Leveraging the flexibility of Islandora to create a dual-use digital repository

Alicia Esquivel

Digital Scholarship Librarian at Paul V. Galvin Library, Illinois Institute of Technology, USA

Alicia Esquivel is the Digital Scholarship Librarian at Paul V. Galvin Library at the Illinois Institute of Technology. She previously worked as a National Digital Stewardship Resident at Chicago Botanic Garden. She holds a MSIS from the University of Texas and a BA from the University of Houston.

Paul V. Galvin Library, Illinois Institute of Technology, 35 West 33rd Street, Chicago, IL 60616, USA
Tel: +1 312 567 6840; E-mail: aesquivel@iit.edu

Tim Fluhr

Systems and Institutional Repository Librarian at Paul V. Galvin Library,
Illinois Institute of Technology, USA

Tim Fluhr is the Systems and Institutional Repository Librarian at Paul V. Galvin Library at the Illinois Institute of Technology. He holds a master of science in information from Florida State University and a master of arts in social justice and community development from Loyola University, Chicago. Prior to joining IIT, he worked in the information technology field for over 15 years.

Paul V. Galvin Library, Illinois Institute of Technology, 35 West 33rd Street, Chicago, IL 60616, USA
Tel: +1 312 567 7997; E-mail: tfluhr@iit.edu

Adam Strohm

Director of University Archives and Special Collections at Paul V. Galvin Library, Illinois Institute of Technology, USA

Adam Strohm is the Director of University Archives and Special Collections at Paul V. Galvin Library at the Illinois Institute of Technology. Adam has also worked at the Newberry Library and Northwestern University Library. He earned his BS from Indiana University of Pennsylvania and his MSLIS from University of Illinois at Urbana-Champaign.

Paul V. Galvin Library, Illinois Institute of Technology, 35 West 33rd Street, Chicago, IL 60616, USA
Tel: +1 312 567 5107; E-mail: astrohm@iit.edu

Abstract   In 2019, Paul V. Galvin Library at Illinois Institute of Technology in Chicago launched a new iteration of the university's institutional repository. This single instance of Islandora serves as both the traditional institutional repository and a digital collections portal for Illinois Tech's University Archives and Special Collections. The migration from an existing DSpace repository was completed in-house by librarians without vendor assistance. This paper describes the migration process from DSpace to Islandora and explores some of the issues and decisions that arose when attempting to use a single instance of Islandora for two sometimes conflicting purposes, including questions of user interface, descriptive metadata and treatment of compound objects.

INTRODUCTION

In the autumn of 2019, Paul V. Galvin Library at Illinois Institute of Technology in Chicago launched a new iteration of the university's institutional repository, repository.iit (https://repository.iit.edu). The project, which took multiple years and the efforts of more than six library professionals, included a migration from DSpace to Islandora, and a streamlining of collection structure and permissions. The migration accomplished a more transformative goal as well: providing the library's University Archives and Special Collections (UASC) unit with its first digital collections platform for digitised and born-digital archival material. These two uses for the new version of the repository were part of the migration process from the start, so software evaluation, project planning and implementation were all completed with these adjacent but not always harmonious use cases in mind. This dual approach introduced complications and concessions into the migration and elongated the project, but the result is an institutional repository with a more sustainable path to future use, added efficiency in library technology maintenance and support, and a success story in collaboration and problem solving during a time of transition and change at Galvin Library.

BACKGROUND

The staff at Galvin Library launched an institutional repository, share.iit, based on the DSpace platform in 2012. Partially as a result of low faculty buy-in, the repository was envisioned and advertised as a home not only for faculty and staff scholarship, but also an ersatz digital asset management system for the university that acted as a place to store digital files and manage access to them. The latter, less traditional use of the institutional repository garnered far more interest, and resulted in complicated collection and permissions structures, with much of the material in the institutional repository restricted to library staff or campus groups with changing membership.

