Strategies and Tools for Metadata Migration Analysis and Harmonization

Anne Washington
University of Houston, USA
awashington@uh.edu

Annie Wu
University of Houston, USA
awu@uh.edu

Santi Thompson
University of Houston, USA
sathompson3@uh.edu

Todd Crocken
University of Houston, USA
tcrocken@uh.edu

Leroy Vallejo
University of Houston, USA
llvallejo@uh.edu

Sean Watkins
University of Houston, USA
slwatkins@uh.edu

Andrew Weidner
University of Houston, USA
ajweidner@uh.edu

Abstract

The University of Houston (UH) Libraries, in partnership and consultation with numerous institutions, was awarded an Institute of Museum and Library Services (IMLS) National Leadership/Project Grant to support the creation of the Bridge2Hyku (B2H) Toolkit. Research shows that institutions are inclined to switch from proprietary digital systems to open source digital solutions. However, content migration from proprietary systems to open source repositories remains a barrier for many institutions because of a lack of tools, tutorials, and documentation. The B2H Toolkit includes general migration strategies and use cases as well as tools specifically designed for transitioning from CONTENTdm, a digital collections management software, to the Hyku digital repository. The toolkit acts as a comprehensive resource to guide migration practitioners in migration planning, metadata analysis and harmonization, and to facilitate the repository migration process. This paper focuses on how the toolkit's metadata guidelines and migration tools aid in migration planning, metadata analysis, metadata application profile development, metadata harmonization, and bulk ingest of digital objects into Hyku.

Keywords: repositories migration; metadata assessment; metadata harmonization; metadata mapping; metadata normalization; open source software

1. Introduction

Digital Asset Management Systems (DAMS) have evolved over time as the technologies that support them have been refined and user needs and expectations have shifted. Not surprisingly, libraries have come to reassess, select, and migrate to a new DAMS based on these changes in technology and user needs. To help institutions overcome barriers in digital repository migration, especially metadata assessment and harmonization, the University of Houston (UH) Libraries have collaborated with partner institutions on an Institute of Museum and Library Services (IMLS) grant project to build the Bridge2Hyku (B2H) Toolkit. Focusing on general information and guides for migration as well as on specific content for migrating to Hyku (formerly Hydra-in-a-Box), an open source digital repository from the Samvera Community, the toolkit will help institutions better understand their digital repository ecosystems and how they can prepare for migration. This paper describes the grant project background and outlines how the B2H Toolkit components,

including metadata strategies for migration and software that assist migration from CONTENTdm to Hyku, can assist practitioners with migration planning and metadata harmonization.

2. B2H Project Background

Many institutions are either investigating or in the process of migrating their DAMS. A research project by Stein and Thompson (2015) explores the motivations for migrating from one DAMS to another. Their findings indicate that the reassessment and migration process often result in institutions moving from proprietary systems, such as CONTENTdm, DigiTool, and Rosetta, toward open source solutions, including DSpace, Fedora, and Islandora. Furthering this trend, IMLS has invested funding in Stanford, the Digital Public Library of America (DPLA), and DuraSpace's collaborative effort to develop Hyku: a turnkey, open source digital repository. While open source repositories such as Hyku offer robust features such as flexibility, scalability and interoperability in making digital objects accessible over the web, institutions cannot take advantage of these benefits as they confront difficulties in legacy data migration, especially those institutions that may lack staff expertise and migration tools. Assuming this trend continues, the need for tools, tutorials, and documentation on how to migrate from a proprietary system to an open source system will grow over time. The B2H Toolkit helps fill the gap of lack of tools and guidance for institutions to migrate to open source repositories.

3. B2H Toolkit

The B2H Toolkit is an open source community resource that assists institutions in their migration from their current digital content delivery system to the open source Hyku platform. The toolkit includes: (1) a B2H GitHub Organization to host a website and software repositories, (2) content migration guidance and documentation, and (3) software to assist data migration practitioners with migration to Hyku.

