Integrating Dataverse and Archivematica for Research Data Preservation

Meghan Goodchild
Scholars Portal &
Queen's University
Canada
meghan@scholarsportal.info
ORCID 0000-0001-7988-8046

Grant Hurley
Scholars Portal
Canada
grant@scholarsportal.info
ORCID 0000-0003-0172-4847

Abstract – Scholars Portal sponsored Artefactual Systems Inc. to develop the ability for the preservation processing application Archivematica to receive packages from Dataverse, a popular repository platform for uploading, curating, and accessing research data. The integration was released as part of Archivematica 1.8 in November 2018. This paper situates the integration project in the broader context of research data preservation; describes the scope and history of the project and the features and functionalities of the current release; and concludes with a discussion of the potential for future developments to meet additional use cases, service models and preservation approaches for research data.

Keywords – research data; Archivematica; workflows; Dataverse

Conference Topics – Collaboration: a Necessity, an Opportunity or a Luxury?; The Cutting Edge: Technical Infrastructure and Implementation

I. Introduction

Between 2015 and 2018, Scholars Portal contracted Artefactual Systems Inc. to develop an integration between Dataverse, a popular repository platform for uploading, curating, and accessing research data, and Archivematica, an open source workflow application for creating preservation-friendly packages for long-term storage and management. Scholars Portal is the information technology service provider for members of the Ontario Council of University Libraries (OCUL), a 21-member consortium of academic libraries in the province of Ontario, Canada.1 Founded in 2002, Scholars Portal is funded by OCUL members and operated under a service agreement with the University of Toronto Libraries. Our services support both research data management via a hosted, multi-institutional instance of Dataverse2 and digital preservation services via Permafrost,3 a hosted Archivematica-based service that pairs with the OCUL Ontario Library Research Cloud (OLRC) for preservation storage.4 The Dataverse-Archivematica integration project was initially undertaken as a research initiative to explore how research data preservation aims might functionally be achieved using Dataverse and Archivematica together. The results of a proof-of-concept phase were developed into a working integration released as part of Archivematica version 1.8 in November 2018. This paper situates the integration project in the broader context of research data preservation in theory and practice; describes the scope and history of the project and the features and functionalities of the current release; and concludes with a discussion of the potential for future developments to meet additional use cases, service models and preservation approaches for research data.

II. Research Data Preservation in Context

In this paper, the term "research data" refers to a broad set of potential outputs from research activities across sectors and disciplines. The key uniting characteristic is that these materials stand as unique evidence supporting a set of research findings, whether scholarly, technical, or artistic [1].

[1] Scholars Portal: https://scholarsportal.info/.

[2] Scholars Portal's Dataverse instance: https://dataverse.scholarsportal.info/.

[3] Permafrost: https://permafrost.scholarsportal.info/.

[4] Ontario Library Research Cloud: https://cloud.scholarsportal.info/.

Furthermore, these data may constitute the research findings themselves, such as in the publication of statistical or geospatial data. The communities of stakeholders who value research findings depend on the maintenance of original data sources in a trustworthy manner that privileges ensuring their continued authenticity, availability and reliability into the future. These concepts have been codified within the sector as the FAIR Principles for research data: findable, accessible, interoperable, reusable [2]. While the FAIR Principles do not specifically cite long-term preservation as a requirement, preservation activities are crucial to the continued ability to discover, access and use research data into the future [3]. The FAIR principles therefore link to the stewardship responsibilities that repositories take on behalf of stakeholders: in order to fulfill the FAIR principles, organizations with access to sustained resources and infrastructure must commit to ensuring the long-term maintenance of the materials under their care.[1] The requirements for this maintenance are outlined in standards such as the Open Archival Information System (OAIS) reference model (ISO 14721)[2] and audit and certification frameworks including CoreTrustSeal,[3] nestor,[4] and *Audit and Certification of Trustworthy Data Repositories* (ISO 16363).[5] Repositories with stewardship responsibilities therefore seek to translate audit and certification requirements into repeatable practices to ensure that data are kept reliably into the future. A series of interrelated stages make up the lifecycle required for responsible data curation and preservation over time, including creation and receipt, appraisal and selection, preservation actions, storage, and access and discovery [4]. One tool that implements some of these stages of the lifecycle is Dataverse.[6]

Dataverse is developed and maintained as an open source project by the Institute for Quantitative Social Science (IQSS) at Harvard University since 2006 [5]. A large open Dataverse instance is hosted by IQSS and Harvard University Library.[7] Fifty individual known installations of Dataverse exist throughout the world as of the time of writing [6]. While Dataverse was developed by members of the social science community, its use is not limited to any specific disciplinary area [5]. Users can deposit and describe their data files using Data Documentation Initiative (DDI)[8] and Dublin Core-compliant standards, as well as certain discipline-specific metadata standards,[9] generate unique identifiers and data citations, and assign access permissions and license terms. Institutions can enable self-deposit or mediated workflows for depositors, and offer Dataverse to faculty members and researchers as a method of fulfilling funder requirements to deposit data in an accessible repository. Published datasets are searchable and downloadable and tabular data files can be explored using visualization tools within the platform itself.

