

# **The ICAT Rules System**

Tom Griffin Computing Group Leader STFC ISIS Neutron and Muon Source NOBUGS 2016 ICAT Workshop

tom.griffin@stfc.ac.uk

### Overview

- · Why rules
- Data policy
- Open tables
- Administration and Data Ingestion
- Facility staff/instrument scientists
- Experiment Team
- · Data Upload
- $\cdot$  Embargo



# Why Rules

- ICAT 3.3
  - Explicit authorisation table + triggers
  - >5,000,000 million entries for 150,000 investigations
  - Limited in scope Investigation & Dataset
  - Hard to change when staff move
- $\cdot$  More flexible
- · Addition of 'InvestigationGroups'



# Some background

- By default everything is closed (deny)
- ISIS data policy requires >100 rules
- Are sanity checked on creation
- Required at every level for **direct** access (see PublicSteps)
- $\cdot$  Can be checked

isAccessAllowed(String sessionId, EntityBaseBean bean, AccessType accessType)



- $\cdot$  ISIS
- <u>http://www.isis.stfc.ac.uk/user-office/data-policy11204.html</u>
- <u>http://tiny.cc/isisdp</u>
- PaNData Common Framwork
- <u>http://wiki.pan-</u> <u>data.eu/imagesGHD/0/08/PaN-data-D2-</u> <u>1.pdf</u>



#### Data Policy

ISIS data management policy

#### 1. General principles

1.1 This data management policy pertains to the curation of and access to experimental raw data and metadata collected and / or stored at the ISIS facility.

1.2 Acceptance of this policy is a condition of the award of beamtime at ISIS.

1.3 Users must not attempt to access, exploit or distribute raw data or metadata unless they are entitled to do so under the terms of this policy.

1.4 Deliberate infringements of the policy may lead to denial of access to ISIS raw data or metadata.

#### 2. Definitions

#### For the purposes of this policy:

2.1 'raw data' are data collected from experiments performed on ISIS instruments. This definition includes data that are created automatically or manually by Facility-specific software and/or ISIS staff expertise in order to facilitate subsequent analysis of the experimental data, unless otherwise agreed.

2.2 'metadata' is information pertaining to data collected from experiments performed on ISIS instruments, including (but not limited to) the context of the experiment, the experimental team (in accordance with the Data Protection Act), experimental conditions and other logistical information.

2.3 the term 'principal investigator' (PI) pertains to the PI identified on the ISIS experiment proposal.

2.4 the term 'experimental team' includes the PI and any other person to whom the PI designates the right to access resultant raw data and associated metadata.

2.5 the term 'on-line catalogue' pertains to a computer database of metadata containing links to raw data files, that can be accessed by a variety of methods, including (but not limited to) www-based browsers.

2.6 the term 'results' pertains to data, intellectual property, and outcomes arising from the analysis of raw data.

2.7 the term 'long-term' means a minimum of ten years.

2.8. the term 'public domain' means belonging to the community at large, unprotected by copyright or patent and subject to appropriation by anyone.

#### 3. Raw data and associated metadata

3.1 Free and commercial access to ISIS

3.1.1 All raw data and the associated metadata obtained as a result of free (non-commercial) access to ISIS, reside in the public domain, with ISIS acting as the custodian.

3.1.2 All raw data and the associated metadata obtained as a result of 'commercial-in-confidence' access to ISIS will be owned exclusively by the commercial user. Commercial users must agree with their relevant instruments scientists how they wish their raw data and metadata to be managed before the start of any experiment.

3.2 Curation of raw data and associated metadata

3.2.1 All raw data will be curated in well-defined formats, for which the means of reading the data will be made available by the Facility.

3.2.2 Metadata that is automatically captured by instruments will be curated either within the raw data files, within an associated on-line catalogue, or within both.

3.2.3 Data will be stored initially on instrument-related computers, and will be migrated or copied to archival facilities for long-term curation.

3.3 Access to raw data and metadata

3.3.1 Access to raw data and metadata beyond the period that it is stored on instrument-related computers will be via a searchable on-line catalogue.

3.3.2 Access to the on-line catalogue will be restricted to those who register with STFC/ISIS as users of the on-line catalogue.

3.3.3 Access to raw data and the associated metadata obtained from an experiment is restricted to the experimental team for a period of three years after the end of the experiment. Thereafter, it will become publicly accessible. Any PI that wishes their data to remain 'restricted access' for a longer period will be required to make a special case to the Director of ISIS.

3.3.4 It is the responsibility of the PI to ensure that the experiment number (RB number) is correctly entered into the metadata for each raw data set, in order to correctly associate each data set with the PI. If this is not done, the experimental team will not be able to access the data via the on-line catalogue and other users may inadvertently be given access rights to the data.

3.3.5 Appropriate STFC staff (e.g. instrument scientists, computing group members) may be given access to any Facility-curated data or metadata for Facility-related purposes. ISIS undertakes that they will preserve the confidentiality of such data.

