PAUL SCHERRER INSTITUT

PSI

Data catalog/Archive team :: Paul Scherrer Institut :: Photon Science Department

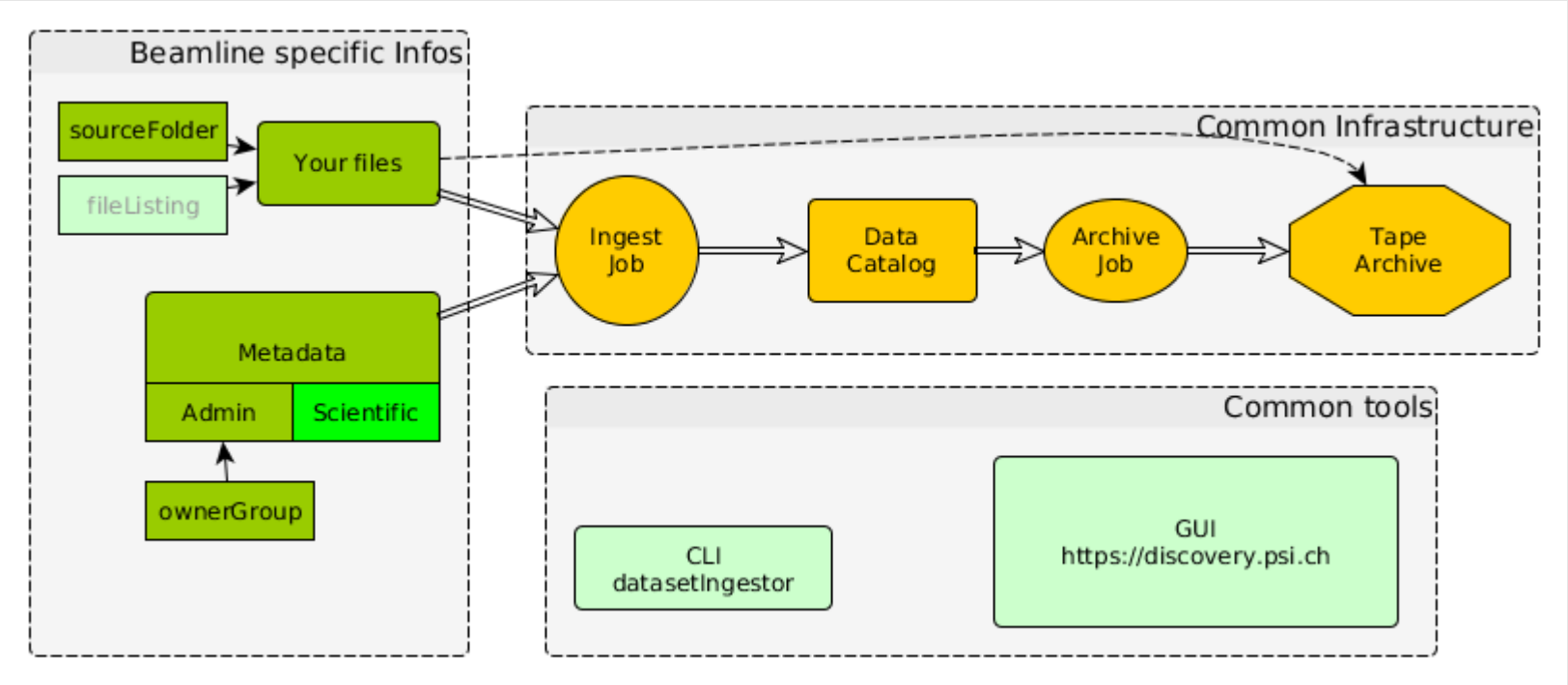
Data Catalog for Scientific Data: How to get started

June 2019

Why bother ?