Dataverse includes a suite of functions that contribute to the ability of a stewarding organization to reliably preserve research data. When it comes to data receipt, it enables efficient capture of materials from a depositor's individual computing systems through user-friendly upload tools, which tackles a major initial barrier of accessing data from the risky (and often inaccessible) environments of personal computers or local network storage [7]. Depositors can also describe and contextualize their submissions through a variety of metadata fields and by

[1] See also the Australian Research Data Commons' *FAIR self-assessment tool*: https://www.ands-nectar-rds.org.au/fair-tool.

[2] ISO 14721:2012 (CCSDS 650.0-M-2) *Space data and information transfer systems -- Open archival information system (OAIS) -- Reference model*.

[3] *Core Trustworthy Data Repositories requirements*, https://www.coretrustseal.org/wp-content/uploads/2017/01/Core_Trustworthy_Data_Repositories_Requirements_01_00.pdf.

[4] nestor seal for trustworthy data archives: https://www.langzeitarchivierung.de/Webs/nestor/EN/Services/nestor_Siegel/nestor_siegel_node.html.

[5] ISO 16363:2012 (CCSDS 652.0-R-1) *Space data and information transfer systems -- Audit and certification of trustworthy digital repositories*.

[6] Dataverse: https://dataverse.org/.

[7] Harvard Dataverse: https://dataverse.harvard.edu/.

[8] Data Documentation Initiative: https://www.ddialliance.org/.

[9] See *Dataverse user guide*, Appendix, "Metadata references," http://guides.dataverse.org/en/latest/user/appendix.html.

linking to related publications and datasets.[1] All user-submitted files receive MD5 checksums upon receipt that can enable verification of file fixity over time [8]. File format identification is also conducted as part of the Dataverse ingest workflow. Processes for file format identification include an internal service to identify tabular data files, the use of JHOVE's[2] file format validation functionality, and if these are unsuccessful, an attempt to identify based on file extension. All of these processes result in the display of the file's MIME type in the Dataverse application's database and interface [9]. The well-formedness and validity of a particular file are not recorded as an output from JHOVE.

The identification of tabular data files informs additional tabular data transformation functions. Tabular data formats (e.g., SPSS, STATA, RData, CSV, and Excel) are converted to non-proprietary tabular text data files (.tab) upon ingest, and citation-related metadata and DDI variable-level metadata are created for the tabular files [10]. Dataverse converts tabular data files as accurately as possible with the caveat that some commercial applications like SPSS have not published their specifications [11]. Tabular files also receive UNF checksums that can be used to verify the semantic content of the derivatives [12]. Users can download the data files in the original format as uploaded by the depositor, and/or in the derivative formats (tab-delimited or RData) created by Dataverse. In addition to exporting the DDI variable-level metadata as an XML file, users can also export a DDI-formatted HTML codebook for the entire dataset that also contains the variable-level metadata.

Initiatives in research data preservation, including those using Dataverse, emphasize the necessity of storing and monitoring datasets independently from the submission and discovery platforms with which users commonly interact. This approach appears to be informed by an interpretation of the OAIS reference model, which emphasizes the flow of received data as Submission Information Packages (SIPs) into stored and monitored units of content information as Archival Information Packages (AIPs) for preservation and Dissemination Information Packages (DIPs) for access. These packages may be logical rather than physical: their components may not have to be stored together so long as the total package can be retrieved and independently understood by members of the designated community [13]. Therefore, institutions could potentially use Dataverse or other repository software as an OAIS-type archive even if that software does not create and maintain physical AIPs. However, OAIS and related certification frameworks also identify in a broad sense what functions preservation systems should perform, and these features may only partially exist in a software package like Dataverse that is designed primarily for receipt, access and discovery. Creating platform-independent preservation packages means that institutions can generate and manage preservation metadata, use more than one managed method for storage, run preservation-supporting functions at ingest and over time, and audit and maintain stored packages without depending on a single system to perform all of these tasks in addition to user-facing functions.

Research on the subject of research data preservation has emphasized the desirability of storing and managing independent preservation packages. A white paper authored by members of the Canadian Association of Research Libraries (CARL)'s Portage Network Preservation Expert Group theorizes the disaggregation of OAIS-type functions among a set of potential preservation service providers who take care of particular functions such as archival storage, while communicating the results of these efforts back to a centralized administrative agency [14]. In the United Kingdom, Jisc's series of three *Filling in the Preservation Gap* reports specifically investigate the use of Archivematica in service of preserving research data.[3] A series of test implementations at the University of York and University of Hull were deemed successful and Archivematica was among the preservation providers tested with the Jisc's Research Data Shared Service pilot [15]. Therefore, Dataverse's functions primarily map to the "Producer" end of the OAIS model, where materials

[1] See also published guidance on metadata in Dataverse: *Dataverse North metadata best practices guide*, https://portagenetwork.ca/wp-content/uploads/2019/06/Metadata_V1.1_EN.pdf.

