Adoption of RDA Interoperability Recommendations

As part of a cooperative effort with the University of North Carolina, NCSA is improving Clowder's interoperability with other research data repositories by adopting the recommendations from the Research Data Alliance as per the findings of their working group detailed here: https://www.rd-alliance.org/system/files/Research%20Data%20Repository%20Interoperability%20WG%20-%20Final%20Recommendations_reviewed_0.pdf

The document primarily describes a particular schema for BagPacks generated from Clowder and the metadata they contain. While Clowder already has some compatibility with BagIt v0.97, it does not meet RDA's recommended standards and the improvements described below will enable greater portability of datasets out of and, significantly, into Clowder instances. The ability to import a BagPack and generate associated Clowder artifacts is a critical component for this that is currently missing.

A. Improving Clowder's exported artifacts (.zip BagPacks)

Section 3 of RDA's recommendations detail the BagPacks Clowder must generate. Key components:

- BagIt v0.97 conformance (implemented)
- One or more metadata XML files in a **/metadata** folder (NEW)
    - **datacite.xml** v4 or newer file is mandatory (detailed below)
    - additional flavors e.g. oai-ore.xml, dc.xml (DublinCore), others are optional
    - **clowder.xml** or other arbitrary files may be defined as well. I think this will be valuable to ensure 1:1 Clowder dataset portability even for Clowder-specific fields that may not be included in datacite.xml or other files. This would be the key file a new Clowder instance would use to recreate all folders, tags, previews, extraction history, etc. for the files in the dataset.
- **fetch.txt** file is supported that contains URLs for every file in the dataset, if we want a metadata-only download option.

The main pieces here are the datacite and (IMO) clowder XML files and deciding if we want to support fetch.txt files.

DataCite v4 has an additional set of standards, awesome!!! https://support.datacite.org/docs/schema-40 Let's take a look.

DataCite XML field	Mandatory?	Clowder mapping notes
identifier	M	DOI, which we generally will not have. This can be discussed, but a default of :none for now is considered valid.
creators	M	dataset creator name (Last, First), with ORCID included if defined in Clowder profile. Need to check if we store Family, Given name as component parts in Clowder, otherwise need to discuss supporting those. If you don't have them filled in, we assume final word of name is Family...?
title	M	dataset name
publisher	M	name of entity that produces the resource, used to formulate citation. Not sure if Clowder is the correct value here, or per-dataset?
publicationYear	M	year of creation
ResourceType	M	Clowder Dataset
description		dataset description
contributors		those who collected, or otherwise contributed to dataset. Clowder dataset contributors list, any file uploaders or metadata uploaders in the dataset or its files?
date		created date YYYY-MM-DD
AlternateIdentifier		Clowder dataset UUID
Format		application/zip
Subjects		Keywords, classifications, key phrases describing resource. Use Tags?
RelatedIdentifier		Clowder base URL, relation isSourceOf
Size		num files
rights		license info, Clowder already has a limited list. Can include URL for license and short name if applicable.
version		
language		
geolocation		
FundingReference		

B. Handling import of BagPacks

The other half of RDA's recommendations are supporting ingest - someone comes along with a compliant BagPack, we upload and use it to generate a new Clowder dataset consuming as much data as we can.

Scenario 1 - It's from Clowder (any instance)

We use clowder.xml to recreate things exactly.

Scenario 2 - It's not from Clowder, but it's compliant.

We recreate as best we can, but need to decide if some datacite.xml fields (such as Version, Language) should be:

- discarded (a full-circle 1:1 preservation is not always possible),
- added as some special metadata,
- added as a new property of the Dataset model.

Scenario 3 - it's not compliant

RDA does specify that we don't need to validate datacite.xml against DataCite, because BagPacks without DOIs would be rejected when they should specifically not be (i.e. for in-progress research data).

But if the BagPack v0.97 requirements are not met, checksums don't match, fetch.txt URLs don't work or a host of other errors, all Clowder artifacts should be erased and an error returned.

This will likely be the most work for this portion - handling resulting errors.