3.3.6 The on-line catalogue will enable the linking of experimental data to experimental proposals. Access to proposals will only ever be provided to the experimental team and appropriate STFC staff, unless otherwise authorized by the PI. 4. Results

#### 4.1 Ownership of results

4.1.1 Ownership of all results derived from the analysis of the raw data is determined by the contractual obligations of the person(s) performing the analysis.

4.2 Curation of results

4.2.1 ISIS will provide a facility for users to upload results and associated metadata to the facility and enable them to associate these results with raw data collected from the Facility.

4.2.2 The upload of results and metadata are subject to volume restrictions.

4.2.3 These results will be stored long-term by the Facility. It will not be the responsibility of the Facility to fully curate this data e.g. to ensure that software to read / manipulate this data is available.

4.3 Access to results

4.3.1 Access to the results of analyses performed on raw data and metadata is restricted to the person or persons performing the analyses, unless otherwise requested by those persons.

#### 5. Good practice for metadata capture and results storage

5.1 The experimental team is encouraged to ensure that experimental metadata are as complete as possible, as this will enhance the possibilities for them to search for, retrieve and interpret their own data in the future.

5.2 ISIS undertakes to provide facilities for the capture of such metadata items that are not automatically captured by an instrument, in order to facilitate recording the fullest possible description of the raw data.

5.3 Researchers who aim to carry out analyses of raw data and metadata which are publicly accessible should, where possible, contact the original PI to inform them and suggest a collaboration if appropriate.

5.4 Pls and researchers who carry out analyses of raw data and metadata are encouraged to link the results of these analyses with the raw data / metadata using the facilities provided by the on-line catalogue. Furthermore, they are encouraged to make such results publicly accessible.

#### 6. Publication information

6.1 References for publications related to experiments carried out at ISIS must be deposited in the STFC e-Pubs system <a href="http://epubs.cclrc.ac.uk/">http://epubs.cclrc.ac.uk/</a> within six months of the publication date, or during any new application for beamtime, whichever is the earlier.

#### 3. Raw data and associated metadata

3.1 Free and commercial access to ISIS

3.1.1 All raw data and the associated metadata obtained as a result of free (non-commercial) access to ISIS, reside in the public domain, with ISIS acting as the custodian.

3.1.2 All raw data and the associated metadata obtained as a result of 'commercial-in-confidence' access to ISIS will be owned exclusively by the commercial user. Commercial users must agree with their relevant instruments scientists how they wish their raw data and metadata to be managed before the start of any experiment.

3.2 Curation of raw data and associated metadata

3.2.1 All raw data will be curated in well-defined formats, for which the means of reading the data will be made available by the Facility.

3.2.2 Metadata that is automatically captured by instruments will be curated either within the raw data files, within an associated on-line catalogue, or within both.

3.2.3 Data will be stored initially on instrument-related computers, and will be migrated or copied to archival facilities for long-term curation.

3.3 Access to raw data and metadata

3.3.1 Access to raw data and metadata beyond the period that it is stored on instrument-related computers will be via a searchable on-line catalogue.

3.3.2 Access to the on-line catalogue will be restricted to those who register with STFC/ISIS as users of the on-line catalogue.

**3.3.3** Access to raw data and the associated metadata obtained from an experiment is restricted to the experimental team for a period of three years after the end of the experiment. Thereafter, it will become publicly accessible. Any PI that wishes their data to remain 'restricted access' for a longer period will be required to make a special case to the Director of ISIS.

3.3.4 It is the responsibility of the PI to ensure that the experiment number (RB number) is correctly entered into the metadata for each raw data set, in order to correctly associate each data set with the PI. If this is not done, the experimental team will not be able to access the data via the on-line catalogue and other users may inadvertently be given access rights to the data.

3.3.5 Appropriate STFC staff (e.g. instrument scientists, computing group members) may be given access to any Facilitycurated data or metadata for Facility-related purposes. ISIS undertakes that they will preserve the confidentiality of such data.

3.3.6 The on-line catalogue will enable the linking of experimental data to experimental proposals. Access to proposals will only ever be provided to the experimental team and appropriate STEC staff, unless otherwise authorized by the PL

#### 4. Results

4.1 Ownership of results

4.1.1 Ownership of all results derived from the analysis of the raw data is determined by the contractual obligations of the person(s) performing the analysis.