[2] JHOVE: http://jhove.openpreservation.org/.

[3] *Filling the preservation gap* project page: https://www.york.ac.uk/borthwick/projects/archivematica.

are negotiated and accepted for ingest and some baseline preservation-supporting functions are performed. Further research is required on how platforms like Dataverse might fulfill the requirements of the Producer-Archive Interface Methodology Abstract Standard (PAIMAS)[1] and Producer-Archive Interface Specification (PAIS)[2] for structuring producer-archive interactions.

Data repositories using Dataverse are taking steps to export data and metadata from Dataverse for additional processing and/or storage, primarily as physical packages. In the Netherlands, DANS' DataverseNL service exports packages using the BagIt specification[3] to their EASY preservation repository [16]. The Qualitative Data Repository (QDR) at Syracuse University is taking a similar approach with the development of a proof-of-concept implementation of exported OAI-ORE metadata and zipped Bags from Dataverse [17]. The Odum Institute at the University of North Carolina uses scripts to push data packages to iRODS,[4] which performs preservation processing and storage replication [18]. The Dataverse software itself also includes the ability to transfer exports as Bags to DuraCloud, a hosted service for replication to cloud storage providers, as well as to the file system[5] [19].

The Dataverse-Archivematica integration takes advantage of the preservation-related actions that Dataverse performs and makes them available to an Archivematica-based workflow to create and store independent preservation packages. The scope and features of this integration are discussed in the following sections.

III. History and Scope of Project

A. *Proof-of-Concept*

In response to growing community interest, Scholars Portal initiated a research project in 2015 to investigate how research datasets stored in Dataverse could be processed into AIPs using Archivematica. Initial project participants included members from Scholars Portal and the University of Toronto, Artefactual Systems, IQSS Dataverse, the University of British Columbia, the University of Alberta, Simon Fraser University, and the CARL Portage Network.

Project participants conducted an initial requirements analysis and proposed a draft workflow. Artefactual Systems developed a prototype of Archivematica that used Dataverse APIs to retrieve datasets for ingest and processing in Archivematica. The proof-of-concept integration was only available through a development branch of Archivematica and presumed an automated workflow in which all datasets in a target Dataverse would be transferred and processed by Archivematica.

The initial project provided an opportunity to explore best practices related to the preservation of research data; investigate how Dataverse handles and stores data and metadata, processes derivatives and versions files, exports data and metadata; and determine how Archivematica could accept and process Dataverse dataset packages. The project also identified the use of the DDI metadata standard within Archivematica METS files for descriptive metadata. Given DDI's capacity to comprehensively describe specific characteristics related to research data for discovery and reuse, this mapping was intended to expand the scope of descriptive metadata in Archivematica METS files and make these files more hospitable to describing research data.

B. *Production Release*

In 2018, Scholars Portal sponsored further development work with Artefactual Systems to improve the original proof-of-concept design and merge it with the public release of Archivematica in version 1.8 (developed and tested using Dataverse version 4.8.6 and above).[6] Four staff members at Scholars

[1] Consultative Committee for Space Data Systems, *Producer-archive interface methodology abstract standard*. CCSDS 651.0-M-1. Magenta book, 2004. https://public.ccsds.org/Pubs/651x0m1.pdf.

[2] Consultative Committee for Space Data Systems, *Producer-archive interface specification*. CCSDS 651.1-B-1. Blue book, 2014. https://public.ccsds.org/pubs/651x1b1.pdf.

[3] *The BagIt File Packaging Format (V1.0)*, https://tools.ietf.org/html/draft-kunze-bagit-17.

[4] iRODS: https://irods.org/.

[5] Duracloud: https://duraspace.org/duracloud/.

[6] Archivematica 1.10 - Dataverse transfers: https://www.

Portal worked directly on the project. The authors served as project leads, including organizing meetings and managing project tasks, communicating with Artefactual, performing testing and analysis, and documenting discussions and results. Amber Leahey (Data & GIS Librarian) provided domain expertise related to research data management, and Dawas Zaidi (Systems Support Specialist) provided systems support. Alan Darnell (Director), Amaz Taufique (Assistant Director, Systems and Technical Operations), and Kate Davis (Assistant Director, Collections and Digital Preservation) provided administrative support. At Artefactual Systems, our primary contacts were Joel Simpson (Operations Manager & Solution Architect) and Ross Spencer (Software Developer). Joel led the requirements analysis process, acted as our main point of contact at Artefactual, tested iterations of the integration, and produced documentation. Ross was responsible for the majority of the software development in collaboration with colleagues at Artefactual. At Scholars Portal, the development project started in April 2018 and concluded with the release of the integration in November 2018. Our key project tasks included identifying and creating test datasets (discussed below), analyzing the outputs and identifying issues, and documenting the integration. The major result of the integration is that Archivematica can be configured to use a connected Dataverse instance as a transfer source location. Datasets are queried and retrieved using Dataverse's APIs and processed using the Dataverse transfer type, which contains specific processing micro-services (described in section IV. below).