The B2H website (http://bridge2hyku.github.io) compiles and outlines content migration guidance and documentation. While software developed for the toolkit is limited to CONTENTdm and Hyku, the website content is intentionally generalizable and applicable to all types of system migrations. It includes information about migration planning, strategies for metadata assessment and harmonization, modifying content, migrating content, and content verification. It also includes worksheets to help users plan their migration, blog posts highlighting approaches to different migration tasks, and migration stories from each partner institution offering advice and lessons learned.

The other major components are CDM Bridge and HyBridge, open-source software that help streamline processes for metadata assessment and content migration and provide required functionality for ingest into Hyku. CDM Bridge is an application designed to export metadata and files out of CONTENTdm into a standardized format for ingest into Hyku. CDM Bridge is easily downloadable and is compatible with both local and hosted instances of CONTENTdm. It connects to the CONTENTdm API to present a custom crosswalking interface for each collection in the repository (See Figure 1). Users map their CONTENTdm metadata fields to another metadata profile (default Hyku) and then export existing metadata and files out of CONTENTdm into an export package compatible with HyBridge. HyBridge is a Ruby gem installed in a Hyku instance that, when given a CDM Bridge export package, imports that package into Hyku and creates digital objects. CDM Bridge and HyBridge are available for download from the B2H GitHub (https://github.com/Bridge2Hyku).

The B2H Toolkit - both the website content and software - was developed collaboratively with grant partners as well as presented to the community where the team gathered feedback and feature requests. Workshops focused on the toolkit, with an emphasis on the general aspects of migration planning and strategy, received positive feedback. The team continues to welcome and incorporate feedback and contributions to the B2H Toolkit in response to users' needs.

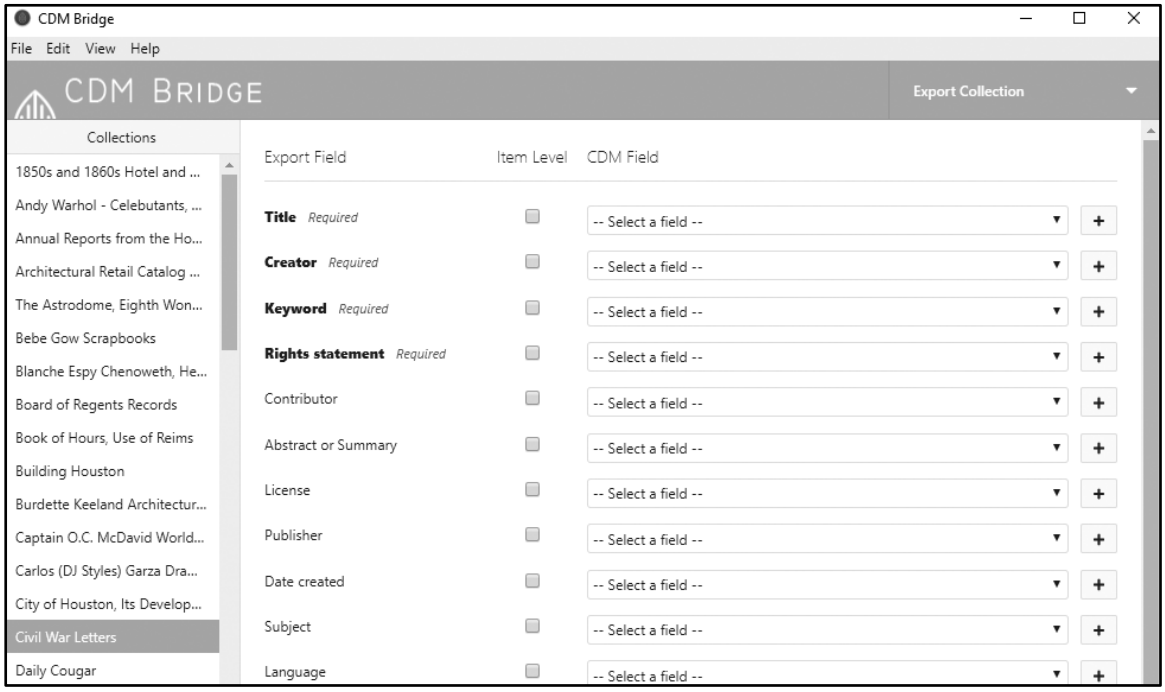

FIG. 1. CDM Bridge metadata mapping interface

4. B2H Toolkit in Practice

The B2H Toolkit aims to clarify and ease the complex process of digital systems migration. The following sections outline how the B2H guidelines and tools can help with migration, particularly how they aid in migration planning, metadata analysis, metadata application profile development, metadata harmonization, and bulk ingest of digital objects into Hyku.

