IJDC | *Conference Paper*

# From Siloed to Reusable: The Opening of Digital Collections at Johns Hopkins University

Kathryn Gucer
Michelle Janowiecki

Johns Hopkins University

**Abstract**

In the past twenty-five years, cross-institutional communities have come together in the creation and use of open source software and open data standards to build digital collections (Madden, 2012). These librarians, developers, archivists, artists, and researchers recognize that the custom-built architectures and bespoke data structures of earlier digital collections development are unsustainable. Their collaborations have produced now-standard technologies such as Samvera, Fedora, GeoBlacklight, Islandora 8, as well as RDF, and JSON-LD among other open schemas. A core principle animating these efforts is reusability: data, schemas, and technologies in the open era must be coherent and flexible enough to be reused across multiple digital contexts. The authors of this paper show how reuse guided the migration of the Hopkins Digital Library from an outdated isolated system to a sustainable interconnected environment in GeoBlacklight, Islandora, with metadata based in Linked Open Data. Three areas of reuse focus this paper: the creation of robust interoperable metadata; the expansion of IIIF functionality to integrate the needs of the Hopkins Geoportal's users; the development of a broadly re/usable data migration module focused on expanding a diverse community of invested users. In focusing on reusability as an organising principle of digital collections development, this case study shows how one digital curation team produced a platform that meets the changing and specific needs of an individual institution, on the one hand, and participated in and furthered the creative coherence of the open communities supporting the team's work, on the other.

*Submitted* 13 March 2022 ~ *Accepted* 22 April 2022

Correspondence should be addressed to Kathryn Gucer and Michelle Janowiecki, 3400 N. Charles Street, Baltimore, MD 21093. Email: kgucer1@jh.edu; michelle.janowiecki@jh.edu

This paper was presented at International Digital Curation Conference IDCC22, online, 13-16 June, 2022

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/



# Introduction

In late 2019, Johns Hopkins University's Sheridan Libraries (henceforth, "the library") began a project to migrate its digital collections, known as JScholarship, from a limited and underused instance of DSpace 6 to a more robust and user-friendly home in Islandora 8.[1] This choice of software vastly expanded and diversified the potential use – including reuse – of the new collections, renamed the Hopkins Digital Library (HDL). In particular, the new design opened up access to levels of the system that had been off-limits and/or inoperable in JScholarship, including the metadata and the digital infrastructure that in JScholarship was hidden under the hood and isolated from other library systems. The authors of this qualitative case study, members of the team at the library that designed and implemented the HDL (henceforth "HDL team"), focus on three ways, in three corresponding sections, that this new digital resource enables its own reuse:

1. Modelling reusable metadata;
2. Fostering Reuse through Batch Ingest;
3. Interoperability with the Hopkins Geoportal.

# Discussion

## Section One: Modelling Reusable Metadata

*This section will cover how our team modelled metadata in the Hopkins Digital Library to improve data interoperability and reuse, paying attention to how these decisions streamline data movement throughout our library ecosystems.*

Migrating to Islandora 8 gave HDL an opportunity to improve the interoperability and reuse of metadata representing its digital assets, benefiting both users and library staff. In particular, three key components of metadata design and implementation improved our overall metadata interoperability and reuse:

1. Representing and exposing our metadata as Linked Open Data through the JSON-LD syntax;
2. Selecting metadata elements from widely adopted schemas and namespaces, and;
3. Referencing controlled vocabularies commonly used within Sheridan libraries and within the archival and library professions at large.

Foremost, our switch from relatively flat, string-centric JSON records to linked data (represented as JSON-LD) presented by the migration to Islandora 8 provides the greatest gains in interoperability (and therefore reuse) of our metadata in the future. By "flat", we reference metadata with values that are not further described through nested properties in hierarchical data structures or by linking out to other entities. Linked data, or data structured as subject-predicate-object triples, is a machine-readable, flexible structure that by its nature, semantically conveys its meaning through named entities and properties with persistent URIs. Compared to the flat, string-dominated JSON records of our previous digital repository, JScholarship, the structure of linked data itself makes our data more consistent and better aligned with the entity-

[1] Islandora 8: https://www.islandora.ca/

relationships found in our other institutional data. Both of these improvements potentially increase the reuse of the data, within the library itself and outside the institution.

