Long Term Archiving Approach @ESRF
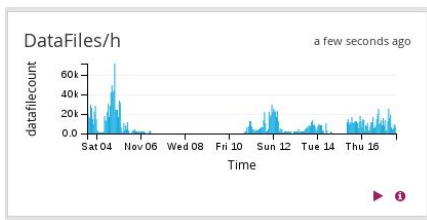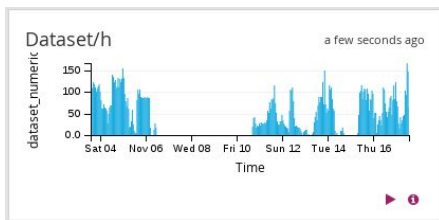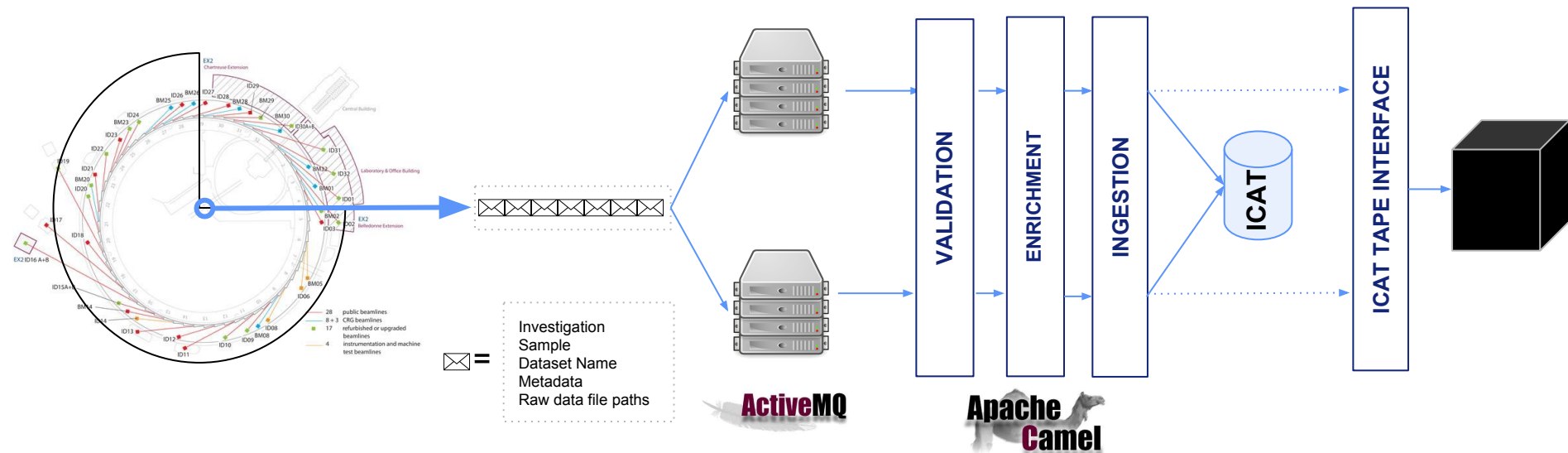
Alex de Maria Antolinos
Software Engineer
Data Manager@Data Analysis Unit
Software Group
ESRF
20.11.1017

Data Policy Short Summary

- Acceptance of the data policy is a **condition** for the **award of beam time**

- ESRF will automatically collect data and metadata for all experiments from all beamlines (including CRGs)

- ESRF is the **custodian** of the **raw data and associated metadata**

- ESRF will store metadata in a metadata catalogue (**ICAT**)

- Experimental team has sole access to the data during the so-called embargo period of 3 years; request to extend the embargo period can be made

- **After the embargo ESRF will make the data "Open Access"**

- **Proprietary data belong by default to the PI and are not archived unless explicitly agreed**

Architecture Overview

Data: from Beamline to Tape

Architecture

Investigation
Sample
Dataset Name
Metadata
Raw data file paths

ActiveMQ

Apache Camel

VALIDATION

ENRICHMENT

INGESTION

ICAT

ICAT TAPE INTERFACE

Dataset/h

a few seconds ago

dataset_numeric

150
100
50
0.0

Sat 04 Nov 06 Wed 08 Fri 10 Sun 12 Tue 14 Thu 16

Time

DataFiles/h

a few seconds ago

datafilecount

60k
40k
20k
0.0

Sat 04 Nov 06 Wed 08 Fri 10 Sun 12 Tue 14 Thu 16

Time

PRODUCERS

CONSUMERS

Archiving

How to archive a dataset?