The integration was designed with a series of assumptions in terms of its design. First, the design presumes a user has an account with a Dataverse instance and has generated an API token (a unique code for authentication). The same or a different authorized user (typically an archivist, librarian, or curator) also has access to an Archivematica instance and wishes to process certain datasets into AIPs for long-term preservation. This assumes the user has obtained the necessary rights and privileges to process and store dataset files independently from Dataverse. Secondly, the current design assumes that the user is interested in selecting specific datasets in a target Dataverse instance for preservation. This assumption conforms to specifications such as CoreTrustSeal that state that repositories must appraise and select data for preservation [20]. The current design does not include an automated function for ingest of all datasets within a Dataverse container, though we acknowledge that this functionality may meet additional use cases.

A single dataset in a Dataverse instance corresponds to a SIP. Individual files cannot be transferred from Dataverse for preservation. However, users can select individual files to be made into a final AIP by using the Appraisal function in Archivematica.[1] At present, only the current version of files and metadata can be selected for preservation, though Dataverse tracks versioning and provenance of metadata and file changes, with all versions retained by the system [21]. Finally, while users may choose to create a DIP as part of the Archivematica workflow, it is assumed that the version available to users in Dataverse will generally remain the one used for access. The scope of the integration did not include communication back with a connected Dataverse to write preservation metadata, or the replacement of user-submitted data with the DIP generated by Archivematica.[2] See section V. below for discussion of features identified for potential future development.

IV. Workflow and Functionality

Fig. 1 presents an overview of the workflow for the integration. Beforehand, an administrator of the target Archivematica installation must configure the Archivematica Storage Service to connect to a specific Dataverse instance. Then, Archivematica's transfer source locations can be set to filter based on query search terms or on a specific Dataverse container using Dataverse's Search API.

archivematica.org/en/docs/archivematica-1.10/user-manual/transfer/dataverse/#dataverse-transfers.

[1] Archivematica 1.10 - Appraisal: https://www.archivematica.org/en/docs/archivematica-1.10/user-manual/appraisal/appraisal/#appraisal.

[2] The latter is the case for the Archidora integration between Archivematica and Islandora. See T. Hutchinson, "Archidora: Integrating Archivematica and Islandora," *Code4Lib Journal* 39, https://journal.code4lib.org/articles/13150.

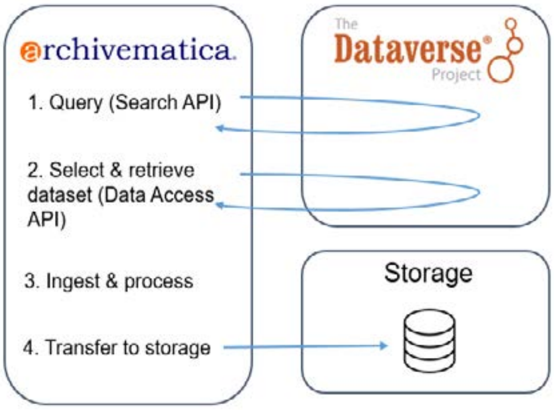


Fig. 1. Workflow for integration between Dataverse and Archivematica

To initiate a transfer, the Archivematica user sets the transfer type to "Dataverse," browses the datasets available in the Archivematica Transfer Browser, selects one dataset and starts the transfer (1). Archivematica uses Dataverse's Data Access API to retrieve a copy of the most recent version of the dataset (2). The package contains the original user-submitted data files, and if the user submitted tabular data, a set of derivatives of the original tabular files along with associated DDI variable metadata and citation metadata files describing the tabular files. Next, the Dataverse transfer type completes a set of preconfigured ingest and processing steps, including Archivematica's processing functions (3). Finally, the AIP is transferred via the Archivematica Storage Service to a connected storage system (4).[1]

Looking at the specifics of the integration, the Dataverse transfer type contains the following preconfigured ingest and processing steps:

- Creation of an initial Dataverse METS XML file describing the dataset as received from Dataverse, which includes a descriptive metadata section mapped to the DDI standard, a list of files grouped by type (original, metadata, or derivative), and a description of the structure of the files provided by Dataverse;
- Fixity checks of data files verified using the MD5 checksums that were generated by Dataverse for all user-submitted files;
- Other standard Archivematica microservices conducted as configured. These services include independent file format identification and validation, which includes mapping identified file formats against PRONOM unique identifiers. Users might also choose to apply Archivematica's preservation normalization policies.[2]

Archivematica produces a final METS and PREMIS-based XML file for the AIP (see Table 1) that copies over the descriptive metadata from the initial Dataverse METS file, outlines the relationships between original and any derivative files resulting from the tabular ingest process in Dataverse, and includes records of any actions undertaken through Archivematica's processing steps. Tabular derivatives created by Dataverse are recorded with an associated PREMIS event labeled as "derivation" in the METS file. The connected Dataverse instance's name and URI is recorded as a linked PREMIS agent in relation to the tabular derivation event.[3] Though Artefactual Systems proposed "derivation" in 2015 as part of feedback on the PREMIS controlled vocabulary for events, it has not yet been implemented in the PREMIS events controlled vocabulary.[4] Derivatives and metadata files are also identified in the METS fileGrp sections within the fileSec section.

