Migration of materials from Zenodo to Invenio RDM

Maxim Potekhin, BNL NPPS

Overview

The goal of this writeup is to lay out the process of migration of materials stored on **zenodo.org** to applications hosted at sites external to CERN (e.g. BNL) which are based on Invenio RDM, and present the results of the **proof-of-concept test** recently conducted. Items to note:

- By "*migration*" we mean harvesting of materials from the data source (**zenodo.org**) as packages which include metadata and the payload, and depositing these into the target repository (**Invenio RDM**), while still leaving the original Zenodo content intact. Notes:
    - It's a good practice with repositories like Zenodo and RDM to follow the central principle of DOI i.e. persistence of data and links to that data – avoidance of broken links; full deletion is impossible by design
    - The assigned DOIs are site-specific i.e. they contain a numerical identifier of the storage site and optionally, a mnemonic reference (which is the case with Zenodo); in that sense, they are not portable
    - An interesting use case is as follows: an item can be uploaded to Invenio RDM with the option "already have a DOI" – and the Zenodo DOI can be specified as such. However, this DOI will resolve to the original repository and the item therein
- There are ways to mitigate side effects of effective duplication of the "legacy" items — see comments in the REST API below
- Both platforms are equipped with the **REST API**. Significantly, this is not an optional interface or an add-on, but the only real interface to either system. An important conclusion: this interface is **complete**. Any function or action is possible with it, either via the Web interface, a Python application or a CLI utility that handles HTTP requests.
- The REST API generates responses according to a well defined JSON schema
- Hence, future migration is possible due to the following

    - Completeness of the REST API
    - Because of the exchange format (JSON), the process of migration is amenable to **automation**
    - Automation can include a Python programmatic interface to form proper HTTP requests, (examples exist) and any other tool or framework for handling HTTP requests.
- The items that have been migrated (duplicated) can be marked as such in their pertinent metadata, which can also contain the DOI of the original submission, in the form of a keyword. This would be a useful and solid policy to adopt.
- The REST APIs of *both* Zenodo service and Invenio RDM:
    - Support download and upload of materials
    - Support definition and modification of metadata and any of its components
    - Support creation and management of Communities (which is not of interest to us at this point)
    - Support auth/auth based on **tokens** that can be obtained from both services prior to data transfer if any sort of auth/auth is needed, such as for data and metadata uploads, or metadata modification
        - – this doesn't apply to read access to public materials which are world-readable
    - Generate responses from their respective servers formatted in JSON, which is trivial to parse in applications and scripts; download links to the payload is enclosed in the JSON-formatted response
    - Use the **same rich schema** for JSON which is necessary for the **completeness** of the API; NB. same JSON schema simplifies the interface between the two systems, including in the context of migration
- Bonus point for Zenodo – organic integration of the (optional) conference information in the Web UI

Zenodo REST API

- Root URL to be used in queries: https://zenodo.org/api/
- An example of a metadata package received when querying an item: https://zenodo.org/api/deposit/depositions/10929194
- NB. The deposit ID in the example above is the meaningful part of its DOI (the suffix) – the ‘zenodo’ string which is a part of the Zenodo DOI is just eye candy that helps brand recognition
- An example of getting the list of records pertaining to a community: https://zenodo.org/api/communities/phenixcollaboration/records

An example of a link to the payload located via a community query:
https://zenodo.org/api/records/11094563/files/ACAT24_PHENIX_DAP_v1.pdf/content

---

Invenio RDM REST API

- Root URL to be used in queries: https://inveniordm.web.cern.ch
- A Python-based example of uploading content to Invenio RDM via the REST API https://github.com/inveniosoftware/docs-invenio-rdm-restapi-example

Obviously, for uploading material some sort of auth/auth becomes mandatory. According to documentation for both Zenodo and RDM,

“The only authentication method supported at the moment for REST API calls is by using Bearer tokens that you can generate at the "Applications" section of your user account's settings of your InvenioRDM instance. There are two ways to pass the tokens in your requests.”

NB. In practice, this seems to work well and the whole scheme is not unique to Invenio RDM in any way.

Outline of the test

The process of migration, as it is defined above, essentially consists of two steps:

- Using the REST API to access the data contained in document collection referred to as “The ePIC Community on Zenodo”, based on a variety of criteria, including a blanket download of all materials (but note extensive search capabilities of Zenodo which makes sophisticated selection possible)
- Using the REST API to upload documents to an Invenio RDM server which in this case was the instance at CERN (however because of the standard API it will work anywhere including BNL)

Both components of this process have been tested, as outlined below.