The share.iit name and URL were changed when anecdotal evidence suggested that a re-branding might find more acceptance by university faculty, but after two years of existence, the DSpace repository was still largely used by library staff, both to disseminate their own work as well as to ingest the electronic theses and dissertations (ETDs) that were the library's responsibility to store and make accessible. The majority of faculty and staff self-deposit came from a few specific departments and faculty members, and outreach efforts were increasingly met with responses that focused on what the system ought to do, or what other university's institutional repositories were able to do. By this point, the library's Digital Initiatives Group began to question the library's ability to sustainably support the software, with no one remaining on staff well-versed in how to maintain, update or troubleshoot the software. The same group also questioned whether DSpace offered the features and experience that the library needed and that users expected, and a team of librarians started to explore other options.

The UASC unit began as Illinois Tech's University Archives, which was formally instituted in 1998. The Archives' collections number over 1,000, containing rich documentation of the history of Illinois Tech, including the Institute of Design, founded by László Moholy-Nagy in 1937 as the New Bauhaus, and the transformation of the Department of Architecture and Illinois Tech campus by modernist legend Ludwig Mies van der Rohe. The collections also document the two predecessor institutions that merged to create Illinois Institute of Technology in 1940: Armour Institute of Technology and the Lewis Institute, both founded in Chicago in the 1890s. Despite a number of systematic efforts to digitise archival material, the archives never had a digital asset management system, or a single platform for digital collections presentation. The DSpace repository launched in 2012 was not seen as a suitable place to make digitised material available due to its small thumbnails, which

forced users to download a file to get a good look at it, and the simplified application of the Dublin Core metadata schema, which did not allow for the contextual information that can be so important in describing archival material. At the same time that the Digital Initiatives Group was looking into other institutional repository possibilities, the group began to think about more comprehensive digital asset management and a means of providing access to digitised and born-digital archival content. It became clear that these two needs would be most efficiently met by the same system.

SOFTWARE SELECTION

A small team was formed to determine the next steps in the process. The team took stock of the library staff's technical expertise and infrastructure with regard to programming language and migration tool proficiency, software platform familiarity, self-hosting capacity and comfort level with open source software. All of these became considerations in the evaluation and selection of a software platform for the institutional repository. A set of user stories was developed to set the scope of the project and define high-level requirements for an end product. The resulting business requirements document was used in combination with the aforementioned internal evaluation to guide the platform search.

Internal evaluation revealed that the library had adequate technical infrastructure and systems administration expertise to self-host. Library staff already had a high degree of comfort working with open source software, the Drupal content management system (CMS) and LAMP software stacks more generally. Additionally, the team felt that there was enough scripting language, data transformation and collective data migration experience on-staff to move from DSpace to another system without the help of consultants or vendors.

Two other factors played a role in the decision. Cost was a necessary consideration:

while a variety of both paid and open source institutional repository platforms were available, the Islandora stack was expected to be the least cost-prohibitive in terms of actual out-of-pocket spending and staff time. Islandora was a free and open source framework, supported by a large and active user community, built on the Drupal CMS that was already being used for the library website.

Secondly, based on the user stories, staffing limitations and internal needs, the migration required a highly customisable product that could meet identified user needs and align with the library's technical infrastructure and expertise. The Islandora framework offered more flexibility than competing products due in large part to its Drupal CMS base. Islandora itself consists of a suite of Drupal modules that communicate with Fedora Repository and Apache Solr while leveraging Drupal's front end for display. Islandora's integrations with Drupal's core functionality can facilitate multiple displays and layouts, supplemental content creation, featured content displays and one-off collection creation. It was determined that these features would allow the library to satisfy the needs of both the institutional repository and UASC in a single site with a unified, Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) compliant metadata schema.

USER EXPERIENCE IN A DUAL-USE ENVIRONMENT

As the same instance of Islandora would be used as both the traditional institutional repository and as a home for digitised archival material, a decision had to be made early in the migration process about how to organise the different types of digital content within the Islandora system. Collections can be set up in Islandora to organise objects based on type of material, format or any other criteria. Illinois Tech's DSpace repository had a complicated hierarchy of communities and collections,

and simplification of this structure was strongly desired. At the same time, there was a need for at least some use of collections to facilitate things like the broad assignment of access restrictions and slight differences in metadata schema application. At the most basic level, it seemed apparent that there should be some sort of division between the scholarly work in the institutional repository and digitised archival material from UASC. Eventually, the decision was made to put ETDs in their own collection as well, the reasoning behind which will be discussed in the section on permissions and access.

