Interoperable REST APIs for large data upload, transfer, and exchange

Leonardo Sala (PSI) and Urs Beyerle (ETHZ)

API Requirements/Features - open for discussion

- Data download
- Data upload (much harder to implement than just download)
- Well defined URIs for download
- User authentication
- Basic checksum test to guarantee successful data transfer
- End to end encryption
- Build in versioning, history (e.g. get latest version of XY)
- Possibility to withdraw data
- …..

Current situation

There is a wide landscape of possibilities to obtain and link open data:

- simple http-based access (ETHZ collections, SciCat, ERIC)
- object storage API (e.g. EnviDat)
- Special solutions like Globus, Rsync, etc.

API access varies:

- specific APIs
  - SciCat: own development
  - ERIC, EnviDat: based on CKAN product
  - ETHZ Research collections: based on DSpace product

Target situation

Ideally only one kind of API should be available to access data, to simplify integration with other tools like Renku and AiiDa. See other initiatives.

One major difference from simple HTTP access would be the introduction of buckets, allowing listing of them, and using them as references in the further integrations + authentication

Possible workflow would require to:

- have an API call to the data repository to get the HTTP pointers to the data (based on e.g Dataset ID) or data bucket
- directly retrieve data from URLs or from buckets list
- have proper authentication in place

Example workpackages

- SciCat: streamline Object storage access
  - at the moment, data can be retrieved in CSCS Object storage system, and an URL is mailed
  - possible example implementations:
    - add in API results the URL links, with expiration time
    - implement S3 buckets in CSCS Object storage for each dataset, where data will be staged and retrieved on request
  - Enable authentication / authorization in object storage
- ETHZ Research Collections
  - Publish DSpace API to request data URLs of public data
  - Explore an access API for Long Term Storage system
- SciCat: generalize data upload
  - at the moment this is based on NFS mounts
  - adding an object storage layer would improve integration with other systems
  - this would also allow data uploads from the browser