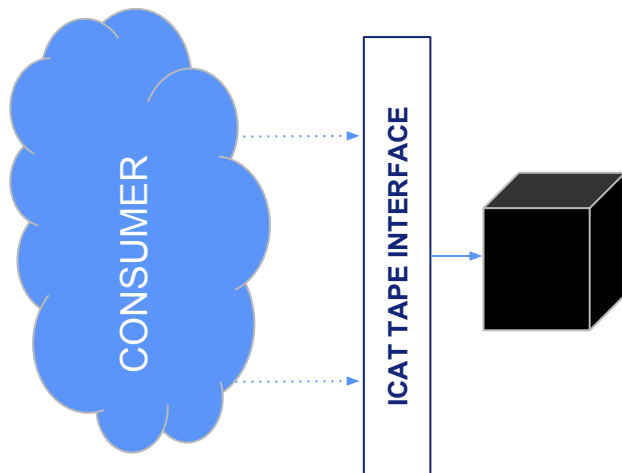1.- Create a file in a specific location

   /data/iti/inq/wait/mx1939.id30a3.67868525.413099800.csv

2.- With a specific file name

   /data/iti/inq/wait/mx1939.**id30a3**.67868525.**413099800**

   Investigation.**Beamline**.datasetId.**size**

3.- And a specific content

```
icat-dataserv:~ % more /data/iti/inq/wait/mx1939.id30a3.67868525.413099800.csv
/data/visitor/mx1939/id30a3/20171117/RAW_DATA/                    /pos9
ref-renato_ts_2_1_master.h5
ref-renato_ts_2_1_data_000001.h5
ref-renato_ts_2_2_master.h5
ref-renato_ts_2_2_data_000001.h5
ref-renato_ts_2_3_master.h5
ref-renato_ts_2_3_data_000001.h5
ref-renato_ts_2_4_master.h5
ref-renato_ts_2_4_data_000001.h5
mx1939-id30a3-ref-renato_ts_2_2158778.h5
```