Finally, the resulting AIP processed from Dataverse is structured in the same general format as other AIPs processed by Archivematica. As shown in Fig. 2, additional metadata files from Dataverse

[1] Storage service, *Archivematica Wiki*: https://wiki.archivematica.org/Storage_Service.

[2] Archivematica 1.10 - Preservation Planning, https://www.archivematica.org/en/docs/archivematica-1.10/user-manual/preservation/preservation-planning/#preservation-planning.

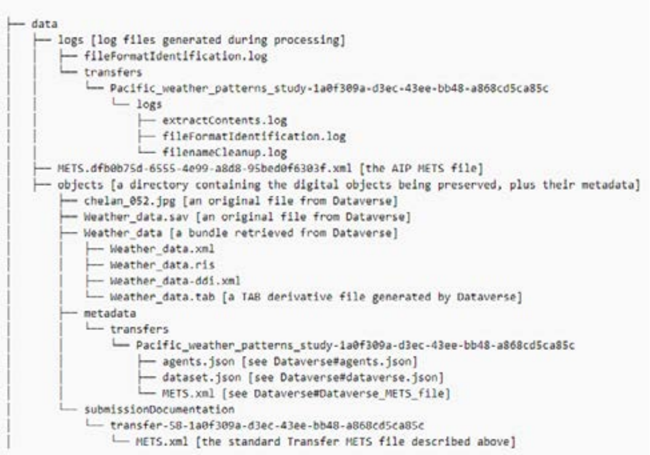
[3] This information is entered as part of the storage service setup in Archivematica and is also stored as file called agents.json in the 'metadata' folder of the AIP: https://www.archivematica.org/en/docs/storage-service-0.15/administrators/#dataverse.

[4] As Evelyn McLellan writes, "The use case is a research data publishing platform that generates tabular file format derivatives from uploaded statistical files. This is not normalization because the purpose is not preservation but rather derivation for the purpose of data manipulation and visualization." See: http://premisimplementers.pbworks.com/w/page/102413902/Preservation%20Events%20Controlled%20Vocabulary. This is as opposed to the "derivation" relationship type referred to in PREMIS s. 1.13.1.

are included, and any originally zipped folders will result in a separate directory within the AIP.

Table 1. Dataverse-Archivematica METS structure overview

METS section	Description
METS dmdSec	Descriptive metadata section in DDI
ddi:title ddi:IDno ddi:authEnty ddi:distrbtr ddi:version ddi:restrctn	DDI fields include: title, unique identifier (e.g., DOI), author(s), distributor (i.e., the Dataverse instance), dataset version, and licenses/restrictions Additional descriptive metadata from Dataverse is stored in the AIP in a file titled "dataset.json" and is referenced using an xlink in the dmdSec of the Archivematica METS file. Any DDI XML files for tabular data files are also linked in the dmdSec
METS amdSec	Administrative metadata section (for original, derivative, metadata, and normalized files)
techMD	Technical metadata (PREMIS), including file format information and extracted metadata
digiprovMD	Provenance metadata, including PREMIS events for derivation (for tabular derived files), ingestion, unpacking bundled files, virus check, fixity check, normalization, and linked PREMIS agents for associated organizations, software, and Archivematica users
METS fileSec	File section defining original files uploaded to Dataverse, derivative tabular files generated by Dataverse, metadata files generated by Dataverse, submission documentation, metadata files and normalized preservation copies generated during Archivematica processing, if applicable
METS structMap	Structural map, showing directory structure of the contents of the AIP



```
├── data
│   ├── logs [log files generated during processing]
│   │   ├── fileFormatIdentification.log
│   │   └── transfers
│   │       └── Pacific_weather_patterns_study-1a0f309a-d3ec-43ee-bb48-a868cd5ca85c
│   │           └── logs
│   │               ├── extractContents.log
│   │               ├── fileFormatIdentification.log
│   │               └── filenameCleanup.log
│   ├── METS.dfb0b75d-6555-4e99-a8d8-95bed0f6303f.xml [the AIP METS file]
│   ├── objects [a directory containing the digital objects being preserved, plus their metadata]
│   │   ├── chelan_052.jpg [an original file from Dataverse]
│   │   ├── Weather_data.sav [an original file from Dataverse]
│   │   ├── Weather_data [a bundle retrieved from Dataverse]
│   │   │   ├── Weather_data.xml
│   │   │   ├── Weather_data.ris
│   │   │   ├── Weather_data-ddi.xml
│   │   │   └── Weather_data.tab [a TAB derivative file generated by Dataverse]
│   │   ├── metadata
│   │   │   └── transfers
│   │   │       └── Pacific_weather_patterns_study-1a0f309a-d3ec-43ee-bb48-a868cd5ca85c
│   │   │           ├── agents.json [see Dataverse#agents.json]
│   │   │           ├── dataset.json [see Dataverse#dataverse.json]
│   │   │           └── METS.xml [see Dataverse#Dataverse_METS_file]
│   │   └── submissionDocumentation
│   │       └── transfer-58-1a0f309a-d3ec-43ee-bb48-a868cd5ca85c
│   │           └── METS.xml [the standard Transfer METS file described above]
│
```