4.2 Curation of results

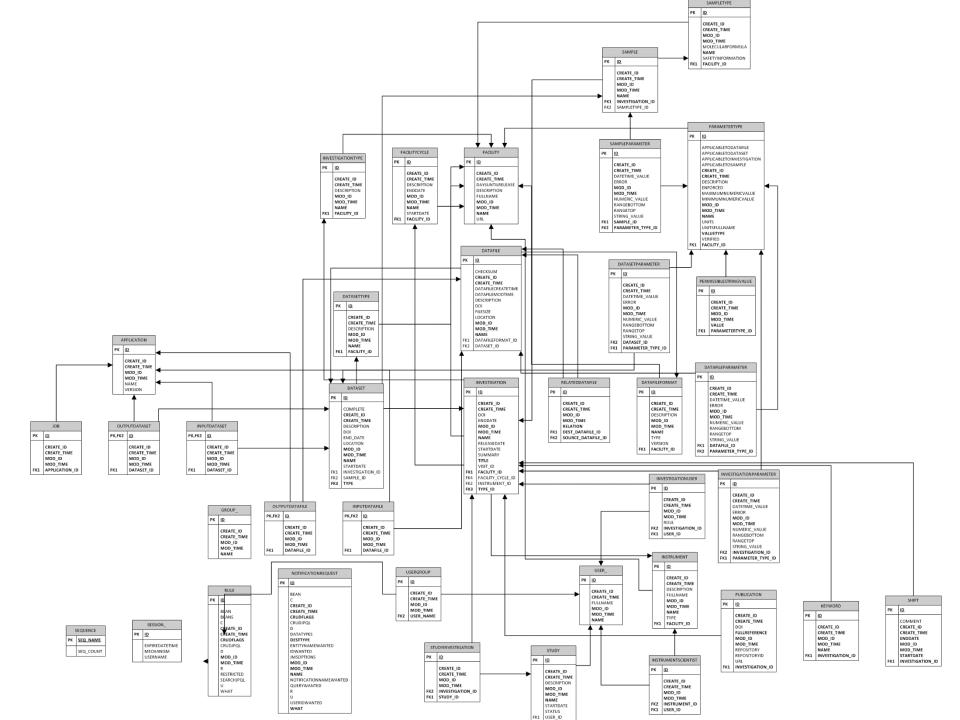
4.2.1 ISIS will provide a facility for users to upload results and associated metadata to the facility and enable them to associate these results with raw data collected from the Facility.

4.2.2 The upload of results and metadata are subject to volume restrictions.

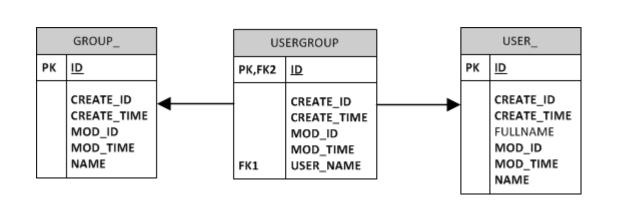
4.2.3 These results will be stored long-term by the Facility. It will not be the responsibility of the Facility to fully curate this data e.g. to ensure that software to read / manipulate this data is available.

4.3 Access to results

4.3.1 Access to the results of analyses performed on raw data and metadata is restricted to the person or persons performing the analyses, unless otherwise requested by those persons.



## **Special Tables**



RULE	
РК	<u>ID</u>
	BEAN BEANS C CREATE_ID CREATE_TIME CRUDFLAGS CRUDJPQL D MOD_ID MOD_ID MOD_TIME R RESTRICTED SEARCHJPQL U WHAT

"root" group



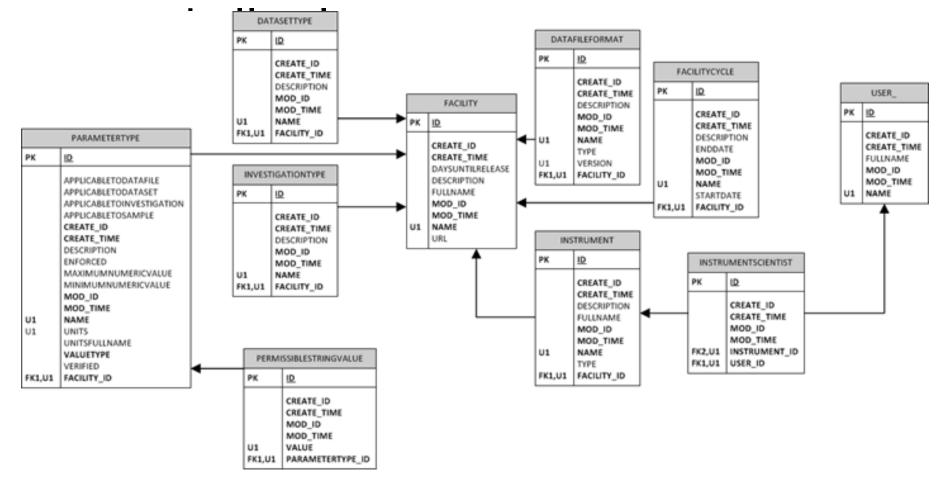
### **Rules Syntax**

#### [Group] CRUD Flags 'What'

Rule myRule = new Rule(); myRule.group = facilityAdmins; myRule.crudFlags = "CRUD"; myRule.what = "Facility";