*Representation of a subject field in JScholarship*

{"key": "dc.subject.other",
"value": "Bronk, Detlev W. (Detlev Wulf), 1897-1975"}

*Representation of a subject field in HDL*

"http://purl.org/dc/terms/subject": [{"@id":
"https://digital.library.jhu.edu/taxonomy/term/936"}]

↓ *Links to*

{"@graph": [{"@id": "http://digital.library.jhu.edu/taxonomy/term/936", "@type":
["http://schema.org/Person"], "http://schema.org/name": [{"@value": "Bronk, Detlev W.
(Detlev Wulf), 1897-1975", "@language": "en"}], "http://schema.org/dateModified":
[{"@value": "2021-10-08T20:35:33+00:00", "@type":
"http://www.w3.org/2001/XMLSchema#dateTime"}], "http://schema.org/sameAs":
[{"@id": "http://viaf.org/viaf/23863944"}, {"@id":
"http://id.loc.gov/authorities/names/n96053230"}]}]}

**Figure 1.** Comparison of metadata in JScholarship and HDL

First, the structure of linked data improves the consistency of our data by representing entities as graphs rather than as strings. For instance, in our JScholarship records, authors were represented only as strings within individual resource records. In JScholarship, this allowed for variety in the representation of a person within the digital repository, as metadata creators often entered slightly different name strings to represent the same entity due to variations in name representation, formatting, or simple typing errors. This practice, however, discouraged accurate collocation of resources by entities (such as an individual author) and decreased findability of resources. Islandora 8 allowed us to implement Taxonomy Vocabularies to represent different types of entities as linked data graphs. Taxonomy Vocabularies in Islandora are entities with a customized set of fields to represent specific types of non-resource entities or controlled vocabularies (Islandora 8 Community, 2022).

**Table 1.** Taxonomy Vocabularies in HDL

| Taxonomy Name | Definition |
|---|---|
| Access Rights | Terms identifying if the access to a resource is public or restricted. |
| Copyright and Use | Terms using RightsStatements.org to "communicate the copyright and re-use status of digital objects to the public." |
| Corporate Body | Terms identifying an organisation or group of persons that is identified by a particular name and that acts, or may act, as a unit. |
| Family | Terms identifying multiple persons related by birth, marriage, adoption, civil union, similar legal status, or who present themselves as a family. |
| Genre/Form | Terms describing a format or artistic category of a resource, not the subject of its contents. |
| Geographic Location | Terms identifying the name(s) of a place or geographic location by which it is known. |
| Language | Terms identifying a system of written, spoken, or signed communication in a resource. |
| Person | Terms identifying an individual or an identity established by an individual, either alone or in collaboration with one or more other individuals. |
| Resource Type | Terms that broadly characterise(s) the content of the resource, regardless of its original or digital manifestation. |
| Subject | Terms that indicate a key topic of a resource (what the resource is about). |

As we are only entering information like author name once (rather than numerous times), linking our resources to taxonomy terms, not strings, greatly improves the consistency of our metadata, and therefore the findability of our resources. Moreover, it allows us to add additional information (properties/predicates) to our entities, providing more entry points for discovery and more context for our resources. While these improvements are standard for many linked data implementations, gains in consistency from linked data improves the overall findability of our resources and therefore, the reusability of our metadata and resources. The structure of linked data also allowed us to model our metadata after the entity-relationships established in our catalogue and ArchivesSpace instances, making the transfer and reuse of data between these sources significantly easier.[2] The metadata group responsible for creating our local Islandora 8 metadata schema proposed taxonomy vocabularies to implement by evaluating the relationships between entities within traditional cataloguing and within ArchivesSpace. For instance, both of these metadata practices created authority records for entities like people and corporate bodies, encouraging us to make taxonomies for these entities that captured similar information.

**Table 2.** Equivalent entities across Hopkins Sheridan Libraries[3]

[2] ArchivesSpace: https://archivesspace.org/

[3] Italicised values are represented in their systems as controlled lists rather than entities with attributes.

| HDL entities | ArchivesSpace entities | Catalogue entities |
|---|---|---|
| Access Rights | *Local access restriction type* | |
| Copyright and Use | Rights Statement | |
| Corporate Body | Agent: Corporate Entity | Authority record: Corporate names<br>Authority record: Meeting names |
| Family | Agent: Family | Authority record: Personal names (type: Family name) |
| Genre/Form | Subject: Genre/Form | Authority record: Genre/form terms |
| Geographic Location | Subject: Geographic | Authority record: Geographic names |
| Language | Language and Script | *MARC List for Languages* |
| Person | Agent: Person | Authority record: Personal names |
| Resource Type | | *Content Types Scheme* |
| Subject | Subject: Topical | Authority record: Topical terms |
| Repository Item | Archival Object<br>Digital Object | Bibliographic record |
| Collection | Resource | Bibliographic record |