4.1. Migration Planning

System migrations happen in a specific context that affects a migration project as a whole, including approaches to metadata analysis and harmonization. Analysis and planning at the outset will help to determine who should be involved, the extent of the work to be done, and limitations on products or processes. The B2H website contains detailed guidance on migration planning including a questionnaire to direct users through an information gathering exercise about their migration context. Briefly, some of the key considerations include library type, size, and budget, as well as the number and skillsets of staff committed to the migration. In addition to these factors, the type of repository—e.g. cultural heritage digital collections, institutional repository, data repository, etc.—and the number and size of the collections to be migrated will shape migration approaches. It is important to identify primary stakeholders and others that will be impacted by the work of migration. In terms of metadata, including these partners in decisions such as metadata mappings and standardization ensures that multiple points of view are taken into account and ultimately improves metadata and the repository as a whole (Washington & Weidner, 2017).

4.2. Metadata Analysis

A migration is a perfect opportunity to assess and improve metadata. Through an analysis of the current state of metadata in a repository, metadata issues can be identified and prioritized. Analysis may include activities such as reviewing metadata schemas and controlled vocabularies used in the repository, and metadata consistency across objects or collections. The B2H website contains suggested methods and considerations for metadata review. For example, inconsistencies and other

measures of metadata quality can be assessed by inspecting metadata reports, comparing metadata values between different items or collections, and comparing the existing metadata to more widely adopted standards and/or standards in place at the institution. Digital collections grow over time, and it is not unusual to have multiple input guidelines and multiple controlled vocabularies, among other changes in standards, used in the same digital repository.

In addition to its role as a main component in the B2H migration path, CDM Bridge is a standalone tool for extracting descriptive metadata from CONTENTdm for review and analysis. Configurable export fields allow users to generate full or custom metadata reports in CSV format for any collection at either the object level or the object and item level (for compound objects). CDM Bridge metadata reports overcome the challenges presented by built-in CONTENTdm metadata exports, which are inflexible and may be difficult to manipulate using spreadsheets, text editors, or other processing software. Reports from CDM Bridge are more configurable, and "Object Type" designations make it easy to filter by top level metadata for compound objects. The reports can be easily manipulated as a spreadsheet, in Open Refine, or by using scripts to identify irregularities and missing values. CDM Bridge can also facilitate collection-level analysis by indicating which records are missing values in a required metadata field.

In addition to the issues identified through the analysis of existing descriptive metadata, other issues will surface once that metadata is considered in relation to an existing or revised metadata application profile.

4.3. Metadata Application Profile Development

The development of a new or revised metadata application profile—the set of metadata elements used in the repository—often accompanies a migration. Institutional and user needs should primarily shape a repository metadata application profile, but the system to which the content will be migrated is often a driving factor. Reworking metadata standards may also include revising input guidelines as well as controlled vocabularies used in the repository. The B2H website includes more information about metadata application development as well as subsequent metadata mapping considerations and approaches. As noted earlier, this work is best completed collaboratively, including stakeholders that represent various users throughout the organization, so the resulting profile and standards take into account different organizational needs.

CDM Bridge can be a helpful tool in the process of assessing and revising a metadata application profile. One of the main functions of CDM Bridge is the ability to map metadata from an existing CONTENTdm repository to the Hyku metadata profile. However, users can configure a custom target set of metadata fields to suit their needs. Users can map their existing metadata to the new profile to determine whether the values fit the requirements of the fields. This mapping exercise will likely reveal additional metadata normalization issues to consider as a part of the migration.