#### · Allow any authenticated reader to



```
List<String> publicTables = new ArrayList<>();
publicTables.add("Application");
publicTables.add("DatafileFormat");
publicTables.add("DatasetType");
publicTables.add("Facility");
publicTables.add("FacilityCycle");
publicTables.add("Instrument");
publicTables.add("InstrumentScientist");
publicTables.add("InvestigationType");
publicTables.add("ParameterType");
publicTables.add("PermissibleStringValue");
publicTables.add("Publication");
publicTables.add("Shift");
```

```
for (String table : publicTables)
{
    Rule rule = new Rule();
    rule.crudFlags = "R";
    rule.what = table;
    rule.setGroup(null);
    port.create(sessionId, rule);
}
```



List<String> publicTables = **new** ArrayList<>(); publicTables.add("Application"); publicTables.add("DatafileFormat"); publicTables.add("DatasetType"); publicTables.add("Facility"); publicTables.add("FacilityCycle"); publicTables.add("Instrument"); publicTables.add("InstrumentScientist"); publicTables.add("InvestigationType"); publicTables.add("ParameterType"); publicTables.add("PermissibleStringValue"); publicTables.add("Publication"); publicTables.add("Shift");



for (String table : publicTables)
{
 Rule rule = new Rule();
 rule.crudFlags = "R";
 rule.what = table;
 rule.setGroup(null);
 port.create(sessionId, rule);

}



```
List<String> publicTables = new ArrayList<>();
publicTables.add("Application");
publicTables.add("DatafileFormat");
publicTables.add("DatasetType");
publicTables.add("Facility");
publicTables.add("FacilityCycle");
publicTables.add("Instrument");
publicTables.add("InstrumentScientist");
publicTables.add("InvestigationType");
publicTables.add("ParameterType");
publicTables.add("PermissibleStringValue");
publicTables.add("Publication");
publicTables.add("Shift");
```

```
for (String table : publicTables)
{
    Rule rule = new Rule();
    rule.crudFlags = "R";
    rule.what = table;
    rule.setGroup(null);
    port.create(sessionId, rule);
```



## **Public Steps**

- Allow access to a related object (attribute)
- · ONLY used in INCLUDE processing.
- Offer a significant speedup on INCLUDE
- Thing -> ThingParameter
- Investigation -> samples, publications, users, investigationInstrument etc



### **Public Steps**

```
String[] parameterBasedPublicSteps = new
String[]{"Investigation", "Dataset", "Datafile",
                "Sample"};
for (String table : parameterBasedPublicSteps)
{
                PublicStep paramPublicStep = new PublicStep
```

}

```
PublicStep paramPublicStep = new PublicStep();
paramPublicStep.origin = table;
paramPublicStep.field = "parameters";
publicSteps.add(paramPublicStep);
```



### **Public Steps**

```
String[] publicStepsFromInvestigation = new
String[]{"samples", "publications", "shifts",
"investigationUsers", "keywords",
"investigationInstruments" };
```

{

```
for (String step : publicStepsFromInvestigation)
      PublicStep invToSomething = new PublicStep();
      invToSomething.origin = "Investigation";
      invToSomething.field = step;
      publicSteps.add(invToSomething);
```



# Breaking down the data policy

- 3.3.5 Appropriate STFC staff (e.g. instrument scientists, **COMPUTING GROUP MEMBERS**) may be given access to any Facility-curated data or metadata for Facility- related purposes. ISIS undertakes that they will preserve the confidentiality of such data
- FacilityAdmins group
  - Do anything to anything
  - To allow fixes, investigations
  - Could do a 'R'-all group too for safety



### Facility Admin Access

```
Group facilityAdmins = (Group)port.search(sessionId,
"Group[name='FacilityAdmins']" ).get(0);
```

```
List<String> adminTables = new ArrayList<String>();
adminTables.add("Application");
adminTables.add("Datafile");
...
```

```
adminTables.add("DatafileFormat");
adminTables.add("User");
adminTables.add("UserGroup");
```

```
for (String table : adminTables)
{
    Rule rule = new Rule();
    rule.group = facilityAdmins;
    rule.crudFlags = "CRUD";
    rule.what = table;
    port.create(sessionId, rule);
```



# Breaking down the data policy

3.3.1 Access to raw data and metadata beyond the period that it is stored on instrument-related computers will be via a searchable on-line catalogue.

