Helmholtz
Zentrum Berlin

Metadata Ingest Using python-icat

Rolf Krahl

ICAT F2F Meeting, 04 May 2023, Berlin

python-icat: icatdump and icatingest

- `python-icat` provides command line scripts `icatdump` and `icatingest` to dump the ICAT content to a flat file and to restore the content from that file respectively.
- Originally conceived as a debug tool for ICAT deployments. Still heavily used in the `python-icat` test suite.
- Supported file formats: YAML and XML. The ICAT data file format is basically a one to one mapping to the ICAT schema.
- The scripts are based on a backend module `icat.dumpfile`. Using that module, it is easy to create scripts that read or write custom file formats.

ICAT Data File Format

Example Data File

```xml
<?xml version='1.0' encoding='UTF-8'?>
<icatdata>
<data>
  <dataset id="Dataset_1">
    <complete>false</complete>
    <description>Dy01Cp02 at 2.7 K</description>
    <endDate>2022-02-03T17:04:22+01:00</endDate>
    <name>r03517</name>
    <startDate>2022-02-03T15:40:12+01:00</startDate>
    <investigation ref="Investigation_name-gate=3A191=2D00002=2D1=2E1=2DP"/>
    <type name="raw"/>
    <datasetInstruments>
      <instrument pid="doi:10.5442/NI000003"/>
    </datasetInstruments>
    <datasetInstruments>
    <datasetTechniques>
      <technique pid="PaNET:PaNET01196"/>
    </datasetTechniques>
  </dataset>
  <datasetParameter>
    <stringValue>NXxas</stringValue>
    <dataset ref="Dataset_1"/>
    <type name="nxs/entry/definition"/>
  </datasetParameter>
</data>
</icatdata>
```

Restricted Ingest File Format

- Problem: the ICAT Data File Format is too powerful: reading that as an ingest file might potentially create any kind of object in ICAT.
- Solution: define a restricted version of the file format for ingestion that is only capable to list what the ingest need to create.

Ingest File Format

Example Ingest File

```
<?xml version='1.0' encoding='UTF-8'?>
<hzbingest version="1.1">
<data>
  <dataset id="Dataset_1">
    <name>r03517</name>
    <description>Dy01Cp02 at 2.7 K</description>
    <startDate>2022-02-03T15:40:12+01:00</startDate>
    <endDate>2022-02-03T17:04:22+01:00</endDate>
    <datasetInstruments>
      <instrument pid="doi:10.5442/NI000003"/>
    </datasetInstruments>
    <datasetInstruments>
    <datasetTechniques>
      <technique pid="PaNET:PaNET01196"/>
    </datasetTechniques>
  </dataset>
  <datasetParameter>
    <stringValue>NXxas</stringValue>
    <dataset ref="Dataset_1" />
    <type name="nxs/entry/definition" />
  </datasetParameter>
</data>
</hzbingest>
```

Ingest Script

- Implement an ingest script based on the `icat.dumpfile` module.
- The script takes the name of an investigation as parameter and allows only ingesting Datasets (and related DatasetInstruments, DatasetParameter, DatasetTechniques) related to this investigation.
- The script first validates the input using XSD. This makes sure that only allowed elements and attributes are used.
- In a second step, the script uses XSLT to transform the input to ICAT Data File format on the fly, adding all elements that should be hard coded.
- The result is fed into `XMLDumpFileReader` defined by `python-icat`.

References

- Ingestion with python-icat. Presentation at ICAT F2F 2016 in Copenhagen.
- `python-icat` documentation:
  - `icat.dumpfile`
  - `icatingest`