Interoperable Data Archiving and Migration Using the RDRI Working Group Recommendations

Jonathan Crabtree (Odum Institute) , Jim Myers, Maxwell Burnett (NCSA), Luigi Marini (NCSA)

Introduction

Replacing bespoke dataset exchange and archiving mechanisms with a standards-based approach is a key step in the maturation of research data management.

This project is developing the means to export and reimport richly annotated datasets, using packages conforming to the RDA's Research Data Repository Interoperability (RDRI) Working Group recommendations[4], in two independent data-focused software systems that are in use today across a broad range of disciplines and in countries around the world.

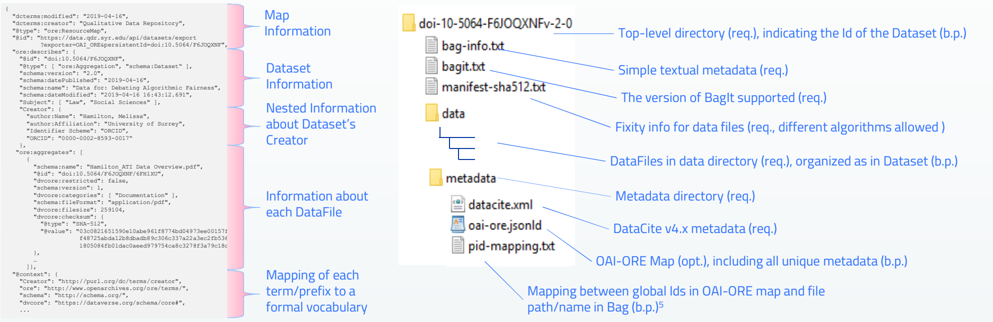
Specifically, we intend to extend Dataverse and Clowder, two open source products with extensive functionality and significantly different internal architectures that we none-the-less believe share enough commonality in their concepts of a dataset to leverage RDRI recommend packaging as a means of exporting, migrating, preserving, and exchanging datasets.

Research Data Alliance Research Data Repository Interoperability (RDRI) Working Group Recommendations

Open Archives Initiative Object Reuse and Exchange (OAI-ORE)[1] – defines a *Map* that *describes* an *Aggregation* that *aggregates* a set of *AggregatedResources*, all of which are described by global identifiers (URIs). Each of these entities can then be annotated with metadata from other vocabularies. An OAI-ORE Map can be represented in RDF, which in turn can be serialized as more readable JSON-LD file[2]:

BagIt[3] – defines a hierarchical structure for storing data and metadata files for preservation and a few mandatory files providing basic descriptive metadata and fixity information.

RDRI WG – recommends [4] using the BagIt and DataCite (v4+) metadata standards and includes OAI-ORE Maps as choice for metadata. It further defines a standard location and naming scheme for metadata files. Together with best practices (b.p.) developed by projects participating in RDA, these provide useful standardization beyond the core BagIt standard.



```
{
  "dcterms:modified": "2019-04-16",
  "dcterms:creator": "Qualitative Data Repository",
  "@type": "ore:ResourceMap",
  "@id": "https://data.qdr.syr.edu/api/datasets/export
          ?exporter=OAI_ORE&persistentId=doi:10.5064/F6JOQXNF",
  "ore:describes": {
    "@id": "doi:10.5064/F6JOQXNF",
    "@type": [ "ore:Aggregation", "schema:Dataset" ],
    "schema:version": "2.0",
    "schema:datePublished": "2019-04-16",
    "schema:name": "Data for: Debating Algorithmic Fairness",
    "schema:dateModified": "2019-04-16 16:43:12.691",
    "Subject": [ "Law", "Social Sciences" ],
    "Creator": {
      "author:Name": "Hamilton, Melissa",
      "author:Affiliation": "University of Surrey",
      "Identifier Scheme": "ORCID",
      "ORCID": "0000-0002-8593-0017"
    },
   "ore:aggregates": [
      {
        "schema:name": "Hamilton_ATI Data Overview.pdf",
        "@id": "doi:10.5064/F6JOQXNF/6FH1XU",
        "dvcore:restricted": false,
        "schema:version": 1,
        "dvcore:categories": [ "Documentation" ],
        "schema:fileFormat": "application/pdf",
        "dvcore:filesize": 259104,
        "dvcore:checksum": {
          "@type": "SHA-512",
          "@value": "03c0821651590e10abe961f8774bd04973ee00157f
                     f48725abda12b8dbadb89c306c337a22a3ec2fb536
                     1805084fb01dac0aeed979754ca8c3278f3a79c18c
        },
        …
      ]},
  "@context": {
    "Creator": "http://purl.org/dc/terms/creator",
    "ore": "http://www.openarchives.org/ore/terms/",
    "schema": "http://schema.org/",
    "dvcore": "https://dataverse.org/schema/core#",
    ...
```

