyright statement.

Most partner institutions indicated in the survey that they will be migrating from CONTENTdm to Samvera (Hyrax, Hyku) system. The following charts display comparison of metadata schema(s) and controlled vocabularies in the current system and the migration target system:

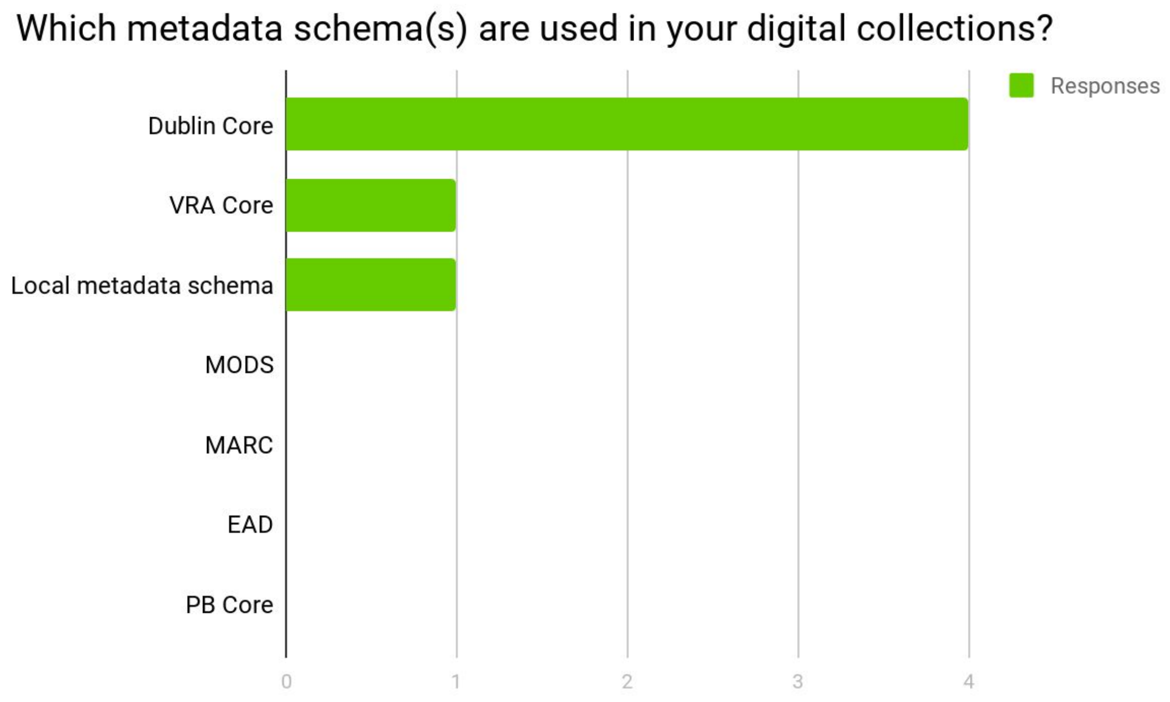

Figure 5: Metadata schema for existing digital collections

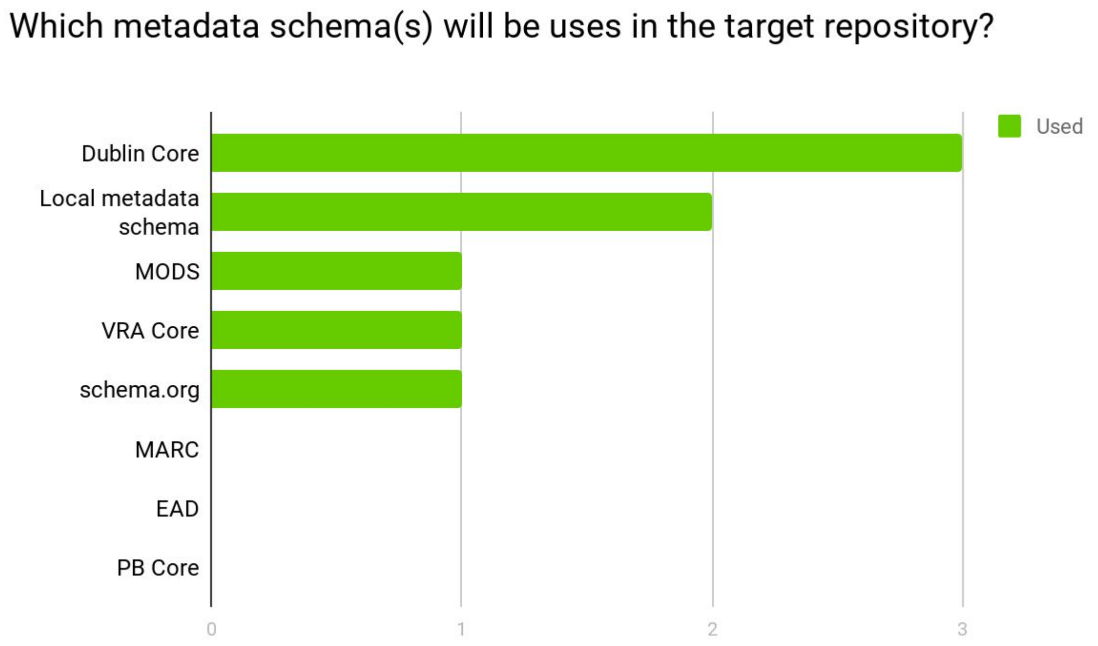

Figure 6: Metadata schema for future digital collections

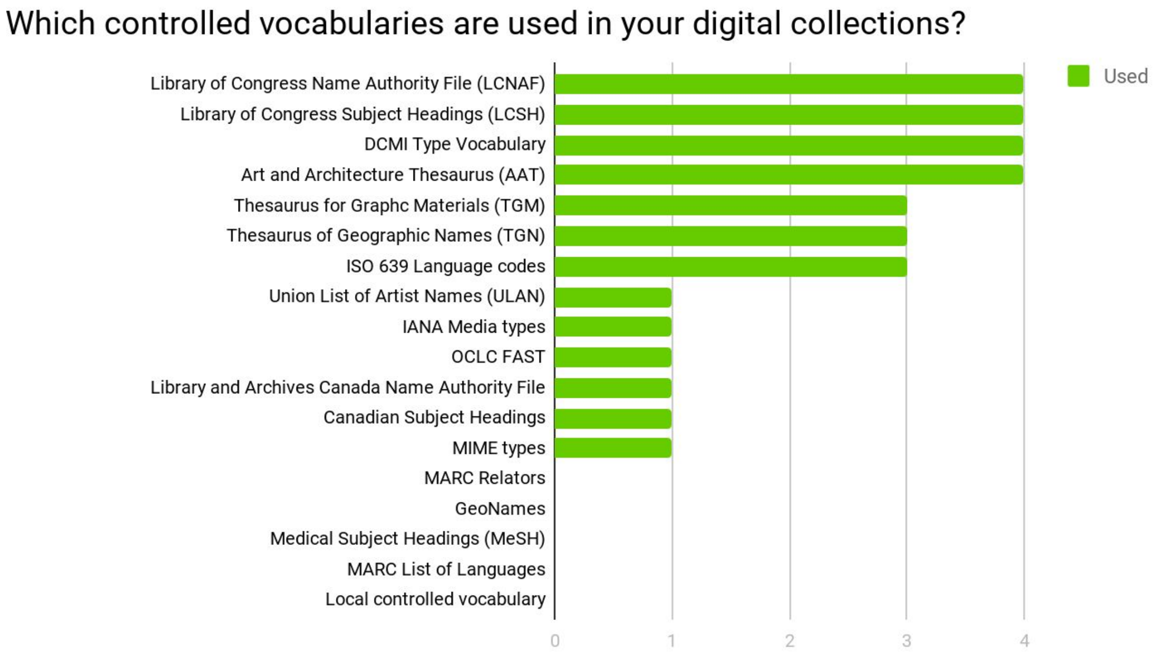

Figure 7: Controlled vocabularies for existing digital collections

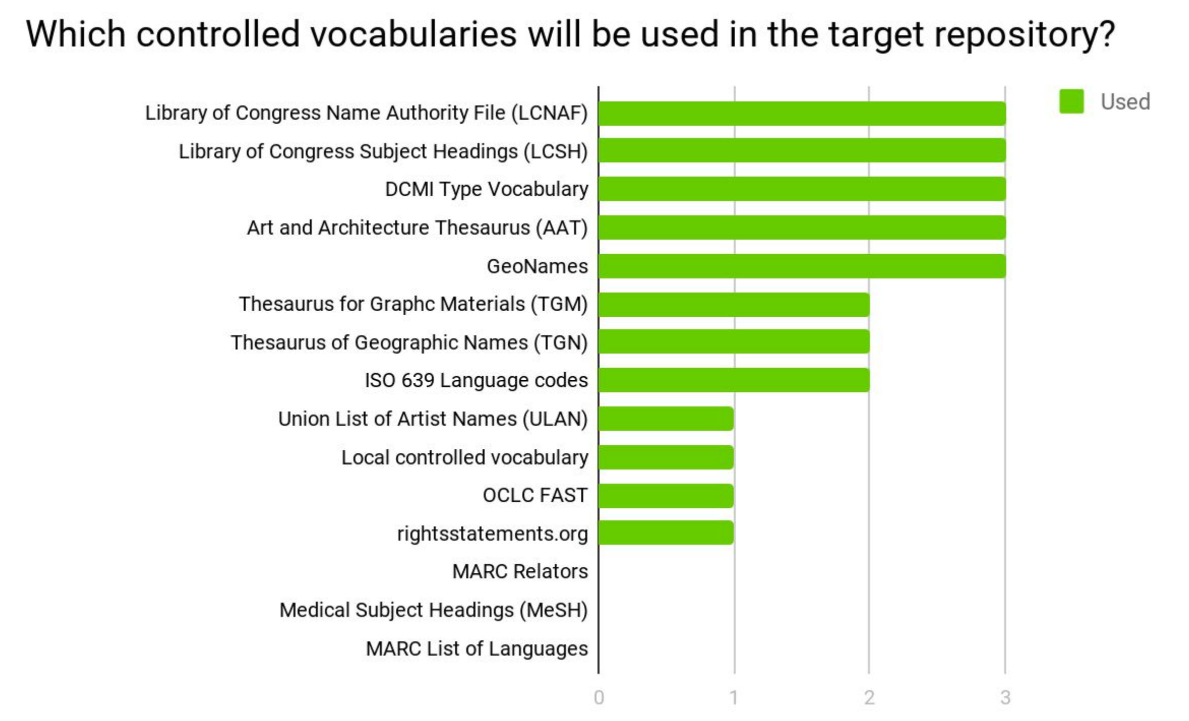

Figure 8: Controlled vocabularies for future digital collections

Survey respondents indicated that a core set of elements are used across all collections. However, the ability to create new labels and fields every individual collection in CONTENTdm has led to wide disparity in metadata fields and labels. For example, a repository may include different labels for the "Creator" field across collections such as "Author," "Scribe," and "Photographer."

