Research Data Repository Interoperability Primer

The Research Data Repository Interoperability Working Group will establish standards for interoperability between different research data repository platforms focusing on machine-machine communication with the primary goal of enabling migration, replication and cross-repository discovery. These standards may include (but are not limited to) a generic API specification and import/export formats summarized in a document serving as an implementation guide for adoption. The scope of this document is to give an overview about targeted use cases, requirements that must be fulfilled to realize the use cases and currently available technologies and standards that might be incorporated. In the following section, the use cases this working group is focussing on including their requirements are described.

Use Cases

1. Migration/Replication of a Digital Object between research data repository platforms

    - Transfer of digital objects from a source to a destination research data repository platform
    - Three different cases are imaginable:
        - Source and destination are the same repository platform with the same or a fully compatible version (same data model)
        - Source and destination are the same repository platform with a different version (different data model, same/similar ontologies)
        - Source and destination are different repository platforms (different data model, different ontologies)
    - Support for legacy platform versions
        - Adoption of results in old platform versions unlikely, but interoperability interfaces are quite important for these platforms
        - Integration of legacy platforms by service/tool approach rather than defining an interface that has to be implemented by a platform itself
        - Necessity of common exchange format used by service/tool between source and destination?
    - Web-based or standalone solution?

2. Retrieval of information related to the platform and/or its contents

- Information about system, e.g. supported interfaces, capacities, access requirements
    - Facilitate registration in a (collection) registry and metadata harvesting
- Information about content
    - Facilitate for metadata harvesting
- Minimal set of mandatory elements but extensible for future use

State of the Art

A. Machine Operable Interfaces

In this section a selection of machine operable interfaces is presented. This includes generic protocol specifications for data/metadata exchange as well as specific programming interfaces implemented in one particular platform that can be adopted by other platforms.

1. OAI-PMH

Type: Protocol
Applicability: Use case 2
Feasibility for research data: No concerns

The OAI-PMH protocol is a standard protocol for metadata harvesting defined by the Open Archives Initiative. The main purpose of this protocol was to harvest publication information from e-print servers and digital repositories. The protocol is intended [...]to provide a low-barrier mechanism for repository interoperability[...] [1] and has been adopted by many (if not all) major repository platforms. The OAI-PMH standard defines an easy-to-implement and easy-to-use HTTP-based interface for requesting metadata records from digital repositories. It delivers metadata records as XML documents containing metadata in arbitrary schemas. Supported schemas can be requested using the OAI-PMH endpoint. Furthermore, metadata records can be organized in collections. For scalability purposes, harvesters can define a from/until range of the modification timestamp of requested records in order to limit the number of harvested metadata documents. In order to limit the amount of returned records, the OAI-PMH provider can also provide a resumptionToken in order to deliver lists of records in multiple requests.

Gaps and Limitations

The standard expects only Dublin Core metadata to be supported. OAI-PMH only supports metadata.

2. SWORD

Type: Protocol
Applicability: Use case 1

Feasibility for research data: A refresh to SWORD is planned by Jisc (UK) in the next few months with the aim to make the standard completely suitable for research data.

SWORD is a lightweight protocol for depositing content from one location to another. It stands for Simple Web-service Offering Repository Deposit and is a profile of the Atom Publishing Protocol (known as APP or ATOMPUB). To make use of SWORD, a repository or deposit client must either currently support sword or implement the SWORD Profile.

Gaps and Limitations

A log of requirements is being kept at in this GoogleDoc. Any gaps that are identified can also be added here to form a development template for the protocol refresh. Currently the protocol supports research data but not metadata.

3. ResourceSync

Type: Specification (latest version 1.1 published February 2017)
Applicability: Use case 1 and 2
Feasibility for research data: Feasible for research data and metadata.

ResourceSync is a framework for resources in source and destination systems to remain synchronized. Several synchronization patterns are supported, included one to many, many to one, and selective synchronization of a subset of resources. A baseline synchronization is equivalent to a migration in the sense that the targeted set of resources in the source system are copied to the destination system. This is done primarily via a Resource List, which is based on Sitemaps. ResourceSync also supports Resource Dumps, which are documents that point to ZIP files containing resource representations and metadata.