- Allows you to store your data permanently and make room on your disk space
- Allows to add metadata to your files to give them more meaning
- Allows you to organize data and make the data findable
- Allows you to publish your data and make it citable via a digital object identifier (DOI) – this is work-in-progress
- Can be used both for raw (experiment) data and derived (analysis results) data
- Helps to fulfill Data Management Plan (DMP) requirements from SNF etc.

How it works: adding data



See https://discovery.psi.ch/help/ingestManual for details

You can run the tools either with your personal account or for the beamline managers using a beamline account provided for each beamline

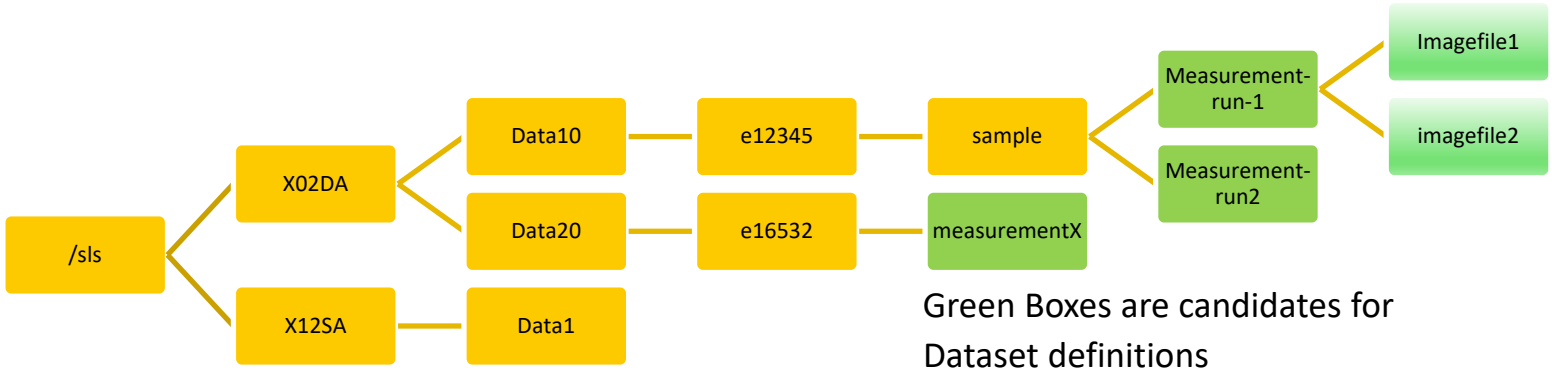
What do you need to do ?

- Often the situation today is like this: you have your data stored on some (central) file server within certain folder locations. You may have additional files “somewhere” annotating this data
- Preparation for Data Catalog is a two step procedure:
  - Step 1: define your **Datasets** (collection of files belonging together)
  - Step 2: Enrich the data with **Metadata**
- Depending on you situation you may add this information manually or automatically.

- See https://discovery.psi.ch/help/ingestManual for all details

1. Define Datasets

- Datasets are the smallest unit for archiving, retrieving and publication
- Create them by defining filelists, e.g. for raw data list all the files that logically belong to a measurement/data taking run. In the simplest case this is just the name of a folder: all files within are taken as the dataset files



- In addition to “raw” Datasets you can create “derived” datasets containing the results of your analysis derived from the raw data. This ingest step is usually done by the user pursuing the analysis

- Warning for raw data : avoid to use the symlinks from the “data” (=eaccount home) folders to the real data locations

Set up metadata.json

- Define data **ownership** by assigning a p-group via Digital user Office DUO.
  - Only members of the pgroup get access to the data. Membership defined by beamline manager BM or principal investigator PI
  - There is a one-to-one mapping from e-accounts to pgroups, e.g. e-account “e12345” has a p-group “p12345”
  - Most p-groups are linked to proposals
- Define minimal administrative metadata (more is possible)
- Put all information into a file “metadata.json” (JSON is a format both easy to read for humans and computer)