Most partner institutions have created local metadata schemas, guidelines, and/or application profiles. They have also tried to enter metadata values consistently across their digital collections, however, there are still inconsistencies in areas such as date formatting, geographic names, title, and description.

Two institutions mentioned that their metadata guidelines/profiles aligns closely with DPLA metadata application profile. Others' are based on a combination of guidelines such as Resource Description and Access (RDA), Describing Archives: A Content Standard (DACS), Descriptive Cataloging of Rare Materials (DCRM), Rules for Archival Description (RAD), Dublin Core (DC) guidelines, Anglo-American Cataloging Rules, Second Edition (AACR2).

For rights metadata, two institutions have used statements from rightsstatements.org along with a locally developed rights statement. One has followed DPLA guidelines, which includes the use of rightsstatements.org. Another institution shared that rights metadata has been entered inconsistently using a variety of textual rights statements entered into different elements across collections: Rights, Use and Access Rights, Access and Use Rights, Use Rights, Access Rights.

Workflow

This section aimed to capture the process by which digital collections are added to an institution's repository. When asked to describe the process of creating a new digital collection, the responses were lengthy and in-depth, but taken in summary, included similar phases:

- Initiation: “request for digital collections”, “a call for digital projects”, “nominations and...suggestions for collections to be digitized”
- Approval: “evaluates, prioritizes, and selects which collections move forward”, “submissions that are approved are discussed… and put into a queue”, “reviewed… [and] operationalized”
- Digitization: “Digitization is done by staff and students in Digital Production department, who add files to storage server.”, “The materials are then given to the Digitization Unit where they are scanned. Files are renamed and put into their proper directory structure.” , “bring in… digitization staff”
- Metadata: “Metadata is then created by staff within Metadata & Discovery Services department.”, “the project goes to metadata where specialist create metadata and controlled vocabulary headings if need be”, “bring in… metadata staff and librarians”

The individuals and departments collaborating on these projects echo the findings in the “Digital Collections” section underscoring the finding that there are many stakeholders to consider when working with digital collections.