IDS

2 levels : pseudo- 3 levels storage

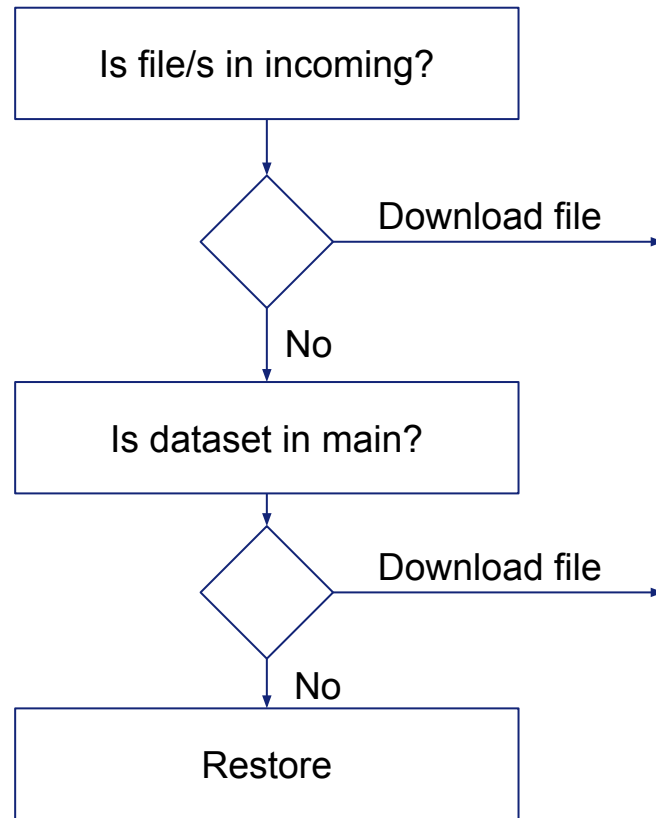# IDS Implementation at the ESRF

- ## IDS/ICAT restricciones
  - Nobody can add a new file to an existing dataset
  - Nobody can create datasets
  - Only RAW data
  - Only files that belong to a dataset
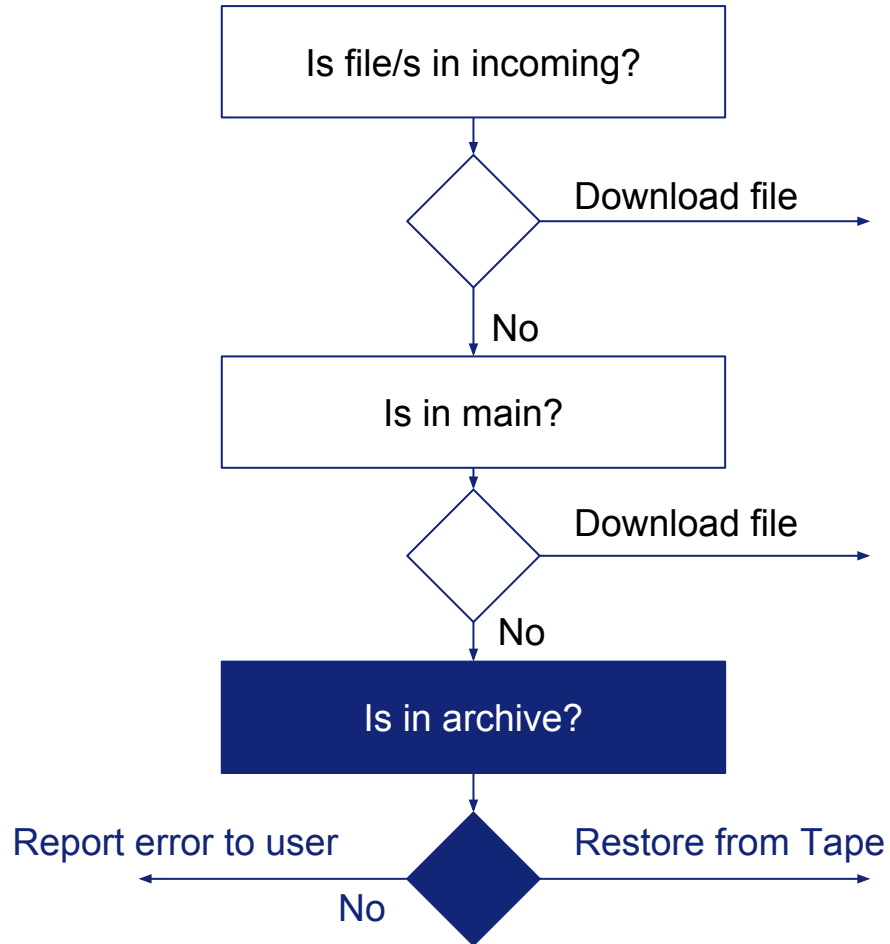
- ## Implementation
  - Storage Unit is DATASET
  - 2-level (pseudo 3)
    - IDS Main area
      - Incoming area (3 months)
      - Main area
    - Archive area

- ## Advantages
  - Files are available as soon as they are registered into ICAT

Is file/s in incoming?

→ Download file

No

Is dataset in main?

→ Download file
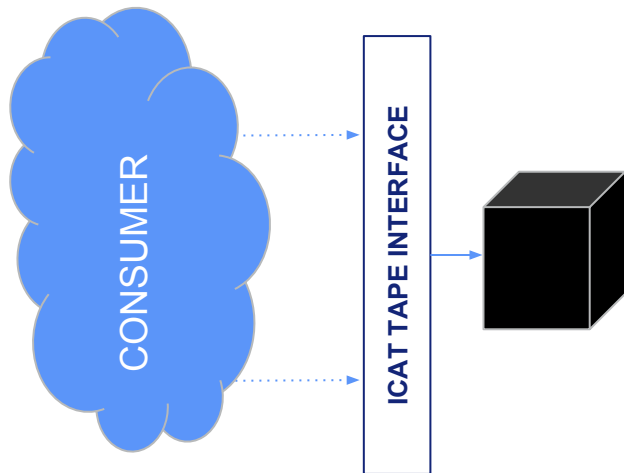
No

Restore

Desired IDS Implementation at the ESRF

How to restore a dataset?

1.- Create a file in a specific location

/data/iti/restore/67868525.csv

datasetId

Thanks!