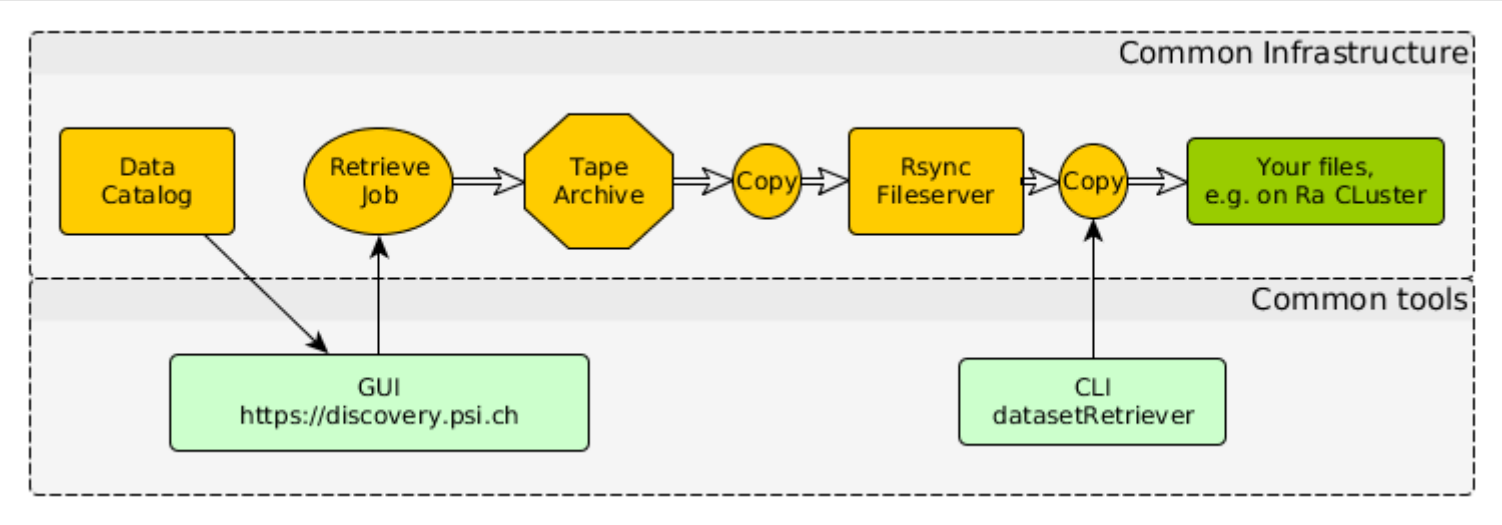
```
{
    "principalInvestigator": "federica.marone@psi.ch",
    "creationLocation": "/PSI/SLS/TOMCAT",
    "sourceFolder": "/data/experiment/run/17",
    "type": "raw",
    "ownerGroup":"p16623“
    ["scientificMetadata “ : optional, see next slide]
}
```

2. Define Scientific Metadata

- The definition of scientific meta data is completely up to the scientific discipline.
- Ideally follow a standard if it exists, e.g. Nexus based HDF5 files
- Scientific metadata can also be added later
- Just an example:

