Science & Technology
Facilities Council

Ingesting Data into ICAT

Kevin Phipps (RAL/STFC)
CLF and ISIS Service Manager
& ICAT Developer

October 2016

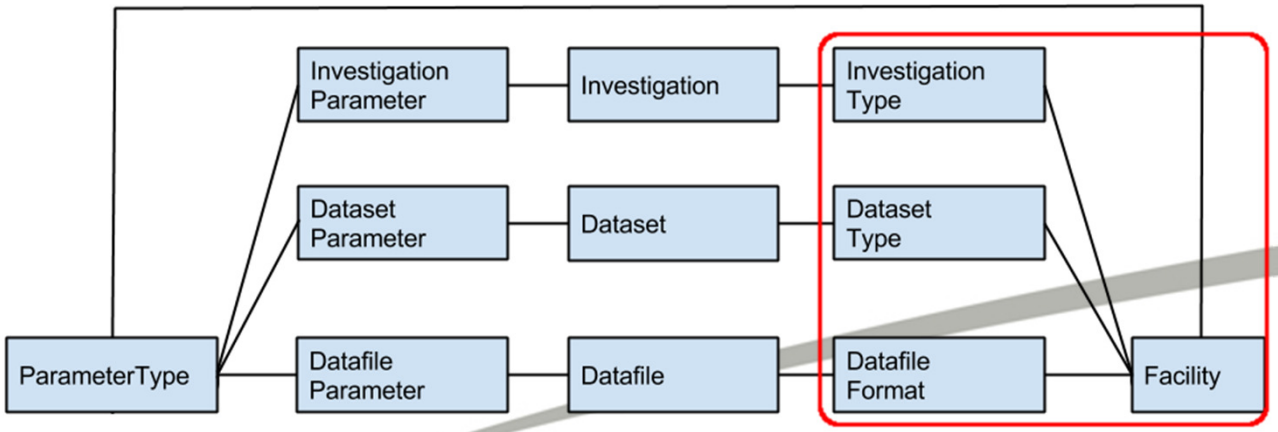
Overview

- Preparing the ICAT
- Setting up Investigations and Users
- Discovering new data
- Making data searchable
- The IDS – ingesting with and without
- Performance and scalability
- ISIS ingestion – a closer look

Science & Technology Facilities Council

Preparing the ICAT

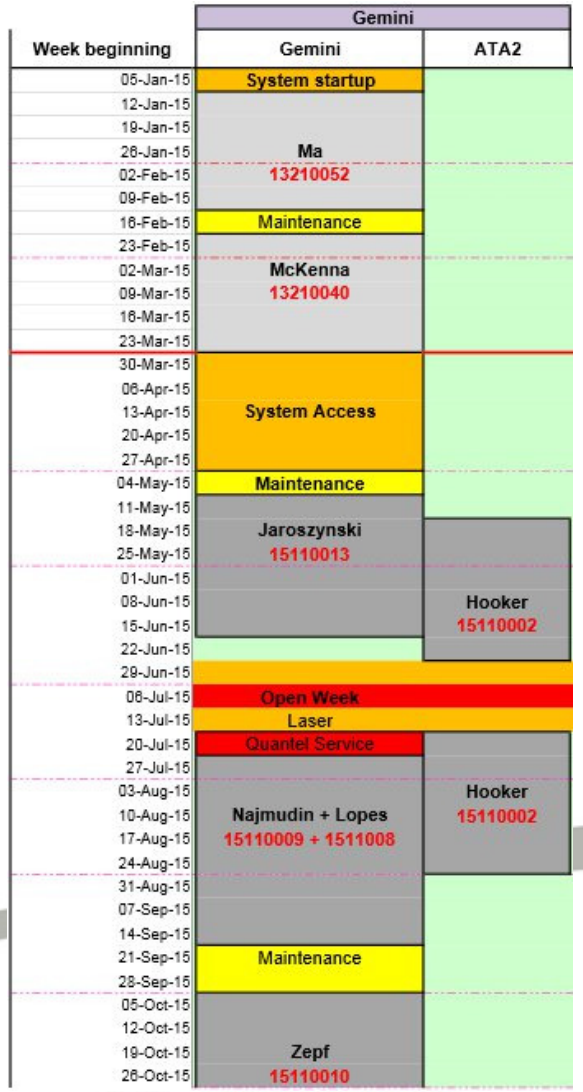
- Functional ingest account - CRU on most entities
- Create Facility, InvestigationTypes, DatasetTypes and DatafileFormats
- Create Instruments (for linking via InstrumentScientist and InvestigationInstrument)
- All just need a name (DatafileFormat also version)
- Code this to make it repeatable