Fig. 2. Example Dataverse-Archivematica AIP structure [22]

V. Testing and Future Development

During the development work, nine test datasets were created in the Scholars Portal Demonstration Dataverse[1] that were representative of the types of datasets deposited in the Scholars Portal production platform. Three of these included tabular data; one included a large collection of text files and images, including zipped packages; and another consisted of geospatial data files. Three others contained Microsoft Office documents, images, and audiovisual materials, respectively. A final dataset contained only metadata (no data files). Our testing focused on ensuring the successful request and receipt of complete data packages from Dataverse to Archivematica and ensuring that an AIP with an accurate Archivematica-generated METS file was created and placed in storage. Processing and configuration details and issues specific to Archivematica, such as file format normalization, were not considered.

The main issues experienced during testing related to unsuccessful Archivematica transfers from Dataverse-originated packages. For example, transfer failures that were the result of files in Dataverse missing checksums or as a result of failed tabular data ingest processes in Dataverse.2 Testing

[1] Scholars Portal Demo Dataverse: https://demodv.scholarsportal.info/.

[2] Dataverse did not implement checksums in versions 3.6 and prior. For a list of known issues experienced during testing, see https://wiki.archivematica.org/Dataverse.

also revealed a number of issues affecting transfers and processing that were outside of the scope of the contracted development. In the following sections, we identify necessary fixes and enhancements in several areas that could be incorporated in future releases.

A. *Metadata*

Currently, only six DDI fields (title, unique identifier, author(s), distributor, dataset version, and restriction) are included in the METS descriptive metadata section (see Table 1 above). Additional DDI fields (abstract, subject, and distDate) were proposed given that the first two of these fields are mandatory in Dataverse, and the third provides additional context. However, the addition of these fields was considered by Artefactual to be outside the scope of the development contract. Rights-related metadata for individual files could also be mapped directly to PREMIS as is supported currently in other Archivematica workflows. Dataverse packages consisting of only metadata currently fail, based on the rationale that there is nothing to preserve if a package does not contain any data files.

B. *Interface*

Several improvements to the transfer browser pane were identified that would facilitate the ability to query and select appropriate datasets within Dataverse, such as showing the dataset version number and the ability to refine searches within the interface. An indication of whether a dataset has already been processed is another potential improvement. The team also outlined the need for stronger error logging and clearer notifications to users based on the issues experienced during testing noted above. Joel Simpson suggested the idea of an additional micro-service for verifying a Dataverse transfer before the transfer begins to make it easier to identify these errors and ensure compliance.

C. *Conformance with Additional Archivematica Functions*

AIP re-ingest functions present in Archivematica do not currently function for Dataverse AIPs. Development of this feature requires further discussion about use cases and approaches, such as whether re-ingest should take into account any updates made to the submitted dataset in Dataverse. The team also noted the potential benefit of relating datasets as part of a larger collection through defining an Archival Information Collection (AIC),[1] a function that needs further development to conform with the Archivematica workflow for creating AICs.

D. *Messaging to Dataverse and DIPs*

Once a dataset has been processed and stored, it would be beneficial for Archivematica to send a notification to the Dataverse platform and surface selected preservation metadata indicating to users that the dataset has been processed by Archivematica. However, this communication mechanism would require development work on both platforms. As mentioned previously in section III.B above, a larger potential development would be the automated replacement of user-submitted data with Archivematica-created DIPs, particularly when normalized access copies of files submitted by depositors might be desired for ease of access for general users. An example would be if a depositor submitted a large TIFF image: Archivematica's access normalization functions could create a smaller version in JPEG format that would be more suitable for general access purposes in Dataverse.

E. *Conformance with External Requirements*

As methods for standardization continue to develop in the field, an additional development opportunity is the ability for Archivematica-created AIPs and DIPs in Bags to be conformant with the RDA Research Data Repository Interoperability Working Group's *Final Recommendations* document. The *Recommendations* specify how repository outputs should be structured to promote data exchange, which could be used for redundant storage or access purposes [23]. Dataverse's Bag export function adheres to the RDA specification [19].

VI. Discussion and Conclusion

Currently, Scholars Portal is hosting a public Archivematica sandbox connected to its demo Dataverse installation with several test datasets.[2] Invitations to participate in testing the sandbox

[1] AIC: https://wiki.archivematica.org/AIC.