#### DataIngest

- Automatic processes
- Powerful account running in many places
- Restrict it if we can



### Data Ingest Account

- · Create a group "DataIngestors"
- Very similar to admin rules just no delete

Application

Datafile DatafileFormat DatafileParameter

Dataset DatasetParameter DatasetType InputDatafile InputDataset Publication RelatedDatafile

Investigation InvestigationParameter InvestigationUser Sample SampleParameter SampleType

Job Keyword ParameterType Study StudyInvestigation User



#### Data Ingest Account

```
Group dataIngestors = (Group)port.search(sessionId,
          "Group[name='DataIngestors']" ).get(0);
List<String> ingestorTables = new ArrayList<String>();
```

```
ingestorTables.add("Application");
...
ingestorTables.add("User");
for(String table : ingestorTables)
{
    Rule rule = new Rule();
    rule.group = dataIngestors;
    rule.crudFlags = "CRU"; //no delete permission for ingestors
    rule.what = table;
    port.create(sessionId, rule);
}
```



## More Complex Rules - JPQL

- Simple rules are.....simple
- More complex rules ("where....", if etc) need JPQL



# (Basic) JPQL in 1 slide

- SELECT ds FROM Dataset ds WHERE ds.name = 'abc'
- SELECT ds.id FROM Dataset ds WHERE ds.name = 'abc'
- SELECT i FROM Investigation i, i.investigationUsers ius WHERE ius.user.name = 'tom'
- SELECT i FROM Investigation i JOIN i.investigationUsers ius WHERE ius.user.name = :user



# (Less basic) JPQL in 1 slide

SELECT i FROM Investigation i WHERE
 i.createTime > {ts 2011-01-15 00:00:00}

 SELECT ds.id FROM Dataset ds, ds.parameters dp1, ds.parameters dp2
 WHERE dp1.type.name = 'A1' AND dp1.numericValue > 50 AND dp2.type.name = 'A2' AND dp2.numericValue > 20

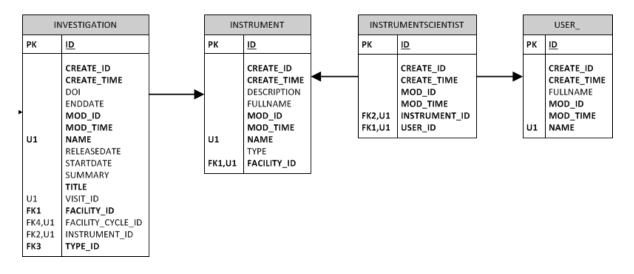


# Breaking down the data policy

- 3.3.5 Appropriate STFC staff (e.g. instrument scientists, computing group members) may be given access to any Facility-curated data or metadata for Facility-related purposes. ISIS undertakes that they will preserve the confidentiality of such data
- Relates to an ICAT Concept : InstrumentScientist



#### Instrument Scientists - Investigation

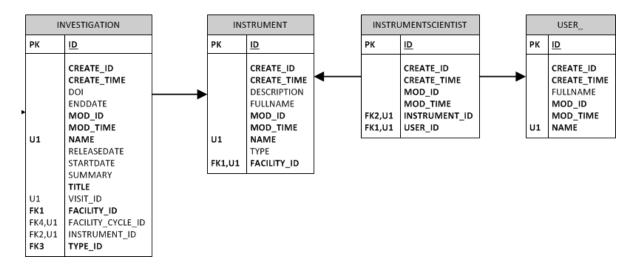


SELECT i FROM Investigation i

- JOIN i.investigationInstruments ii
- JOIN ii.instrument inst
- JOIN inst.instrumentScientists instSci
- JOIN instSci.user u
- WHERE u.name = :user



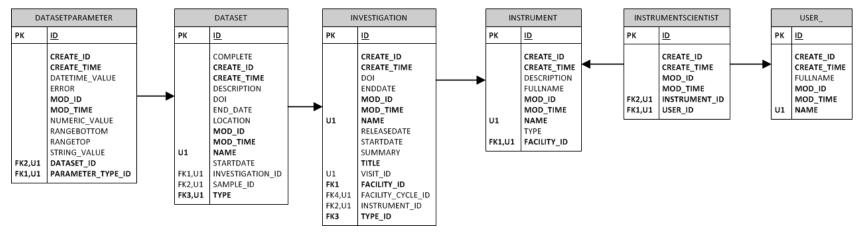
#### **Instrument Scientists - Investigation**



```
Rule isInv = new Rule();
isInv.crudFlags = "R";
isInv.what = "SELECT i FROM Investigation i JOIN i.investigationInstruments ii
JOIN ii.instrument inst JOIN inst.instrumentScientists instSci JOIN instSci.user u
WHERE u.name = :user";
isInv.group = null; //rules applies regardless of group membership (not explicitly
needed)
port.create(sessionId, isInv);
```