Science & Technology Facilities Council

Investigations and Users

- Set up in advance of experiment
- Export data from User Office system
- Nightly/weekly job
- Create Investigations (name, visitId, title)
- Link Investigation to Instrument(s)
- Create User (name = ldap/rqw38472, fullName and email useful)
- Link Investigation to User and assign a role (for use in Rules)



Discovering new data

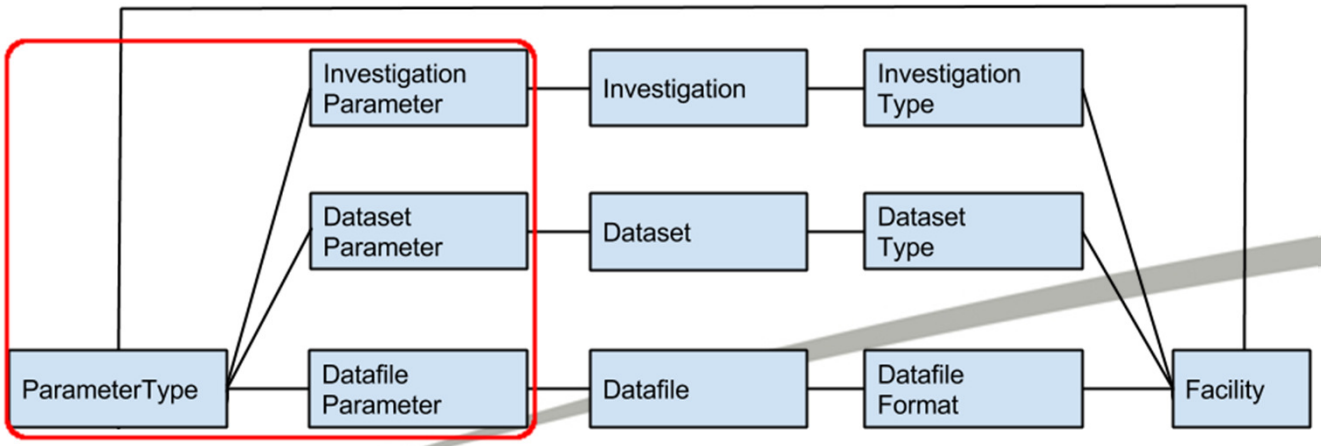
- Methods in use:
  - single input directory
  - filesystem watcher
  - trigger/drop files

```
20150329 GS 00114835 S_LEG2_GREEN_NF.dat
20150329 GS 00114835 S_LEG2_GREEN_NF.mdt.xml
20150329 GS 00114835 S_LEG2_GREEN_NF.png
20150329 GS 00114835 S_PUMP_TIMING_TRACE.dat
20150329 GS 00114835 S_RED_TIMING_TRACE.dat
20150329 GS 00114835 S_SCINTILLATOR_TRACE.dat
```

- Which Investigation?
  - use the file name or path
  - metadata in the file or drop file
  - file timestamp and Instrument
- What is a Dataset?
  - shot, run, scan etc.

Making data searchable

- Use Parameters (Investigation, Dataset and Datafile) to store metadata
- Name-value pairs of type NUMERIC, STRING, DATE_AND_TIME
- Nothing in a Datafile is searchable unless you create DatafileParameters