Downstream component of the test

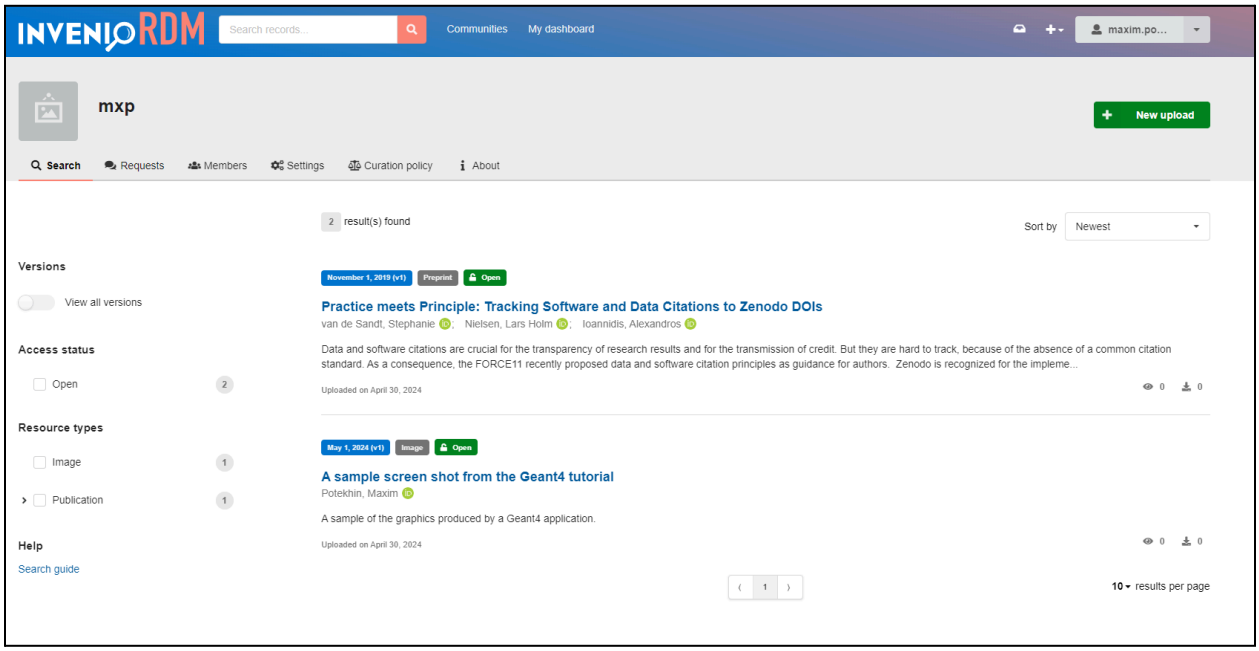
This component of the test contained the following elements

- Use the PHENIX Community on Zenodo as a test case because it comprises a fair amount of data (675 entries)
- Perform a query and receive a JSON-formatted file describing the collection of the PHENIX documents; the size of the file thus received was 2.5MB
- Select a few tests documents from the JSON description and verify the metadata for consistency vs what is seeing on the Zenodo website
- Use the metadata contained therein to pull a few “payloads” of choice, using the REST API, with utilities such as curl and wget
- Check the contents to make sure they are as expected

Upstream component of the test

This component of the test contained the following elements

- Create a test community on Invenio RDM
- Obtain a “Bearer token” from the Invenio RDM service (auth/auth)
- Use the existing example of a Python script which exists on GitHub, based on the “requests” package, to form a correct HTTP request necessary for the upload, perform the upload (running the app from the command line)
- From the Web UI, confirm acceptance into the community (this can also be done programmatically from a script but would require additional coding)
- Check the result (see the screenshot below)



Summary

- The initial test as described above confirmed the feasibility of a future migration from Zenodo to Invenio RDM.
- One element that can be improved is the direct upload to the target community w/o any further action. Additional documentation is needed (a question has been posted on the Invenio forum, information requested). However, at present there is a workable solution for the migration between these two systems.
- Started receiving technical advice from the CERN experts, on the Invenio forum, which is encouraging.