Additionally, our selection of metadata elements from widely adopted namespaces like Dublin Core, Schema.org, and BibFrame gives our data semantic clarity as well as providing easy mapping between equivalent properties in other namespaces. Additionally, the prevalence of elements from these namespaces within the GLAM (Galleries, Libraries, Archives and Museum) community helps in the transfer of our metadata from non-linked data formats, given the proficiency of many information professionals with these namespaces and pre-existing community mappings between traditional bibliographic and archival metadata and namespaces like Dublin Core (Smith-Yoshimura, 2018).

Finally, selecting FAIR-compliant controlled linked open data vocabularies for subject, genre/form, and name authorities from prevalent vocabularies in our institutional bibliographic and archival data greatly improves the ability to harvest data for digital assets from pre-existing records — another important example of reuse of legacy data – while improving the interoperability of our metadata within the GLAM community. For instance, the strong interoperability of our Islandora 8 schema with MARC metadata allowed us to easily reuse data from 300 pre-existing MARC records to describe our collection of digitised videos from the Science Review, a historically notable educational television show created by the university in the 1950s and 1960s. Because we created our Islandora schema with bibliographic metadata as a reference, we were able to fairly seamlessly reuse high-yield metadata like date information, description information, name authorities from the Library Congress, and Library of Congress Subject Headings (which we converted to FAST headings) from already well-described catalogue records. Many other resources within our digital repository are also, at least partially, represented in our catalogue as MARC records or within ArchivesSpace as components of

archival collections. Harvesting metadata from these records sources for reuse in our digital repository has always played a role in our workflows, but the implementation of linked data has greatly expanded the possibilities to reuse FAIR-compliant controlled vocabularies terms already used by our archivists and librarians. This greatly aids in the reuse of URIs within our institutional information ecosystem.

## Section two: Interoperability with the Hopkins Geoportal

*This section will discuss how Hopkins Digital Library makes use of the IIIF standard to reuse images from its collection in an adjacent Geoportal project.*

Islandora 8 also provides an immediate opportunity for the reuse of our digital collection through the implementation of the International Interoperability Framework (IIIF) standard through the image viewer OpenSeadragon. In 2019, Data Services at Sheridan Libraries began working on an implementation of a GeoBlacklight repository locally called the Hopkins Geoportal, to serve as a specialised, geospatial-centric access platform for its rich collection of digitised maps and geospatial data.[4] GeoBlacklight can host a large amount of geospatial data not easily available in other formats, such as vector datasets and georeferenced rasters, that give it distinct GIS-focused utility. It can also link to and display data from other GeoBlacklight repositories, seen in examples such as the Big Ten Academic Alliance Geoportal.[5] As a result, institutions that use the software link to items' library catalogue records through their URIs, and to digitised images from their collections using the Mirador viewer to display them via IIIF manifests. It is these manifests that HDL shares with the Hopkins Geoportal, which reuses them. This means that digitised maps within Hopkins Digital Library in Islandora 8 can be easily reused through IIIF APIs in the Geoportal without storing large duplicate images in multiple repositories. Moreover, it allows these images to be accessed and viewed into two different contexts: in a general context for non-geospatial audiences and in a domain-specific context for GIS and map researchers, who may choose to overlay the map images with other kinds of geospatial data.

## Section Three: Fostering Diverse Community Reuse through Batch Ingest

*In this section, the authors discuss the development and initial reception of an ingest service within the administrative layer of HDL, with an eye towards the divers and creative reuses it has engendered.*

A core principle in the design of HDL is a robust and diverse community of users. Reuse, in all senses of this word, is essential to this community. The more people who iteratively engage with our collections, the more that the community of users, and their investment in the collections, grows. As this formulation implies, the relationship between community and reuse is symbiotic and circular, so any component of the new system that facilitates wider and repeated use helps develop the community. One component of the system that the HDL team specifically designed with an eye toward community development and reuse is the batch ingest service. This distinct component within the system allows librarians and other staff to create and update whole collections or portions of collections greater than one item. In doing so, it constitutes a huge improvement on the old system, JScholarship, and enables several important new kinds of reuse by and between two important communities, as detailed in the rest of this section:

1. The external community of librarians and software engineers who develop and support the open-source software in which HDL is built, Islandora 8;