The other question in this section addressed digital preservation. Each institution includes digital preservation in their current workflow.

Migration

Largely, partner institutions plan to migrate from CONTENTdm to Samvera applications Hyku, Hyrax, and/or Avalon. One partner is still evaluating options including Islandoa, OCLC-hosted CONTENTdm, and TIND Digital Archive.

Responses in this section indicating metadata schemas and controlled vocabularies partners plan to use in their target repository can be found in the Metadata section above.

When asked to describe the digital collections rework planned before, during, or after the migration process, respondents shared the following activities:

- Metadata remediation
    - Standardizing metadata elements and values
    - Mapping values to new elements in the target repository
- Reprocessing materials
    - Rescanning materials
    - Rerunning OCR
- File management and renaming
- Digital preservation
- Other activities
    - Enhancing complementary library records such as a finding aids
    - Moving content out of library's digital collections to another management system

The final three questions of the survey were about the size and scope of each institution's migration efforts. B2H partners plan to have three to ten employees working on their migration. These employees would be from departments such as: Digital Strategies, Metadata and Discovery Services, Metadata and Digitization Services, Web and Application Development, Library Technology Services, Systems, and Special Collections. Respondents estimated their migration timelines to between periods as short as "at least 6 months" and as long as "2-3 years."

Survey Implications

The survey results provide the B2H team vital pieces of information for the development of use cases to drive migration tool evaluation as well as for enhancement of B2H tool features and functionality. The survey results also inform the components needed for successful migration strategy.

Implications for Migration Tools

1. The migration tool(s) need to accommodate images, text and audio/visual materials in numerous access file types.
2. The migration tool(s) should accommodate single, hierarchical, and multipart data model structures.
3. The migration tool(s) should be flexible enough to accommodate multiple metadata schema and standards.
4. The migration tool(s) should reconcile numerous controlled vocabularies including emerging vocabulary such as GeoNames.
5. The migration tool(s) should account for various system dependencies for digital collections

6. The migration tool(s) should allow for file renaming in transit.

Implications for Migration Strategy

1. To achieve efficiency in migration, the B2H Toolkit should include documentation on migration planning including content and metadata analysis, mapping of metadata elements and values, metadata standardization and digital preservation considerations.
2. The B2H Toolkit should include documentation on the migration workflow including metadata cleanup and remediation, materials reprocessing (rescanning and rerunning OCR), and step-by-step migration instructions.
3. It will be beneficial to include a bibliography of best practices for digital collections migration in the B2H Toolkit.

CONCLUSION

While the survey provides the team with data needed to evaluate, refine, and develop the migration tool and strategy, it also presents limitations that are important to note. This survey collected digital collections data from four grant partner institutions, which are medium size to large academic libraries. This data could be enriched by increasing participation from more diverse set of institutions, such as public libraries, smaller institutions, museums and other cultural heritage organizations. To address some of these limitations, the project team is exploring the possibility of working with a consortia to better understand the migration needs of different institutions in a hosted environment.

This survey reflects the first step towards building the migration toolkit. While the project team is working to solidify toolkit, we are anticipating that this toolkit can be expanded to address digital collections migration from other repositories beyond CONTENTdm to Hyku. The project team will also encourage active participation from the Samvera open source community to broaden the scope of this toolkit.

APPENDIX: Digital Collections Survey Questions

Institutional Characteristics

What best characterizes your institution’s type?

- Independent research library/archives
- Private college/university library
- Public college/university library
- Public library
- Government library/archives
- Regional consortium
- Historical society
- Museum
- Other

Number of staff/librarians (FTE)

Number of staff/librarians supporting digital collection management (FTE)

Departments of staff/librarians supporting digital collection management

Number of IT staff/librarians (FTE)

Approximately what number of FTE of local IT staff/librarians time is devoted to digital collection management per year?