There were significant user experience questions inherent in the use of collections to separate objects in the repository: how to curate the experiences of users searching or browsing on one or the other 'sides' of the repository, and how apparent the division between the two ought to be to the front-end user. Drupal's flexibility meant that it would be possible to create a repository landing page on top of Islandora that offered highly customised entry points for users. The implementation team looked at the landing pages of other institutional repositories that included archival material, including Florida State University's Diginole (https://fsu.digital.flvc.org/), which seemed to have the most directly relevant example of a division between the institutional repository and archival collections. In a more explicitly bifurcated example like Diginole, the entry points for the institutional repository and UASC collections seemed likely to lead users more directly to what they are looking, but there was concern that too much division between the two could cut down on serendipitous discovery of potentially relevant objects from different sides of the repository. Ultimately, setting up a distinction between these collections via multiple entry points seemed more meaningful to library staff than a general user. The decision was made to create a single landing page, with a repository-wide search box that would return objects across the institutional repository, regardless of collection. Smaller links for other means of returning objects were added below the search box, giving who knew what they wanted the option to focus on archival material, department-specific scholarship, or ETDs right away (Figure 1). The Islandora Collection Search module (https://github.
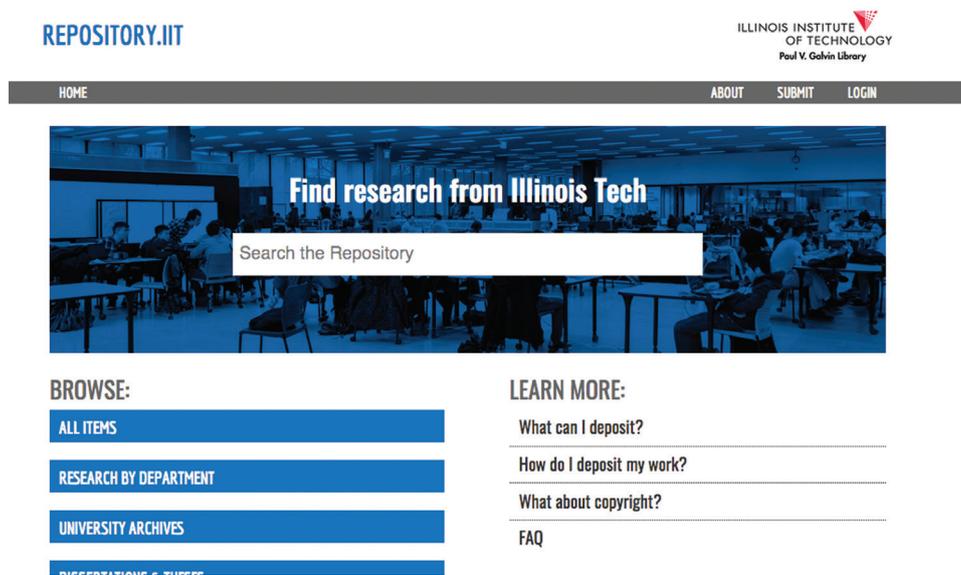
Figure 1: The repository.iit landing page

com/discoverygarden/islandora_collection_search) was used to place a block at the top of most pages that permitted users to search within specific collections or across the whole repository at any point in their visit.

METADATA

Metadata was another challenge in the two-pronged approach. One of the reasons that Islandora was chosen was its use of the Metadata Object Description Schema (MODS) and the opportunity for more detailed description of digitised archival material than DSpace's application of Dublin Core was able to offer. But using the same MODS schema to describe things on both 'sides' of the repository meant that the eventual solution would require making some concessions in specificity for both the more traditional institutional repository submissions and archival content. Galvin Library's former Metadata and Discovery Librarian created a comprehensive metadata registry, and was able to find appropriate applications of unused or repeatable fields to express archival metadata that would be irrelevant or non-existent in institutional repository objects, such as archival collection, or a link to the relevant finding aid in UASC's Archivesspace database. The Islandora Multi-Importer module (https://github.com/mnylc/islandora_multi_importer) provides the ability to create multiple metadata twig transformation templates and choose which file to apply on ingest, so MODS records for institutional repository material and archival material can be created in alignment with the specific applications of MODS developed for each. The use of repeating fields, however, impacted Solr's faceting: results from both sides of Islandora would show up in the same facets, and this necessitated more general labels for fields that were used in slightly different ways for different types of objects. For example, the MODS <relatedItem type="host"> element was used in archival objects to define the archival collection of which the object was a constituent and in institutional repository objects for the conference or project for which a paper was written (as seen in the excerpt of the site's metadata registry in Figure 2). The 'Part Of' label that we used for this field was an imperfect, but acceptably broad descriptive term for both uses.