2. The internal community of library staff at JHU.

[4] GeoBlacklight: https://geoblacklight.org/

[5] Big Ten Academic Alliance Geoportal: https://geo.btaa.org/

## External reuse: participating in the Islandora/Drupal community

Islandora 8 is both an open-source software and a community of librarians and developers grounded firmly in the principle of reuse, a distinctive feature that the designers of HDL considered in their choice of this resource over others. The Islandora Foundation describes its software as providing "open-source digital asset management" that enables people to "collaboratively author, manage, discover, and steward digital collections across data types and knowledge domains" (islandora.ca). Currently, over three hundred institutions use Islandora to host their digital collections, mostly in the United States and Canada. This already substantial community of Islandora users is, however, only a sliver of a much larger community that Islandora taps into: the more than one million software designers and engineers worldwide who support the content management system Drupal (drupal.org). Drupal provides the graphical user interface (GUI), frontend design, and much of the data modelling for Islandora 8. Like a Lego set, Drupal is made up of modules and Islandora's components are also Drupal modules. Simply put, Islandora *is* a reuse of Drupal. Islandora's integration with Drupal not only means that the two softwares are coextensive, it also means that the Islandora 8 community can piggyback on the vast resources in the Drupal community, including Drupal documentation and the Drupal developer network and the channels of communication within it. In tapping into the Drupal community, Islandora has a potential for reuse of Drupal in the broadest sense, a potential that HDL's designers recognized and used to develop the ingest module, as described in more detail below.

## Internal Reuse: Building a diverse community at JHU

Early in the design of HDL, both the librarians and the software engineers familiar with JScholarship's limitations recognized that, in order to expand use in the new system, we would have to broaden access among library staff to the tools for batch ingest and update (itself a form of reuse) of collections. The use of those tools in JScholarship was limited to staff members in two roles:

1. A digital metadata specialist with knowledge of the scripting language Python, the command-line, and access to the DSpace API;

2. A software engineer. The fact that the librarians, collections owners, archivists, and curators–subject specialists who owned and best understood these collections–were unable to create and update them quietly discouraged the collections' use and reuse.

In order to give more curatorial control to these staff members, the HDL team created a batch ingest service which provides a user-friendly GUI (Figure 2) for librarians to create and update their collections. The technological components of this service are, as described above, themselves a reuse of the open-source content management system, Drupal. More precisely, the ingest services reuses three Drupal modules: Migrate Source UI, Drupal migrate module, and the migrate source csv module.[6] In combination, the HDL team's customised reuse of these three modules, enables library staff to create and update collections in the system using spreadsheets (see list of spreadsheets circled in red in Figure 2. in a .csv format. This method of ingest is librarian- and archivist-friendly because these staff are trained to organise data in these formats. The service does not require command-line knowledge or running Python scripts and thereby expands use and reuse to far greater numbers of staff.

[6] Migrate Source UI: https://www.drupal.org/project/migrate_source_ui; Drupal migrate module: https://www.drupal.org/project/migrate; migrate source csv module: https://www.drupal.org/project/migrate_source_csv