Annual Library Budget

Digital Collections

Number of digital collections to be migrated

Number of digital objects to be migrated

Total size (TB) of digital objects to be migrated

Work Types

- Single sided photograph
- Single sided document
- Multi-page document
- Single audio

- Multi-part audio
- Single video
- Multi-part video file
- Hierarchical work
- Multiple file types (e.g. audio/video file with image or PDF)
- Other

Access file types

- jpeg
- tif
- mp4
- wav
- pdf
- Other

Who are the primary stakeholders for your digital collections?

this may include particualr departments in your organization, organization administration, community organizations, etc.

Which of these best describe the intended audience(s) for your digital collections?

- Undergraduate students
- Graudate students
- Faculty/Instructors
- Librarians and library staff
- University staff
- University alumni
- Non-university affliated researchers
- General public
- Other

Describe the administration surrounding your digital collections

What structures, committees, etc. are in place related to the administration of digital collections? This may include individuals or groups that create policy, have technical administrative control over the repository, etc.

Repositories

What system(s) are used to manage the digital collections to be migrated?

- CONTENTdm
- DSpace
- Islandora

- Sufia
- Avalon
- ILS
- Locally developed solution
- Other

Describe your digital collection system dependencies

Describe the systems involved and how they interact. For example, the ILS pulls digital collection data from an API, etc.

Describe the data model(s) supported by your your current repository

Consider describing the most complex object you have in your repository. How many levels of hierarchy does it have? Is metadata stored for each level?

Metadata

Which metadata schema(s) are used in your digital collections?

- Dublin Core
- MODS
- MARC
- EAD
- PB Core
- VRA Core
- Local metadata schema
- Other

Are the same metadata elements used across your digital collections?
If no, describe how metadata elements differe across collections.

Which controlled vocabularies are used in your digital collections?

- Library of Congress Name Authority File (LCNAF)
- Library of Congress Subject Headings (LCSH)
- Thesarus for Graphic Materials (TGM)
- MARC Relators
- DCMI Type Vocabulary
- Art and Architecture Thesaurus (AAT)
- Union List of Arist Names (ULAN)
- GeoNames
- Medical Subject Headings (MeSH)
- MARC List of Languages

- ISO 639 Language codes
- Local controlled vocabluary
- Other

Do you have local metadata input guidelines and/or a Metadata Application Profile?
If yes, please provide a link to your documentation if available.

Have metadata values been entered consistently across your digital collections?
If no, describe how values differ across collections.

Does your metadata, elements and/or values, align with any other standards or best practices? For example, does it align with the DPLA metadata application profile? DACS?

How do you indicate copyright in your digital collections?

Describe the metadata field(s) used, controlled vocabulary used (if applicable), and any other practices around rights metadata.

Workflow

Describe the process at your institution for creating a new digital collection
Please include general information about the process as well as the positions and departments that are involved. Examples of points to include are: the number and positions of those who add digital objects and metadata, whether there is centralized and/or formalized control over ingest and metadata production, etc.

Is digital preservation part of your digital collections workflow?

Migration

What system(s) will you be migrating your digital collections to?

Describe the data model(s) supported by the target repository
Include description or link to repository data model documentation. Consider describing the most complex object supported by the repository. How many levels of hierarchy does it have? Is metadata stored for each level?

Which metadata schema(s) will be uses in the target repository?

- Dublin Core
- MODS
- MARC
- EAD
- PB Core
- VRA Core

- Local metadata schema
- Other

Which controlled vocabularies will be used in the target repository?

- Library of Congress Name Authority File (LCNAF)
- Library of Congress Subject Headings (LCSH)
- Thesarus for Graphic Materials (TGM)
- MARC Relators
- DCMI Type Vocabulary
- Art and Architecture Thesaurus (AAT)
- Union List of Arist Names (ULAN)
- GeoNames
- Medical Subject Headings (MeSH)
- MARC List of Languages
- ISO 639 Language codes
- Local controlled vocabluary
- Other

Describe the digital collections rework you plan to undertake before, during, or after the migration process
For example, metadata rework, file management rework, etc.

Number of staff/librarians supporting migration (FTE)

Departments of staff/librarians supporting migration

Describe the timeline established for completing the migration