COLLECTION VIEWS AND SOLR SEARCHES

Islandora and its suite of modules offer a few options for collection views, including a Solr-powered one intended to mimic the look and feel of a page of search results. But, the search results interface offered by this view could not replicate the look of an actual Solr search on the same site. Exploration of how other Islandora users had handled collection views found that the developers of Islandora at the Islandscholar

Fedora IndexField name	MODS field	Item view label	Definition
mods_relatedItem_host_title_titleInfo	<relatedItem type="host"><titleInfo><title>	Part of	Title of journal, conference that object is associated with, OR , name of UASC collection or other collection to which the original material belongs

Figure 2: An excerpt from the repository.iit metadata registry, showing a single field with multiple uses

repository at University of Prince Edward Island (http://www.islandscholar.ca/) had already implemented one of the solutions brainstormed by the implementation team by using canned Solr searches as substitute collection views, and making the default view essentially unreachable unless someone keys in the URL by hand. It was decided that a Solr search would harmonise best with the rest of the site, and would also be more scalable as collections grew. This approach was taken for ETDs, as well as the topical sub-collections eventually created within the UASC collection (Figure 3). To get descriptions of the UASC sub-collections for these canned searches, URL-specific Drupal blocks were created and triggered to appear when a URL contains a search string for a particular collection. This solution felt over-engineered and inelegant, but it was preferable to the other options, and any complexity was masked to the user, for whom the search and browse experience would be harmonised across the site.

PERMISSIONS AND ACCESS POLICIES

The permissions in the DSpace repository were highly configured to meet the needs of different shareholders over the years, including department-specific user groups created for objects not meant for public sharing. This resulted in a large permissions matrix that was becoming unwieldy to manage and would be difficult to maintain going forward in Islandora. To cut down on

Figure 3: The University Archives and Special Collections landing page, including topical sub-collections

administrative duties regarding permission management, the team decided to create a stricter policy for user level permissions in Islandora. In the spirit of open access, items in the Islandora repository would only have three permission levels: open access, IIT-only access (with only metadata visible to users outside of Illinois Tech) and embargoed access. Islandora's integration with the Central Authentication Service (CAS) protocol meant that authenticating users and creation of user accounts could be offloaded on the university's existing CAS directory, something that was not possible in DSpace. During migration, authors of items with unique or complicated permissions conditions were contacted directly about the migration and changes in permission levels. These objects were then either made more openly accessible in Islandora or, if it was determined that ingest into DSpace was not meant to make these objects accessible, but only as backup file storage, removed from the repository completely.

The IIT-only access level was created largely to comply with existing restrictions put upon the ETDs, which the library was obliged to restrict to current Illinois Tech students, faculty and staff. This was achieved by making the original PDF bitstreams in the ETDs collection available only to users who authenticated via the university's CAS system.

For the archival materials, high-resolution TIFF and WAV files have been restricted to UASC staff and repository administrators with lower-resolution derivatives available for all to download. There were cases, however, in which some digitised or born-digital archival material could not be discoverable on the web, usually due to issues of uncertain rights or terms agreed upon with copyright holders. Leaving these objects outside of Islandora, however, would mean maintaining two systems for digital asset management and retrieval, so a UASC staff portal was created, with any item visibility or discoverability within the collection restricted to UASC staff. This also provided a holding pen of sorts for any objects that had been ingested and needed to be held back from public visibility temporarily for any reason. This staff portal collection has been used sparingly, but it has already been of help in touchy issues of donor relations or murky copyright situations.