Map Information

Dataset Information

Nested Information about Dataset's Creator

Information about each DataFile

Mapping of each term/prefix to a formal vocabulary

```
doi-10-5064-F6JOQXNFv-2-0
    bag-info.txt
    bagit.txt
    manifest-sha512.txt
    data
    metadata
        datacite.xml
        oai-ore.jsonld
        pid-mapping.txt
```

Top-level directory (req.), indicating the Id of the Dataset (b.p.)

Simple textual metadata (req.)

The version of BagIt supported (req.)

Fixity info for data files (req., different algorithms allowed)

DataFiles in data directory (req.), organized as in Dataset (b.p.)

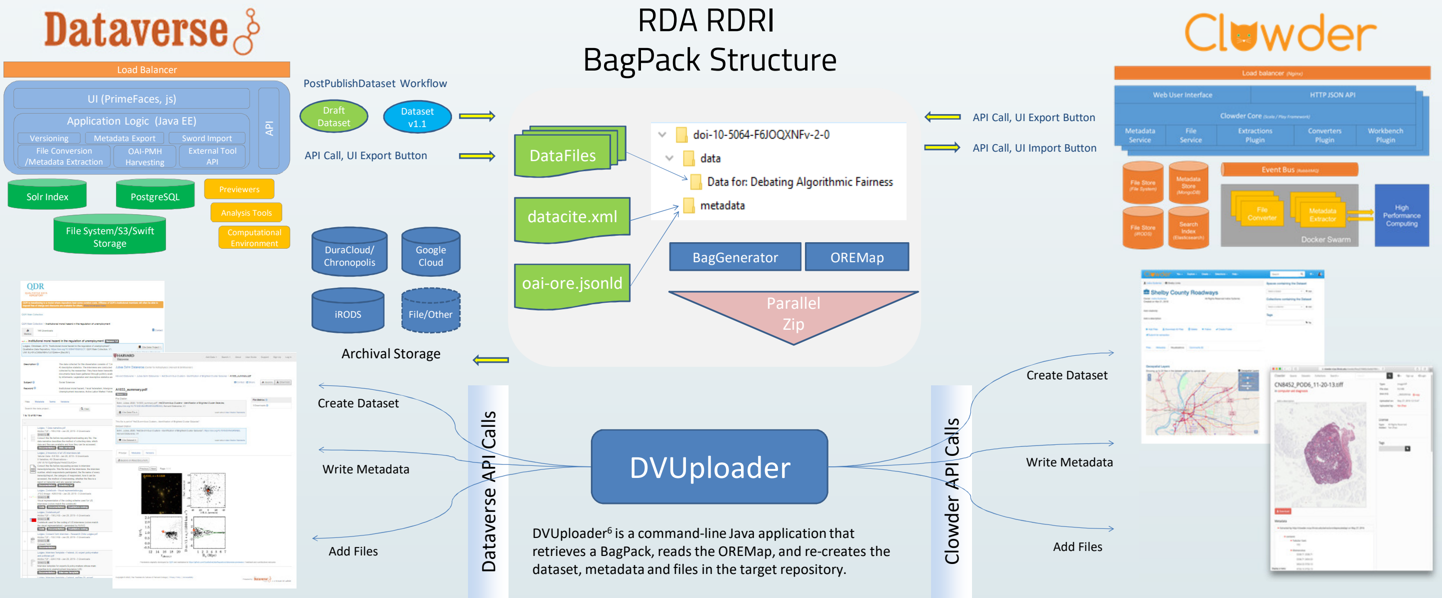
Metadata directory (req.)

DataCite v4.x metadata (req.)

OAI-ORE Map (opt.), including all unique metadata (b.p.)

Mapping between global Ids in OAI-ORE map and file path/name in Bag (b.p.)[5]

Conceptual Overview



DVUploader[6] is a command-line Java application that retrieves a BagPack, reads the OREMap, and re-creates the dataset, metadata and files in the target repository.