[2] Archivematica Demo Sandbox, *Spotdocs Wiki:* https://spotdocs.scholarsportal.info/display/DAT/Archivematica+Demo+Sandbox.

and to provide feedback were shared with regional, national and international groups related to research data management, digital preservation, and archives, as well as Dataverse users and Archivematica users. Community testing is crucial to provide further information about how different users might use the integration and to identify additional needs from the community. This feedback will be used to inform future platform enhancements and contribute to the ongoing discussion surrounding best practices for preserving research data. We hope that others interested in using these tools will bring additional use cases and sponsor additional developments to improve the integration. Community members who test and implement the integration on their own infrastructure will also provide new perspectives related to its capacity and limitations in different contexts.

This research and integration work contributes to ongoing research and discussions surrounding research data preservation. Several challenges exist in this area, particularly in relation to forming research data preservation policies and strategies. A recent Jisc report *What to Keep* outlined use cases for research data retention and considerations for this emerging field, noting that the practice and procedures—the what, why, how long, and where—are still evolving [24]. Another challenge in developing policies and strategies relates to the heterogeneity of research data, resulting in a large number of data types and file formats, as well as discipline-specific practices and protocols. The *Science Europe Guidance Document: Presenting a Framework for Discipline-specific Research Data Management* provides a useful guidance framework for protocols within various research domains, informed by the FAIR principles, applicable laws, regulations, and standards [25]. The significant differences across disciplines suggest inherent difficulties in developing general policies and strategies for multi-disciplinary data repositories. Increasing our shared knowledge of various curation and preservation workflows would help to ensure that the tools and policies developed in these areas assist in properly managing different types of data for the long term.

Finally, additional research and requirements-gathering needs to be conducted in the area of service models and policy development to understand how preservation approaches can flow from individual researchers to institutions and repositories that are tasked with stewarding research data, and onto potential to shared infrastructures. In addition to connecting the technical pieces of infrastructure, the stewarding institution or organization would need to develop and manage policies and costs for long-term storage and maintenance. For example, OCUL institutions that subscribe to Permafrost would have access to Archivematica instances that could be configured to their institutional containers as part of Scholars Portal Dataverse platform. In this case, datasets processed as AIPs could be stored on the OLRC and managed by the library. Other users may host Archivematica locally or take advantage of other service arrangements and still be able to connect to a target Dataverse instance of their choice. The integration also presents opportunities for centralized, collaborative services that offer Dataverse, Archivematica, and preservation storage as a service model, and therefore a consequent requirement to develop appropriate agreements and governance models for shared services.

Overall, the Dataverse-Archivematica integration project aims to connect several pieces of the research data management ecosystem, drawing on best practices and standards in the archives and digital preservation communities, and to contribute to the development and enhancement of features within these two platforms.

Acknowledgements

The authors are grateful for funding from OCUL to pursue this project and thank the Scholars Portal staff, members of Artefactual Systems, the initial project participants, and the community for their support of the project. Our thanks go to: Allan Bell (UBC), Eugene Barsky (UBC), Peter Binkley (University of Alberta), Eleni Castro (Boston University, formerly IQSS), Alan Darnell (Scholars Portal), Kate Davis (Scholars Portal), Philip Durbin (IQSS), Alex Garnett (SFU), Geoff Harder (University of Alberta), Chuck Humphrey (Retired, formerly CARL Portage), Larry Laliberte (University of Alberta), Amber Leahey (Scholars Portal), Victoria Lubitch (Scholars Portal), Steve Marks (University of Toronto), Evelyn McLellan (Artefactual Systems), Umar Qasim (University of Alberta), Joel Simpson (Artefactual Systems), Justin Simpson (Artefactual Systems), Ross Spencer (Artefactual Systems), Amaz Taufique (Scholars Portal), Leanne Trimble (University of Toronto), and Dawas Zaidi (Scholars Portal). Finally, the authors would like to thank the three anonymous reviewers of the manuscript for their valuable comments.

References

[1] CASRAI, Research data, Aug. 12, 2015. Accessed on: Nov. 1, 2019. [Online]. Available: https://dictionary.casrai.org/Research_data

[2] GO Fair, FAIR Principles. Accessed on: Nov. 1, 2019. [Online]. Available: https://www.go-fair.org/fair-principles/

[3] M.D. Wilkinson et al., "The FAIR Guiding Principles for scientific data management and stewardship," Scientific Data 3, Mar. 15, 2016. doi:10.1038/sdata.2016.18

[4] Digital Curation Centre, DCC curation lifecycle model. Accessed on: Nov. 1, 2019. [Online]. Available: http://www.dcc.ac.uk/resources/curation-lifecycle-model

[5] M. Crosas, "The Dataverse Network: An open-source application for sharing, discovering and preserving data," D-Lib Magazine, vol. 17, no. 1-2, January-February 2011. doi:10.1045/january2011-crosas