DEPOSIT

The DSpace repository featured a self-deposit process that asked authors to create their own metadata, select a Creative Commons licence, and upload their own objects to the repository without the need for input or reconciliation by library staff. This self-deposit model frequently involved some level of intervention by library staff, especially for first-time depositors, either in the form of discussions and quality control prior to ingest, or, less frequently, alterations to metadata or other edits after ingest. The initial migration plan was to retain the self-deposit model in an attempt to maintain an 'on-demand' deposit process. The implementation team, however, was unable to use Islandora's XML Forms functionality to create an ingest form that aligned with the library's application of the MODS metadata standard. The field logic required for determining how to handle multiple instances of certain fields proved a challenge to map out in the XML Forms interface, and there were applications of controlled vocabularies and element attributes that proved too complicated for what the implementation team was able to achieve with the form creator. This not only made the XML forms module untenable for self-deposit by users, but also for deposit by library staff. The decision to move to a mediated deposit was made easier by literature that indicated that self-deposit by university faculty remains low, especially in the early days of a repository, and even at institutions with an open access policy mandating that faculty deposit their work

into an institutional repository.[1,2] A Google form was created to gather metadata from submittors. If necessary, the Digital Scholarship Librarian works with submittors to ensure that the data and object are ready for ingest, and then adds the data to a comma–separated values (CSV) file to be used by the Islandora Multi–Importer module, which provides a graphical user interface for Islandora's batch functionality.

OBJECT AND METADATA MIGRATION

As the DSpace and Islandora data models handle compound objects differently, it was apparent from the start of this project that the data migration would require more than a single export and load routine. The DSpace compound objects, as an example, could not be exported and ingested in the same manner as simple, single–file objects and therefore each type would require a different ingest workflow. All DSpace metadata also required being transformed from Dublin Core to MODS. Additionally, some DSpace objects included a second XML metadata file containing local metadata which needed to be added to their respective MODS record.

Fortunately, IIT's DSpace objects were split into collections of generally similar items. The implementation team utilised DSpace command line functionality to export items and constituent metadata into collection–level directories containing object–level subdirectories. Exported collections were then inspected for uniformity and outlier object directories using a series of simple Python test scripts.

Foremost, it was necessary to identify and separate compound and single file objects because different methods for ingest were required for each type of object. A file counting script was employed to split collections into sets of simple (single file) and compound (multiple file) objects. Additional scripts were employed to locate archived or zipped objects and to separate objects by

internet media type. Example results of the preprocessing can be seen in Figure 4.

As Figure 4 displays, collections were broken down into directories of uniform objects. Metadata transforms, however, were still required. A general Extensible Stylesheet Language Transformations (XSLT) transform was created to convert Dublin Core records to MODS. Variants of the generalised XSLT were created and ran against each set of like objects in order to account for differences in metadata per collection. Single–file objects and MODS records were then zipped and batch ingested via the Islandora batch module. Compound objects, unfortunately, required manual ingest.

The final migration task required reconciling the persistent URLs of

```
Masters Projects

|------------------------> Simple
|                              |--------> PDF
|                              |--------> Other
|
| -----------------------> Compound
                               |--------> PDF
                               |--------> XLSX
                               |--------> PDF


ETDs

|------------------------> Simple
|                              |--------> PDF
|                              |--------> Other
|
| -----------------------> Compound
                               |--------> PDF
                               |--------> XLSX
                               |--------> PDF
```
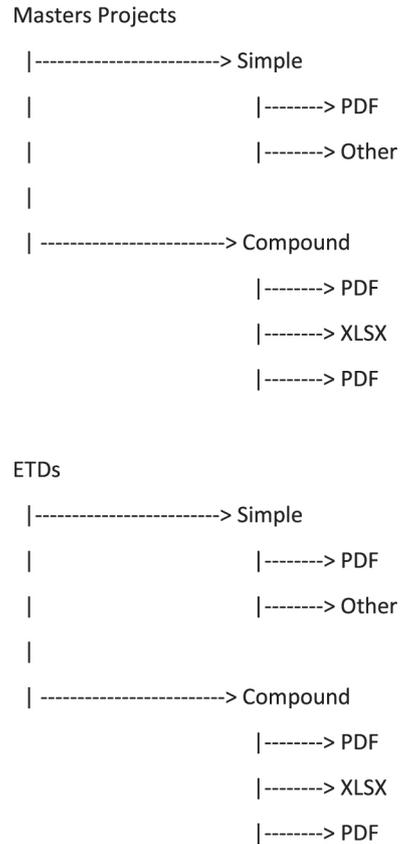
Figure 4: Results of a scripted preprocessing routine used to categorise electronic theses and dissertations (ETDs) and masters project collection items into sets of simple and complex objects