Gaps and Limitations

Embedded metadata is limited, provided as attributes in an XML tag. Complex metadata must be provided externally and referenceable as an IRI in the XML. Resources must have HTTP URIs and synchronization occurs over HTTP.

4. Linked Data Platform (LDP)

Type: Specification
Applicability: Use case 1
Feasibility for research data: No concerns

The Linked Data Platform (LDP) defines a set of rules for HTTP operations on web resources, some based on RDF, to provide an architecture for read-write Linked Data on the web. This standard can help promote interoperability by establishing clear rules for interactions between clients and servers.

Gaps and Limitations

LDP focuses on client/server interactions on the linked data web.

5. DARIAH Storage API

Type: API
Applicability: Use case 1
Feasibility for research data: no concerns for data, not feasible for metadata

The DARIAH RESTful Storage API is based on the REST architectural style; resources are addressed by URLs, and the actions to perform on a representation will be performed by using some of the basic HTTP methods to achieve basic CRUD operations on the resources (create, retrieve, update, and delete). The DARIAH Storage API is used by the DARIAH-DE Repository and the DARIAH-DE Geo-Browser/Datasheet Editor.

Gaps and Limitations

The DARIAH Storage API is only for transparent data storage/access; metadata handling is not part of the API.

6. DARIAH-DE Repository API

Type: API
Applicability: Use case 1 and 2
Feasibility for research data: no concerns

The DARIAH-DE Repository API is used by the DARIAH-DE Repository as a layer above the Storage API for OwnStorage and PublicStorage management and access. Metadata can be retrieved/exchanged as well.

Gaps and Limitations

This is a single solution instead of a generic approach, it would only apply to adopters of the DARIAH-DE Repository.

7. OpenAIRE API

Type: API
Applicability: Use case 1 and 2
Feasibility for research data: No concerns for metadata, not feasible for data

The OpenAIRE API is focused on digital repositories system to system interoperability for open access to scientific publications, data and projects outcomes and reports from EC, FP7, ERC, Horizon 2020, H2020. Software implementation is a part of an official release of DSpace (open source digital repository). OpenAIRE Guidelines for Data Archives being harvested by OpenAire are available online, and a Validator service allows one to test a repository's compatibility with the OpenAIRE Guidelines. If validation succeeds the data source can be registered for regular aggregation and indexing in OpenAIRE. OpenAIRE allows for registration of institutional and thematic repositories registered in OpenDOAR,

research data repositories registered in re3data, individual e-Journals, CRIS, aggregators and publishers. OpenAIRE compliant repositories include: DSpace, Eprints, Invenio CDS, Zenodo.

Gaps and Limitations

The OpenAIRE API is based on OAI-PMH, DataCite v3.1, previously used Dublin Core Metadata (with modifications and extensions). It only supports metadata.

8. VOSpace, the IVOA interface to distributed storage

Type: Protocol
Applicability: Use case 1, Use case 2
Feasibility for research data: Could be reused for other domain than astronomy

VOSpace is the IVOA interface to distributed storage. It specifies how VO agents and applications can use network attached data stores to persist and exchange data in a standard way.

A VOSpace web service is an access point for a distributed storage network. Through this access point, a client can:

- add or delete data objects in a tree data structure
- manipulate metadata for the data objects
- obtain URIs through which the content of the data objects can be accessed

VOSpace does not define how the data is stored or transferred, only the control messages to gain access. Thus, the VOSpace interface can readily be added to an existing storage system. When we speak of "a VOSpace", we mean the arrangement of data accessible through one particular VOSpace service. Each data object within a VOSpace service is represented as a node and has a description called a representation. Nodes in VOSpace have unique identifiers expressed as URIs in the `vos' scheme.