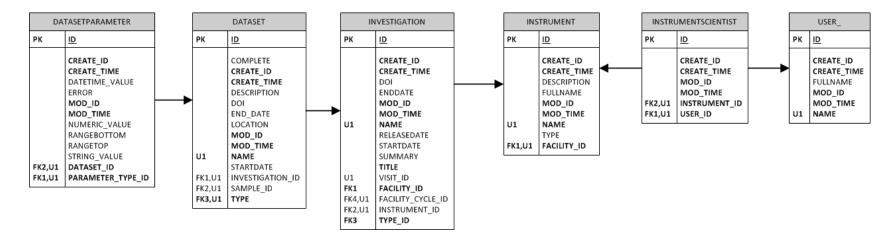
#### Instrument Scientists - Dataset



- SELECT d FROM Dataset d
- JOIN d.investigation i
- JOIN i.investigationInstruments ii
- JOIN ii.instrument inst
- JOIN inst.instrumentScientists instSci
- JOIN instSci.user u
- WHERE d.name='raw' //ISIS
- AND u.name = :user";



#### Instrument Scientists - Dataset



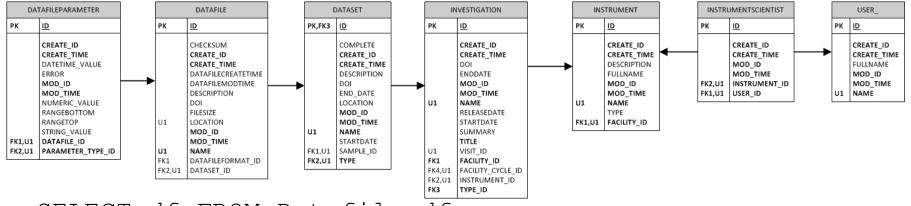
#### Rule isDs = new Rule();

isDsParam.crudFlags = "R";

isDsParam.what = "SELECT d from Dataset d JOIN d.investigation i JOIN i.investigationInstruments ii JOIN ii.instrument inst JOIN inst.instrumentScientists instSci JOIN instSci.user u WHERE d.name='raw' AND u.name = :user"; port.create(sessionId, isDsParam);



#### Instrument Scientists - Datafile



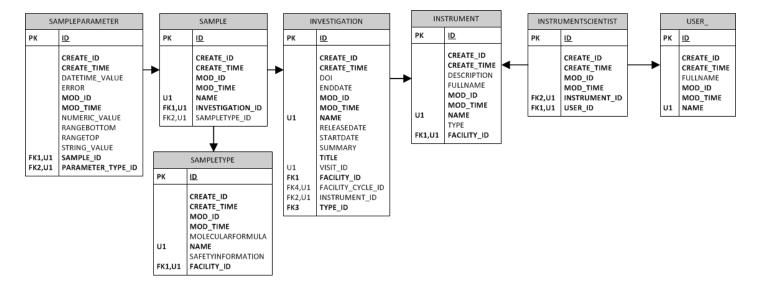
SELECT df FROM Datafile df

- JOIN df.dataset d
- JOIN d.investigation i
- JOIN i.investigationInstruments ii
- JOIN ii.instrument inst
- JOIN inst.instrumentScientists instSci
- JOIN instSci.user u
- WHERE d.name='raw'

AND u.name = :user";



#### Instrument Scientists - Samples



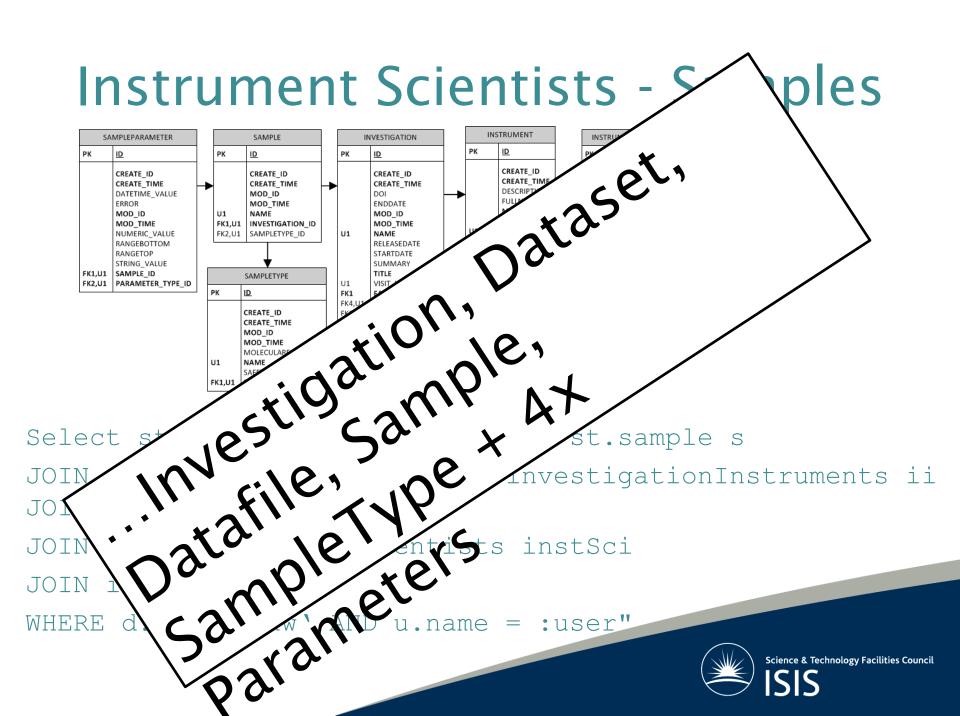
Select st from SampleType st JOIN st.sample s