4.4. Metadata Harmonization and Migration

With an understanding of the existing repository metadata, and with a (potentially revised) detailed metadata application profile, the work of metadata cleanup and harmonization can begin. Metadata strategists must determine which issues will be addressed, when they will be addressed and how. The B2H website characterizes metadata normalization as including both harmonization—or alignment with a metadata application profile and other standards—and enrichment, reconciling data values with new controlled vocabularies and/or adding URIs to metadata.

There are a variety of approaches to metadata rework. Tools such as OpenRefine help with metadata analysis, but can also be used to normalize field values within and across collections. OpenRefine can also be used to enrich metadata with URIs for linked data controlled vocabularies. Modifying titles and descriptions often requires manual analysis and rework. To standardize dates, a combination of manual and automated approaches may be used for analysis and validation. Fortunately, the field of metadata assessment and harmonization is active and provides practitioners

with information and tools for these processes. The B2H website directs users to these resources to aid in their metadata normalization efforts.

Metadata can be normalized at different stages of a migration: before content is moved out of the existing repository, in transit (once it has been moved out of the existing repository, but is not yet in the new repository), and after the content has been migrated to the new repository. There are costs and benefits to each of these approaches, with time and system limitations often being deciding factors. For example, metadata rework done before a migration improves the quality of the data going into the new repository, but could extend the migration timeline. In some cases, it may not be possible to edit the metadata in the current system because of functional limitations.

Unfortunately, bulk metadata editing in CONTENTdm is not always possible; the current version allows for some collection by collection bulk edits, but does not provide an accessible solution for flexible metadata bulk editing and reimport. To address these limitations, the B2H software was designed to make it possible to edit metadata in transit. After a collection is exported from CDM Bridge, metadata in the resulting CSV can be modified, enriched, and then ingested along with the digital object files into Hyku using HyBridge.

5. Conclusion

Using the B2H Toolkit, practitioners can overcome their migration barriers while, at the same time, advance open and sustainable digital collections principles. In addition to serving as a model for a collaborative, open source solution that makes migration a reality for institutions with limited resources, the B2H Toolkit functions as a key component of an efficient and effective metadata assessment and harmonization process. These actions have direct benefits for repository users and metadata practitioners. Assessing and harmonizing metadata improves its consistency, accuracy, and relevancy, which can increase the rate of discoverability for users searching for digital objects. Additionally, these benefits can help repository managers and metadata librarians create standardized metadata with consistently applied controlled vocabularies and metadata fields, making long-term maintenance more manageable and the prospect of future migrations more efficient. By taking advantage of the B2H Toolkit functionality, including its migration documentation and software tools, metadata practitioners contribute to an open, community-based solution for addressing the emerging needs of metadata migration planning, assessment, and harmonization.

Acknowledgements

The authors would like to acknowledge that this project was made possible in part by the Institute of Museum and Library Services National Leadership Grant LG-70-17-0217-17. The views, findings, conclusions or recommendations expressed in this paper do not necessarily represent those of the Institute of Museum and Library Services. The authors would like to thank the project partners and community stakeholders, fully listed on https://bridge2hyku.github.io/partners, for their work on this project.

References

Bridge2Hyku Project Team. (2018). Bridge2Hyku website. Retrieved, April 24, 2019, from http://bridge2hyku.github.io.

Bridge2Hyku Project Team. (2019). Bridge2Hyku Github repository. Retrieved, April 24, 2019, from https://github.com/Bridge2Hyku.

Stein, Ayla & Santi Thompson. (2015, September/October). Taking control: identifying motivations for migrating library digital asset management systems. D-Lib Magazine, 21(9/10). Retrieved from http://www.dlib.org/dlib/september15/stein/09stein.html.

Washington, Anne & Andrew Weidner. (2017). Collaborative Metadata Application Profile Development for DAMS Migration. Proceedings of the International Conference on Dublin Core and Metadata Applications, 2017, 117-119. Retrieved, April 24, 2019, from http://dcpapers.dublincore.org/pubs/article/view/3861.