migrated objects with their new location in Islandora. Objects migrated to Islandora each included persistent Handle URIs in their metadata. A Solr query was constructed to create a CSV containing the Handle URI and Islandora Persistent Identifier (PID) of each object in the repository. A Python script was used to generate a batch Handle generation file.

COMPOUND OBJECTS

DSpace makes it possible to attach multiple files or bitstreams to a single object, which is known in DSpace parlance as a compound object. As Islandora is based in Fedora, however, it adheres to a different system in which every file is its own Fedora object, and therefore its own Islandora object. A significant number of objects in the Illinois Tech DSpace repository utilised the compound object format, with a single item containing as many as eight constituent files and a single metadata record describing them all. For the most part, these items were migrated into Islandora using the compound object solution pack (https://wiki.lyrasis.org/display/ISLANDORA/Compound+Solution+Pack), which requires creating a parent item and metadata record and attaching each bitstream as a child of that parent. The parent and each of the children are required to have their own metadata record, so the compound objects moving from DSpace to Islandora needed additional metadata records for each attached file that ended up being almost identical records in Islandora (Figure 5). Some objects, such as datasets, were zipped into a single object and ingested as a binary file instead. The migration process for these multi-file objects could not be done in bulk as Islandora required an item structure of parents and children. The team was able to export the files and metadata records from DSpace and use an XSLT script to transform the records from Dublin Core to MODS. Copies of the MODS records were then

made to correspond to each file in an object. The MODS records were edited to append the file name to the title field of the objects. These MODS records were then compiled into folders for every object. This resulted in folders with nearly identical metadata records, one for each file of a compound object. Student workers were trained to assist in the manual ingest of items which required a single metadata record to be added as a parent object followed by the rest of the children objects added in bulk and associated to that parent.

TAKEAWAYS

The migration from DSpace to Islandora took three years from the first installation of a test instance of Islandora until the official launch of the repository in the autumn of 2019. Staff turnover was part of the reason: the implementation team started as a trio, then grew to six (including a summer-long practicum student), and ended as a trio again, with only one librarian a constant throughout the process. Although the migration took longer than expected, the team was able to work iteratively and adaptively, responding to shifting needs and expectations within the library in a way that would not have been as possible with an external vendor working on the project. Through engaging with the Islandora community, including the Islandora Institutional Repository Interest Group, the team was able to identify other library's solutions to problems and triage possible solutions easily.[3] There were desired enhancements and additions to Islandora's functionality as an institutional repository that the team wanted to implement that were shelved due to time and/or staffing constraints or the team's desire to limit customisation to theme files as much as possible for the sake of sustainable and easy software updates in the future. Ideas were suggested that would have been possible with the creation of custom Drupal modules,
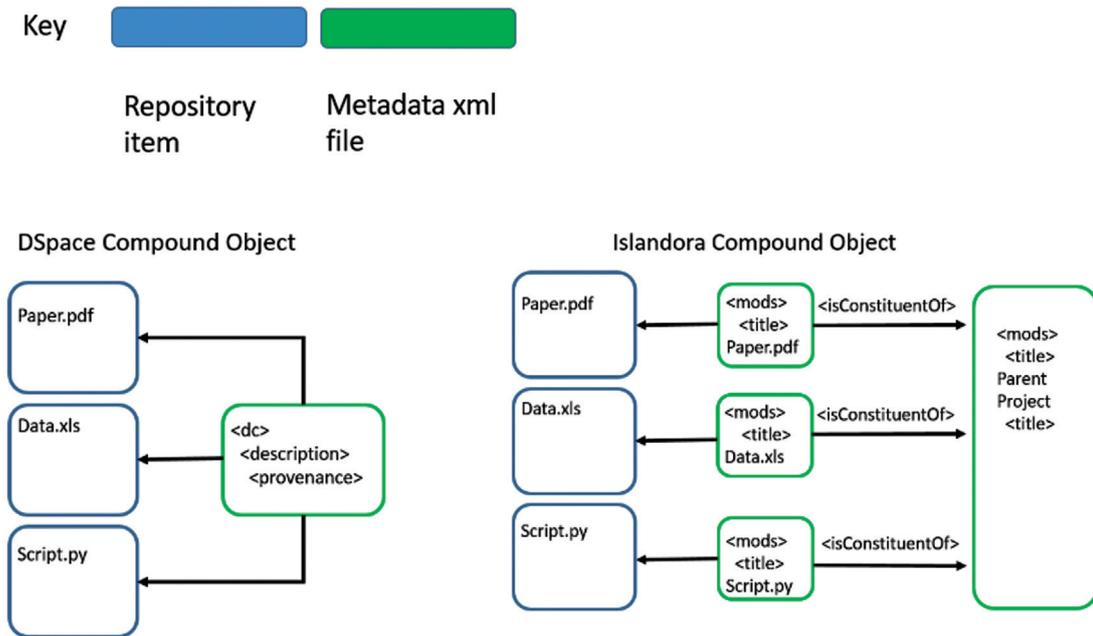
Figure 5: Comparison of compound-object structures in DSpace and Islandora