JOIN s.investigation i JOIN i.investigationInstruments ii JOIN ii.instrument inst

- JOIN inst.instrumentScientists instSci
- JOIN instSci.user u

WHERE d.name= 'raw' AND u.name = :user"





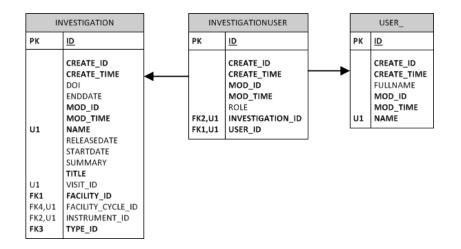
#### Breaking down the data policy

3.3.3 Access to raw data and the associated metadata obtained from an experiment **is** restricted to the experimental team for a period of three years after the end of the experiment.

 Relates to an ICAT Concept : InvestigationUser OR InvestigationGroup



#### **Experiment Team**

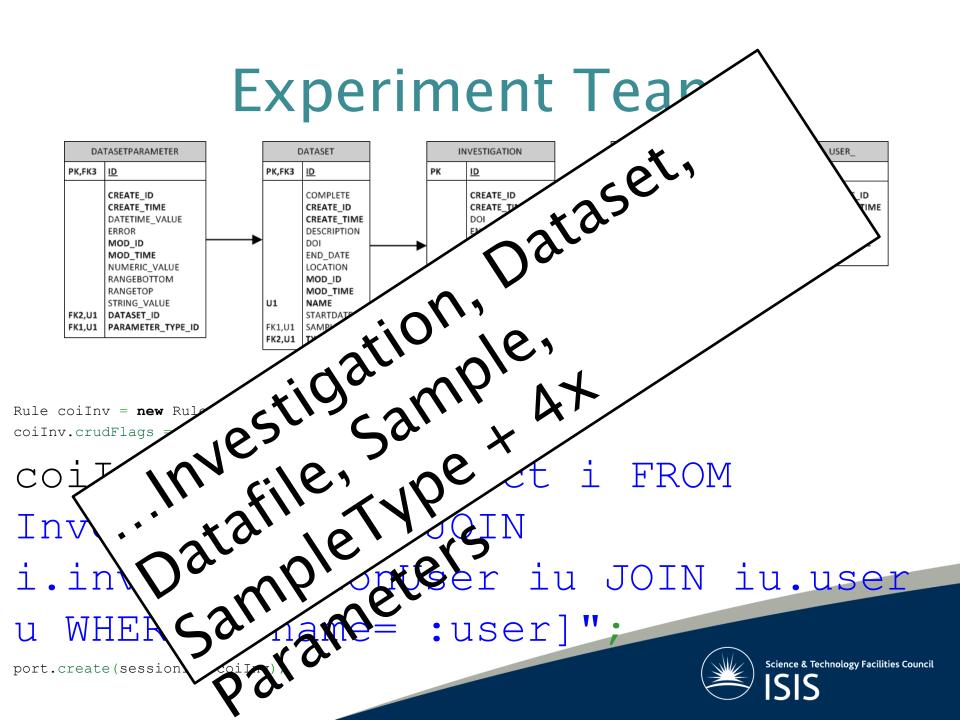


```
Rule coiInv = new Rule();
coiInv.crudFlags = "R";
```

```
coiInv.what = "Select i FROM
Investigation i JOIN
i.investigationUser iu JOIN iu.user u
WHERE u.name= :user]";
```

port.create(sessionId, coiInv);





#### Breaking down the data policy

#### 4. Results

4.1 Ownership of results

4.1.1 Ownership of all results derived from the analysis of the raw data is **determined by the contractual** 

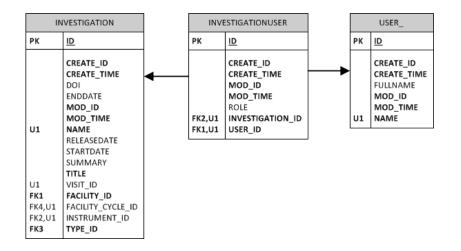
**obligations** of the person(s) performing the analysis.

4.2 Curation of results

4.2.1 ISIS will provide a facility for users to upload results and associated metadata to the facility and enable them to associate these results with raw data collected from the Facility.