Gaps and Limitations

No default storage is associated. VOSpace must be implemented over an existing data storage (simple filesystem, database, etc.) Main focus is on data, the only metadata covered a simple string-based properties.

B. (Meta-)Data Formats and Models

This section contains (meta-)data formats and schemas that can be used in order to exchange information in a standardized way. However, in addition to the format there will still be the need of some machine operable endpoint reading and writing the presented standard (meta-)data formats.

1. OAI-ORE

Type: Format

Applicability: Use case 1 and 2
Feasibility for research data: As research data objects are often composed of separably accessible components, OAI-ORE is a good match to research data.

ORE is a standard for describing aggregations of data and the semantic relationships between its components. It leverages RDF to describe the relationships and supports multiple serialization formats (including XML, Atom, and JSON-LD). For some communities--most notably, DataONE--an ORE file is an important component of their BagIt profile, allowing them to describe the roles of the different data files in a bag which are not otherwise described by the standard BagIt metadata and manifest.

Gaps and Limitations

Describing hierarchical collections with ORE can be challenging; however, some patterns are emerging for doing this. One example can be found here.

2. METS

Type: Schema
Applicability: Mainly use case 1, but also use case 2
Feasibility for research data: no concerns

[...]The Metadata Encoding & Transmission Standard (METS) schema is a standard for encoding descriptive, administrative, and structural metadata regarding objects[...][2] It was defined for libraries, but it is not restricted to them. Apart from the actual standard METS provide the possibility to define profiles, which are human readable documents supporting authors and programmers while creating and processing METS documents following a particular profile. Information about mandatory elements and supported schemas for different sections can be provided in the profile. Due to a very flexible schema METS documents may hold detailed information about all aspects of a digital object, e.g. descriptive metadata, administrative metadata and structural information together with URLs referring to data streams.

Gaps and Limitations

Due to its flexibility it is up to a profile to define mandatory elements and requirements towards an accepted METS document. The only aspect that is defined by the standard is the overall structure of the document. In order to support METS, e.g. to expose a METS representation of a digital object via OAI-PMH, this group would have to find a consensus on a METS profile defining the minimum requirements that have to be fulfilled in order to be able import a digital object from a METS document.

3. Re3data Schema

Type: Schema
Applicability: Use case 2

Feasibility for research data: No concerns, schema is dedicated to research data repositories

The re3data.org metadata schema contains metadata properties describing a research data repository such as its general scope, content and infrastructure as well as its compliance with technical, quality and metadata standards. [3] It represents a standard to describe a research data repository and ensures interoperability between research data repositories and re3data registry. The documentation of the schema as well as examples and a running registry with plenty of registered research data repositories is available via re3data.org. For the goals of this working group the pure re3data schema can be a (very first) start into interoperability of repository information (use case 2).

Gaps and Limitations

The focus is on the repository, not on the content. Information about content must be retrieved using one of the repository APIs listed in the re3data document. Supporting the re3data schema won't automatically allow the repository to register an instance at re3data.org. The registration process is triggered manually by filling out a web form.

4. DataCite Metadata Schema 4.0

Type: Schema
Applicability: Use case 1 and 2
Feasibility for research data: No concerns

The DataCite Metadata Schema is a list of core metadata properties chosen for an accurate and consistent identification of a resource for citation and retrieval purposes, along with recommended use instructions. It provides a set of mandatory, recommended and optional elements, widely used by data providers to register DOI-Names. The resource that is being identified can be of any kind, but it is typically a dataset.
The DataCite REST API is a common API to get all metadata from DataCite. The API is generally RESTFUL and returns results in JSON. The API follows the JSONAPI specification.

Gaps and Limitations

DataCite is only one standard out of many. An (incomplete) list of them can be found in the RDA Metadata Directory.

5. Portland Common Data Model

Type: Model
Applicability: Use case 1
Feasibility for research data: No concerns

The Portland Common Data Model (PCDM) is a flexible, extensible domain model that is intended to underlie a wide array of repository and DAMS applications. The primary

objective of this model is to establish a framework that developers of tools can use for working with models in a general way, allowing adopters to easily use custom models with any tool. Given this interoperability goal, the initial work has been focused on structural metadata and access control, since these are the key actionable metadata.