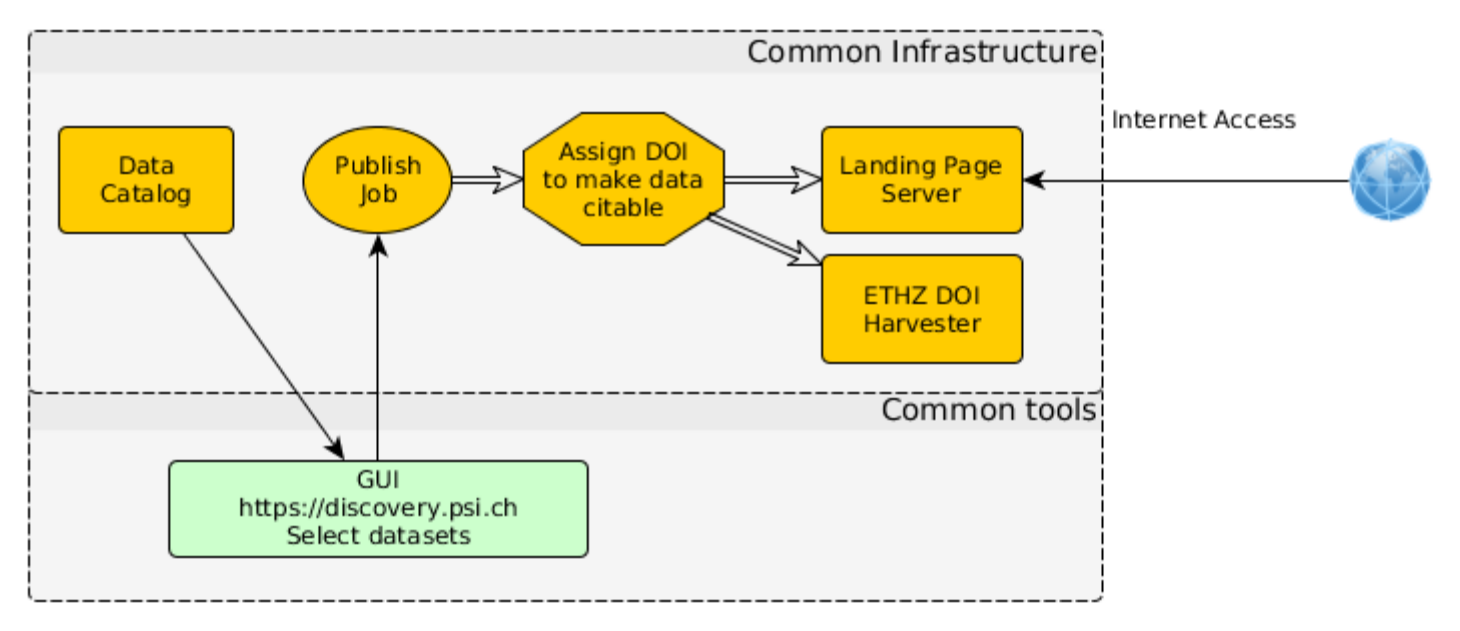
```
“scientificMetadata": {
    "beamlineParameters": {
        "Monostripe": "Ru/C",
        "Ring current": {
            "v": 0.402246,
            "u": "A"
        },
        "Beam energy": {
            "v": 22595,
            "u": "eV"
        }
    },
    "detectorParameters": {
        "Objective": 20,
        "Scintillator": "LAG 20um",
        "Exposure time": {
            "v": 0.4,
            "u": "s"
        }
    }…
}
```

How it works: retrieving data



Retrieve Jobs are typically done for a larger set of files which should be subject to further data analysis. Often they are fetched for a group of people, e.g. to get all data of a previous measurement period.

How it works: publishing data (WIP)



- Process is driven by researchers who want to publish in a journal
- They define a list of datasets as “published data”, for which a DOI is assigned.
- Data will be published on landing page server https://doi.psi.ch

Next steps

- Answer questionnaire (needed for rollout planning)
  - https://docs.google.com/forms/d/e/1FAIpQLSd3p0TvXy4p8_ty3nsPYp4nhtKfVuTq4SmmvKHGwlxVjLS2iw/viewform?vc=0&c=0&w=1&usp=mail_form_link
  - e.g. to define contact persons, suitable dates for first ingest test, automation requirements at beamline etc

- Discuss with your colleagues
  - if you want to ingest historic data as well
    - data migration of data still on disk
    - data migration from old archive will be treated as a dedicated project
  - Discuss which of the newly created data should be ingested to the data catalog and at what time (directly after data taking ? Automatically or manually ?)
  - Get in contact with us (via datacatalog@psi.ch, currently Luke Gorman, Peter Huesser, Edgar Barabas, Stephan Egli)