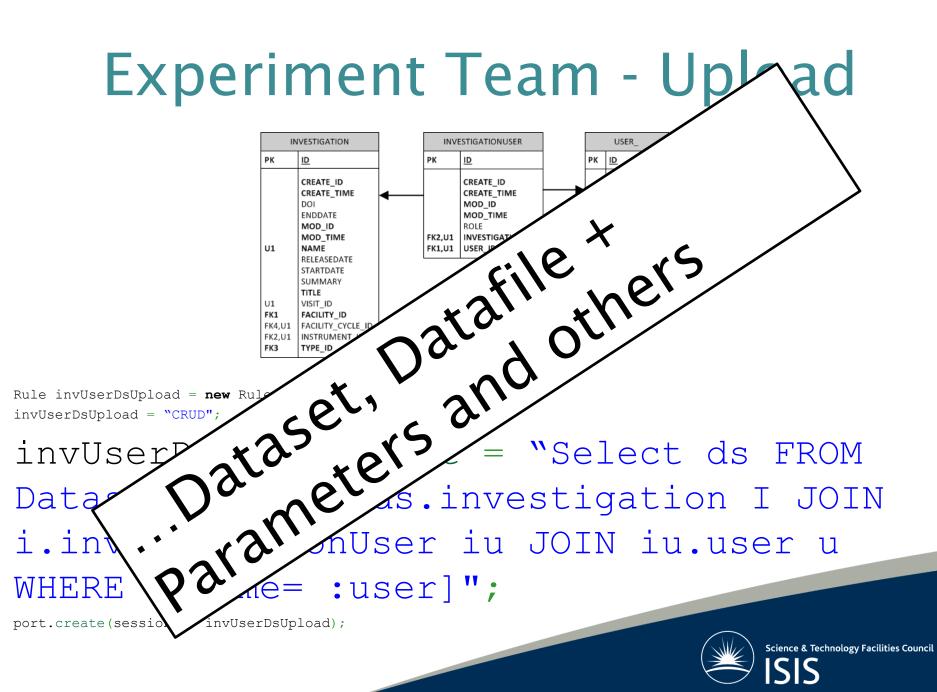
#### **Experiment Team - Upload**



```
Rule invUserDsUpload = new Rule();
invUserDsUpload = "CRUD";
```

invUserDsUpload.what = "Select ds FROM Dataset ds JOIN ds.investigation I JOIN i.investigationUser iu JOIN iu.user u WHERE ds.name <> `raw' and u.name= :user]"; port.create(sessionId, invUserDsUpload);





# Breaking down the data policy

3.3.3 Access to raw data and the associated metadata obtained from an experiment is restricted to the experimental team for a period of three years after the end of the experiment. Thereafter, it will become publicly accessible. Any PI that wishes their data to remain 'restricted access' for a longer period will be required to make a special case to the Director of ISIS.

 Can't use simple dates = Investigation.release\_date



### Embargo – without extension

rule.setCrudFlags("R");

rule.setWhat(

"Select i FROM Investigation i JOIN i.Facility f where

i.enddate + f.daysuntilrelease] < Now ");</pre>



#### Embargo - allowing extensions

```
rule.setCrudFlags("R");
rule.setWhat("SELECT i FROM Investigation i WHERE
i.releaseDate < CURRENT_TIMESTAMP");</pre>
```

Must set Investigation.releasedate as part of data ingest

Rules would need to consider non-facility owned data dataset.type = 'raw'



#### Public Access After Embargo

• Read for all authenticated users

```
SELECT i FROM Investigation i WHERE i.releaseDate < CURRENT_TIMESTAMP
```

```
SELECT dfp FROM DatafileParameter dfp
JOIN dfp.datafile df
JOIN df.dataset d
JOIN d.investigation i
WHERE d.name = `raw'
AND i.releaseDate < CURRENT_TIMESTAMP</pre>
```

Investigation, Dataset, Datafile, Sample, SampleType
+ 4x Parameters



#### Public Access After Embargo

rule.setCrudFlags("R");

rule.setWhat ("SELECT i FROM

Investigation i JOIN

- i.InvestigationType it WHERE
- it.name IN calibration',

`commissioning'"



#### **DOI Account**

 Requires read access to generate landing pages

SELECT i FROM Investigation i WHERE i.doi IS NOT NULL

SELECT ds FROM Dataset ds WHERE ds.doi IS NOT NULL

#### DOI creation runs as Data Ingestor



## **Delegating Permissions**

 Allow a data owner to grant other users access to their data

4.3.1 Access to the results of analyses performed on raw data and metadata is restricted to the person or persons performing the analyses, unless otherwise requested by those persons.



## **Delegating Permissions**

InvestigationGroup?

Talk to Rolf!

Allow PI to update the membership of that group

Requires a group per investigation/dataset

ISIS implements this via the business system + ingest



#### Summary

- Rules syntax = JPQL
- Examples given of implementing a PaNdata like policy

   <u>https://github.com/icatproject-contrib/example-rules</u>
- All tables need to be considered
- Limitations and scalability
- Must be carefully thought out



#### Questions...