Gaps and Limitations

The Portland Common Data Model is only one example of a well-accepted domain model. There are many more, e.g. NetCDF (Network Common Data Format), HDF5 (Hierarchical Data Format), which may have a different perspective.

C. Custom Tools

In this final section custom tools and other contributions not fitting in sections A and B are located.

1. Import/Export via BagIt

Applicability: Use case 1
Feasibility for research data: no concerns

The Fedora Import/Export utility is a tool that allows repository resources to be exported from Fedora as serialized RDF, optionally packaged using the BagIt standard. These exported resources can then be imported into a different Fedora repository or an external preservation system. While this functionality is currently limited to the previously mentioned use cases, it could be expanded to support import/export from/to other repository platforms as well. The primary advantage to this approach over a generic API specification is ease of implementation - the external tool would only need to be made to work with an existing repository's API rather than modifying the core repository code (which may be difficult or impossible for existing implementations with little to no technical support for upgrades and modifications). Apart from the explicit Fedora use case, there are also solutions used by DataONE and NIST that are based on BagIt accompanied by an ORE manifest for data package exchange. More details can be found at researchobject.org and the Research Object BagIt Archive.

Gaps and Limitations

The utility is currently focused on importing/exporting to/from Fedora, but could be expanded to support other repository platforms.

2. Dat Protocol

Applicability: Use case 1
Feasibility for research data: The focus is on syncing and versioning datasets between systems

Dat is a protocol designed for syncing folders of data, even if they are large or changing constantly. Dat uses a cryptographically secure register of changes to prove

that the requested data version is distributed. A byte range of any file's version can be efficiently streamed from a Dat repository over a network connection. Consumers can choose to fully or partially replicate the contents of a remote Dat repository, and can also subscribe to live changes. To ensure writer and reader privacy, Dat uses public key cryptography to encrypt network traffic. A group of Dat clients can connect to each other to form a public or private decentralized network to exchange data between each other. A reference implementation is provided in JavaScript.

Gaps and Limitations

Dat is un-opinionated about metadata, only provides unstructured file synchronization, and seems to be experimental at the moment.

3. The Data Documentation Initiative (DDI)

Applicability: Use case 1
Feasibility for research data: No concerns

The Data Documentation Initiative (DDI) is an international standard for describing the data produced by surveys and other observational methods in the social, behavioral, economic, and health sciences. DDI is a free standard that can document and manage different stages in the research data lifecycle, such as conceptualization, collection, processing, distribution, discovery, and archiving. Documenting data with DDI facilitates understanding, interpretation, and use -- by people, software systems, and computer networks.

In constructing DDI special care was taken to review related standards as well as previous versions of DDI in order to provide clear mapping to the contents of outside standards or to incorporate content where appropriate. Over 25 standards were evaluated. DDI 3 currently has mapped relationships to the following standards: DDI Codebook, Dublin Core and MARC, GSIM (General Statistical Information Model), ISO/IEC 11179, ISO 19118 - Geography, SDMX, METS and PREMIS.

DDI is a very flexible and complex standard that may be used in "customized" ways that best answers specific needs. DDI profiles allow different agents or agencies to specify exactly how they use the DDI XML format, and thus help achieve seamless transfer and interoperability of DDI instances. A DDI profile describes the subset of valid DDI objects used by an agency for a specified purpose. This is documented in a DDI-XML format, which allows a set of declarations to be made, identifying specific fields in the DDI which are "Used" or "Not Used". Various other qualifications can be made to restrict or default permitted values for specific elements, and human-readable documentation can be added.

Gaps and Limitations

Various versions of DDI Lifecycle and Codebook are in use.

RDA Groups with Overlapping / Complementary Work

Please follow this link to see related RDA groups.

Platform Capability Matrix

Please follow this link to see the platform capability matrix.