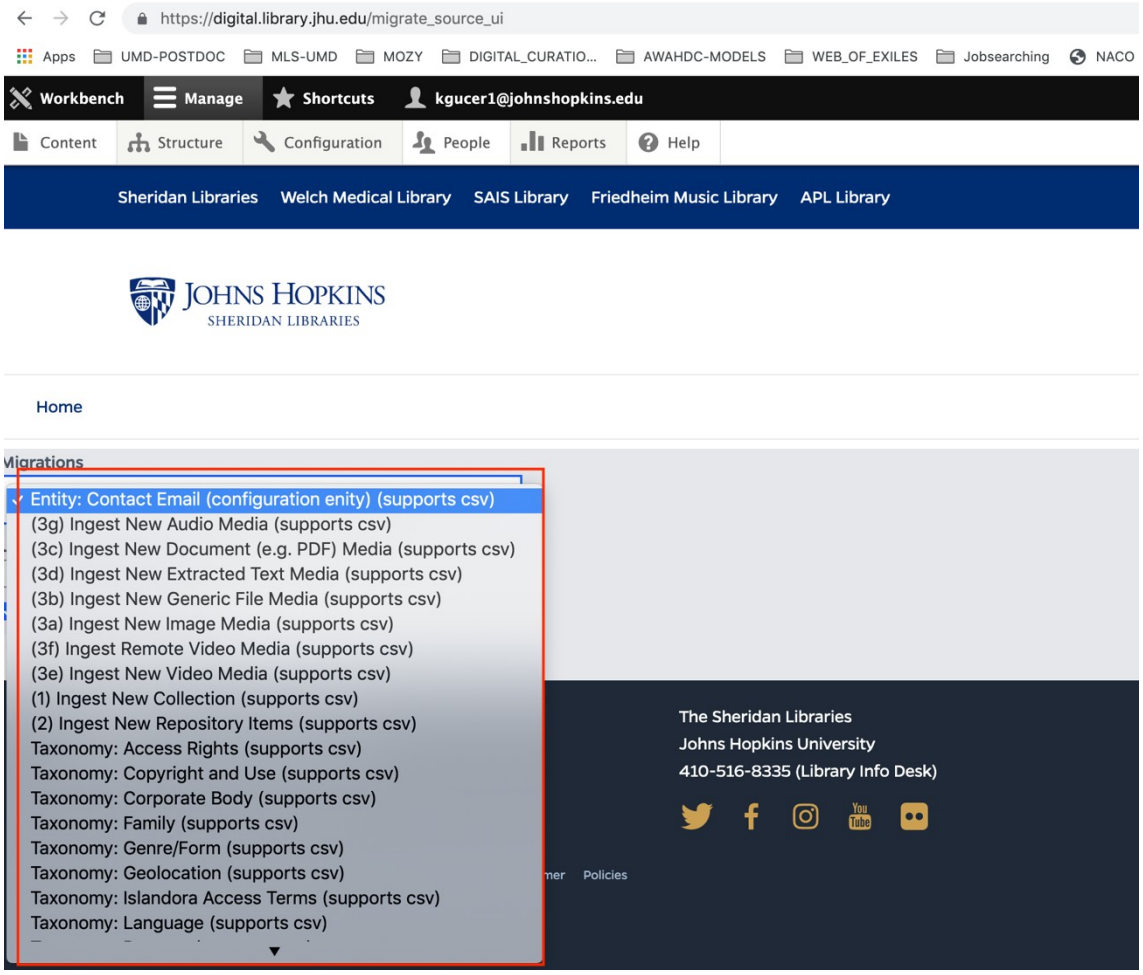


**Figure 2.** HDL Ingest GUI with list of spreadsheets highlighted in red.

Among the many advances over JScholarship in HDL, this ingest service enables library staff to iterate and build on one another's work and the very specialised knowledge that goes into it. For instance, the ingest service provides the interface by which these staff members create and reuse a common reservoir of taxonomy vocabularies, such as "person," "subject," and "corporate body" as described in section one. The batch ingest service defines and sets in motion a collective workflow by which librarians create taxonomy terms that, for example, distinguish Jane Seymour, the modern-day television and film actor, from Jane Seymour, the queen of England and second wife of Henry VIII. Embedded in the system are instructions and prompts making it clear how to make such entities independent, transparent, and reusable by structuring them according to external standards, including FAST and VIAF, among others.[7] Most importantly from the perspective of reuse, these taxonomy terms are stored in this system (Figure 3) in a list that staff can easily consult and update. This common, reusable knowledge store makes for a huge advance over the workflow in JScholarship, where only one digital metadata specialist could make bulk changes to taxonomy terms and there was no transparent reservoir that staff could access. With only one person at the gate, so to speak, to batch ingest via API in JScholarship, this gateway became a bottleneck. Updating one subject taxonomy in the system – for instance disambiguating Jane Seymour (the actor) from Jane Seymour (the queen) – could take this specialist over a month, because they would have to hunt down every instance of this name, consult the subject specialists separately, and then write, test, and execute one or more Python scripts to implement these changes. By contrast, in HDL, there is far less room for

[7] FAST: https://fast.oclc.org/searchfast/ and VIAF: https://viaf.org/

error and ambiguity in creating and structuring these terms, as the workflow enables specialists to create robust uniquely identified terms that library staff can collectively verify and update.