but these were not pursued unless there was an existing module that would achieve the desired result. The list of outstanding potential future enhancements provides a roadmap for continued work on Islandora and serves as a guide in evaluating the prospect of a migration to the next version of Islandora.

CONCLUSION

The process of migrating from DSpace and implementing an Islandora repository on the Illinois Tech campus was not always an easy one, but the resulting product is one that the library has been proud to promote on campus, and one that has garnered more staff and faculty interest than has been evinced on campus since the initial launch of DSpace in 2012. UASC has, as of early 2020, ingested more than 900 objects into the repository, expanding exponentially the digital archival content that is publicly available in a single

searchable platform. The institutional repository migration has already resulted in improved efficiency and interoperability in library systems and workflows, and the Islandora repository will be a central part of the library's technology plans and infrastructure moving forward. The strategies for collaboration across library units that were developed as part of this migration have already been used in other digital projects at the library and have given staff and library leadership confidence that Galvin Library can engage in broad, collaborative projects in a thoughtful, collegial, and mutually beneficial way.

Even as the Islandora repository is encouraging new collaborations between the library and campus partners and engaging more faculty interested in using the institutional repository, Galvin Library's Digital Initiatives Group is already beginning to strategise for the eventual migration to Islandora 8, a major update of the Islandora

infrastructure launched in June 2019, and the inevitable challenges and improvements it will bring. The cleanup and normalisation work that was completed as part of the migration from DSpace to Islandora has made data more understandable and interoperable. This will not only make the eventual move to Islandora 8 an easier one, but provides opportunities for making such data discoverable in systems outside of the Galvin institutional repository, and the team is eager to explore other ways to leverage data interoperability in the future both internal and external to Galvin Library and Illinois Tech.

References

1. Tillman, R.K. (2017) 'Where are we now? Survey on rates of faculty self-deposit in institutional repositories', *Journal of Librarianship and Scholarly Communication*, Vol. 5, No. 1, eP2203, available at: https://doi.org/10.7710/2162-3309.2203 (accessed 28th June, 2020).
2. Zhang, H., Boock, M. and Wirth, A.A. (2015) 'It takes more than a mandate: Factors that contribute to increased rates of article deposit to an institutional repository', *Journal of Librarianship and Scholarly Communication*, Vol. 3, No. 1, p.eP1208, available at: http://doi.org/10.7710/2162-3309.1208 (accessed 28th June, 2020).
3. GitHub (n.d.) 'Islandora Institutional Repository Interest Group', available at: https://github.com/islandora-interest-groups/Islandora-IR-Interest-Group/blob/master/README.md (accessed 28th June, 2020).