Interoperability: Opportunities and Challenges

We anticipate numerous benefits from this work including shared code that will reduce per-repository development, archiving capabilities that support administrators in obtaining certification such as the Core Trust Seal, and flexibility for users to migrate data and to leverage the unique capabilities of different repositories independent of any requirements with respect to publication and long-term storage. We also recognize that Dataverse and Clowder differ in multiple ways that will present interoperability challenges that will ultimately need to be addressed by RDA and the repository community:

- **Storage Architecture:** document versus SQL databases
- **Support for Versioning:** Dataverse assigns a single DOI across versions, Clowder does not directly support versioning
- **File Identifiers:** Dataverse can optionally assign DOIs at the file level, Clowder assigns UUIDs
- **Vocabulary/profile alignment:** Clowder and Dataverse do not have 100% overlap in 'core' metadata and both allow customization to support new terms
- **Value assumptions:** Both repositories make assumptions on the structure of at least some metadata values for display and/or internal or third-party tool functionality (e.g. author attributes, metadata used in computations)
- **Controlled vocabularies:** Clowder can leverage external web vocabularies whereas Dataverse currently relies on locally managed metadata blocks
- **Derived Metadata:** Both systems store metadata generated by third-party tools that could conflict with that normally generated locally from file contents
- **Ancillary files:** Both systems create (different) files that can be considered metadata: thumbnails, previews, converted formats, provenance, structured (XML) metadata, etc.
- **Access Control/Data Sensitivity:** Dataverse supports per-file access control, terms of use that must be accepted to access files and is exploring sensitivity-related Data Tags that have associated storage and access restrictions.
- **Scalability:** Clowder supports datasets with millions of files, deep folder hierarchies, and tera- to peta-scale data volumes.

Work Plan

This project leverages prior work to develop OAI-ORE and BagIt export/import functionality for Clowder in the NSF SEAD DataNet project and work to update and adapt that functionality to provide archival export from Dataverse supported through the Qualitative Data Repository (https://qdr.syr.edu). The RDA Adoption project now supporting work within the Global Dataverse Community Consortium (GDCC) and at the National Center for Supercomputing Applications (NCSA) is intended to deliver the following outputs.

Current Status	Future Options
Development	Round-trip export/import of BagPacks in Clowder and Dataverse. Exports will include, at a minimum, all metadata and data files required to regenerate the original dataset version in the originating repository and may include derived metadata/ancillary files that represent value added by the originating repository. Dataverse will initially support import via the stand-alone DVUploader application whereas Clowder will implement import initiated via Clowder's web interface.
Testing	We intend to demonstrate proof-of-concept interoperability between Dataverse and Clowder and will test with datasets of varying complexity between instances of the same repository (e.g. Dataverses with different installed metadata blocks) and between repositories. RDRI-conformant BagPacks from other repositories may also be tested.
Implementation/ Deployment	Dataverse and Clowder both intend to include the developed functionality for community use in future releases. GDCC further anticipates working with Dataverse community members including Harvard University, the Odum Institute, the Qualitative Data Repository (QDR), and the Texas Digital Library (TDL) to implement dataset archiving leveraging RDRI-conformant BagPacks in their production systems.
Outreach/ Documentation	We anticipate a combination of community presentations, online documentation, and/or conference posters/papers describing this work, providing information to users and administrators, and describing the benefits and limitations discovered. We also anticipate providing a suite of test BagPacks that can be leveraged by future implementations.

Acknowledgements & References

This work is funded as an RDA Adoption project through support received from the U.S. National Science Foundation (NSF) for the Research Data Alliance United States (RDA/US). The authors thank the Dataverse and Clowder Teams and Communities for helpful discussions and support in integrating this work into these applications and for their interest in adopting it in their operations.

[1] Open Archives Initiative Object Reuse and Exchange, https://www.openarchives.org/ore/
[2] ORE User Guide - Resource Map Implementation in JSON-LD, http://www.openarchives.org/ore/0.9/jsonld
[3] The BagIt File Packaging Format (V1.0), https://tools.ietf.org/html/draft-kunze-bagit-17
[4] Research Data Repository Interoperability WG Final Recommendations, https://www.rd-alliance.org/group/research-data-repository-interoperability-wg/outcomes/research-data-repository-0
[5] Package Serialization Using BagIt, https://releases.dataone.org/online/api-documentation-v2.0.1/design/DataPackage.html
[6] DVUploader, a Command-line Bulk Uploader for Dataverse, https://github.com/IQSS/dataverse-uploader