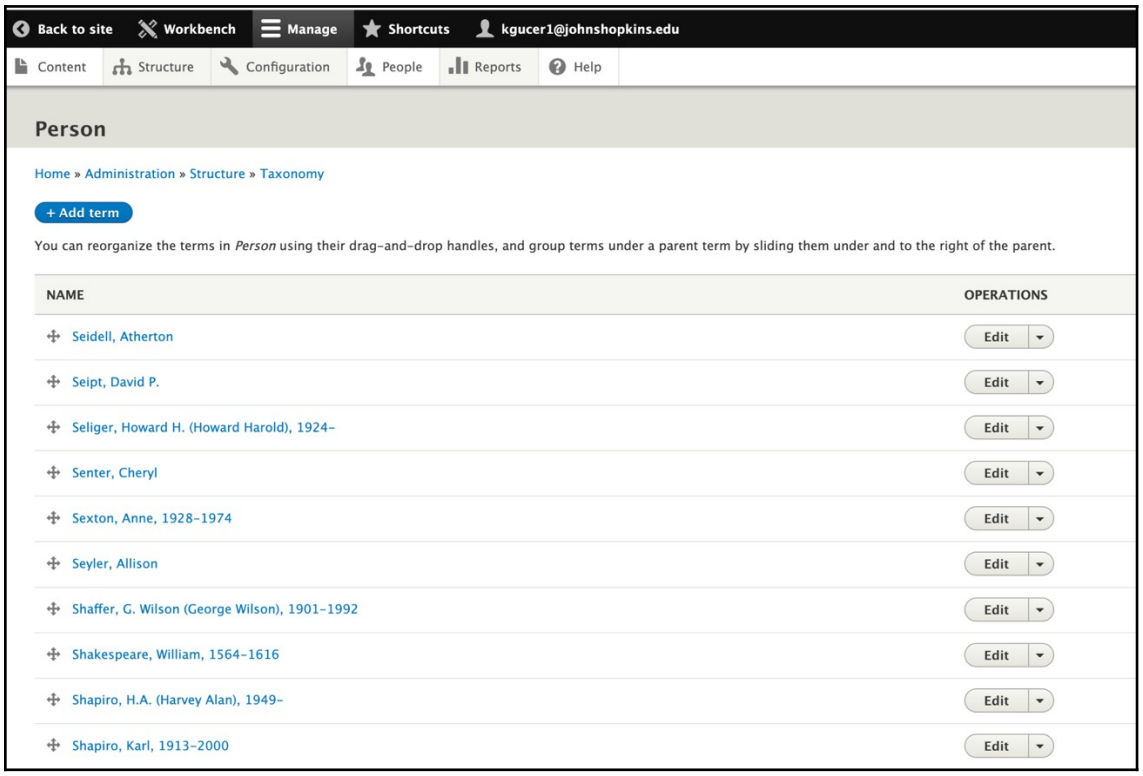


**Figure 3.** Terms from Person entity in Hopkins Digital Library.

# Conclusion

In telling the story of how one team incorporated reuse, and the related FAIR principle of interoperability, into the digital collections development at Johns Hopkins University, this case study illuminates the practical application of reusability as an organising principle of a digital asset management system. It addresses core themes in digital collections development including: balancing reusability between wide and narrow audiences; evolution in tools for enabling reuse and; how GLAM institutions are employing the notion of collections as data to reinvent engagement with people as users and re-users. Moreover, our work comes at an opportune moment, when many of those GLAM institutions are eyeing newly introduced digital collections in the open source software and open standards the authors discuss – including the Carnegie Mellon Digital Collections[8], University of Toronto at Scarborough Digital Collections[9], and the University of Nevada, Las Vegas[10]. For these potential re-users, this paper illuminates in detail the process by which a similar institution opened up their digital collections to a broader and more diverse re-/usership than ever before.

8 https://digitalcollections.library.cmu.edu/node/23
9 https://core.digital.utsc.utoronto.ca/
10 http://digital.library.unlv.edu/node/6

## Acknowledgements

The authors would like to acknowledge the following colleagues at JHU who helped to build the elements of HDL discussed in this paper: John Abrahams, Aaron Birkland, Lena Denis, Kristin Diehl, Jared Galanis, Elliot Metsger, Anna Loewenthal, Bethany Seeger, and Hanh Vu.

## References

[website] Islandora 8 Community. (2022, April). *Metadata in Islandora*. Islandora Documentation. https://islandora.github.io/documentation/user-documentation/metadata/

[journal article] Madden, L. (2008). Applying the digital curation lessons learned from American Memory. *International Journal of Digital Curation*, 3(2). doi: 10.2218/ijdc.v3i2.63

[journal article] Smith-Yoshimura, K. (2018). Analysis of 2018 International Linked Data Survey for Implementers. *Code4Lib Journal*, 42. Retrieved from https://journal.code4lib.org/articles/13867