[6] Dataverse, Dataverse repositories: A world view. Accessed on: Nov. 1, 2019. [Online]. Available: https://services.dataverse.harvard.edu/miniverse/map/

[7] T. H. Vines et al., "The availability of research data declines rapidly with article age," Current Biology vol. 24, no. 1. doi:10.1016/j.cub.2013.11.014

[8] Dataverse, Dataverse installation guide: Configuration – Database settings - :FileFixityChecksumAlgorithm, Oct. 23, 2019. Accessed on: Nov. 1, 2019. [Online]. Available: http://guides.dataverse.org/en/latest/installation/config.html?highlight=md5#filefixitychecksumalgorithm

[9] Dataverse, Dataverse installation guide: Installation, Oct. 23, 2019. Accessed on: Nov. 1, 2019. [Online]. Available: http://guides.dataverse.org/en/latest/installation/installation-main.html

[10] Dataverse, Dataverse user guide: Tabular data ingest, Oct. 23, 2019. Accessed on: Nov. 1, 2019 [Online]. Available: http://guides.dataverse.org/en/latest/user/tabulardataingest/index.html

[11] Dataverse, Dataverse user guide: Tabular data ingest - SPSS, Oct. 23, 2019. Accessed on: Nov. 1, 2019. [Online]. Available: http://guides.dataverse.org/en/latest/user/tabulardataingest/spss.html

[12] Dataverse, Dataverse developer guide: Universal Numerical Fingerprint (UNF), Oct. 23, 2019. Accessed on: Nov. 1, 2019. [Online]. Available: http://guides.dataverse.org/en/latest/developers/unf/index.html

[13] B. Lavoie, The Open Archival Information System (OAIS) reference model: Introductory guide (2nd edition). Great Britain: Digital Preservation Coalition, 2014. doi: 10.7207/twr14-02

[14] U. Qasim, C. Davis, A. Garnett, S. Marks, and M. Moosberger, Research data preservation in Canada: A white paper. Portage Network, 2018. doi:10.14288/1.0371946

[15] J. Mitcham, C. Awre, J. Allinson, R. Green, and S. Wilson, Filling the digital preservation gap: A Jisc research data spring project. Phase Three report. Jisc, 2016. doi: 10.6084/m9.figshare.4040787

[16] V. Tykhonov, P. Doorn, and M. Wittenberg, The development of DataverseNL data repository as structured data hub, 2017. Accessed on: Nov. 1, 2019. [Online]. Available: https://www.slideshare.net/vty/dataversenl-as-structured-data-hub

[17] James Myers, Data and metadata packaging for archiving, Jan. 10, 2019. Accessed on: Nov. 1, 2019. [Online]. Available: https://github.com/QualitativeDataRepository/dataverse/wiki/Data-and-Metadata-Packaging-for-Archiving

[18] J. Crabtree, R. Moore, D. Sizemore, Odum Institute iRODS policies to support preservation, 2016. Accessed on: Nov. 1, 2019. [Online]. Available: https://slideplayer.com/slide/14487733/

[19] Dataverse, Dataverse installation guide: Configuration - DuraCloud/Chronopolis integration, Oct. 23, 2019. Accessed on: Nov. 1, 2019. [Online]. Available: http://guides.dataverse.org/en/4.17/installation/config.html#bagit-export

[20] Core Trustworthy data repositories requirements, Nov. 2011. Accessed on: Nov. 1, 2019. [Online]. Available: https://www.coretrustseal.org/wp-content/uploads/2017/01/Core_Trustworthy_Data_Repositories_Requirements_01_00.pdf

[21] Dataverse, Dataverse user guide: Dataset + File Management – Dataset versions, Oct. 23, 2019. Accessed on: Nov. 1, 2019. [Online]. Available: http://guides.dataverse.org/en/latest/user/dataset-management.html#dataset-versions

[22] E. McLellan, "Dataverse: AIP Structure," Archivematica [wiki], Nov. 17, 2015. Accessed on: Nov. 1, 2019. [Online]. Available: https://wiki.archivematica.org/Dataverse#AIP_structure

[23] RDA Research Data Repository Interoperability Working Group (WG), Final Recommendations. Accessed on: Nov. 1, 2019. [Online]. Available: https://www.rd-alliance.org/system/files/Research%20Data%20Repository%20Interoperability%20WG%20-%20Final%20Recommendations_reviewed_0.pdf

[24] Neil Beagrie, What to keep: A Jisc research data study, Feb. 2019. Accessed on: Nov. 1, 2019. [Online]. Available: https://repository.jisc.ac.uk/7262/1/JR0100_WHAT_RESEARCH_DATA_TO_KEEP_FEB2019_v5_WEB.pdf

[25] Science Europe, Science Europe guidance document: Presenting a framework for discipline-specific research data management, Jan. 2018. Accessed on: Nov. 1, 2019. [Online]. Available: https://www.scienceeurope.org/wp-content/uploads/2018/01/SE_Guidance_Document_RDMPs.pdf