Science & Technology Facilities Council

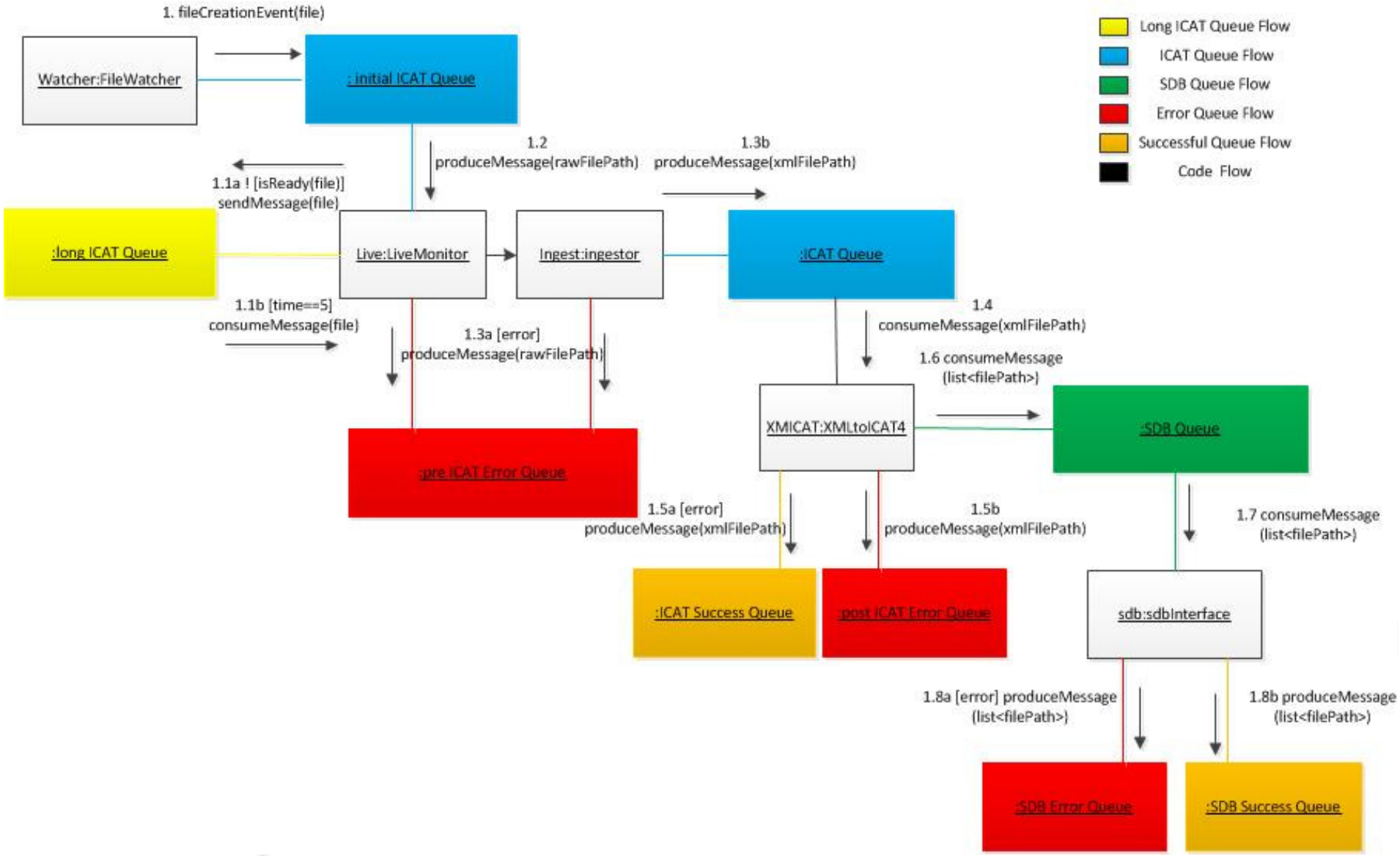
The IDS – ingesting with or without

- With:
  - for 2 level storage
  - if the data needs moving anyway
- Without:
  - large disk storage accessible from instruments
  - just create a Datafile entry in ICAT
  - add any metadata as DatafileParameters

Performance and scalability

- Cache entities used frequently – DatasetTypes, DatafileFormats, ParameterTypes
- Cache recently used Investigations and Datasets
- Make it multi-threaded – one thread for queueing plus multiple ‘worker’ threads
- Design around an enterprise queueing system and run multiple ingestion program instances

ISIS ingestion – a closer look



Conclusions

- Import Investigation and User information from a User Office
- Decide how searchable data needs to be and use Parameters accordingly
- Size of available file stores and/or budget will determine the IDS setup
- Design for performance and scalability

Science & Technology Facilities Council

Thanks for listening.

Questions ?