Science and Technology Facilities Council

ICAT Update

PaNOSC WP3 Catalogue Integration
Best Practices Meeting
May 2021

Stuart Pullinger
The ICAT Collaboration

ICAT Collaboration

- **ICAT** has been developed for over 10 years
- **ICAT** is in production at several facilities – in both **PaNOSC** and **ExPaNDS**.
- Recently ALBA and CERIC have joined the ICAT Colaboration

Science and Technology Facilities Council

Agenda

1 ICAT Overview

Architecture – Components – Data Model

2 Schema Additions

Techniques – Data Publication – And more

3 APIs & DataGateway

DataGateway API – PaNOSC Search API – DataGateway

4 Mapping Facility Entities to Schema and OAI-PMH

From ICAT Schema to OAI-PMH

5 Future Plans

OpenID Connect – Improve Search – Cloud

Science and Technology Facilities Council

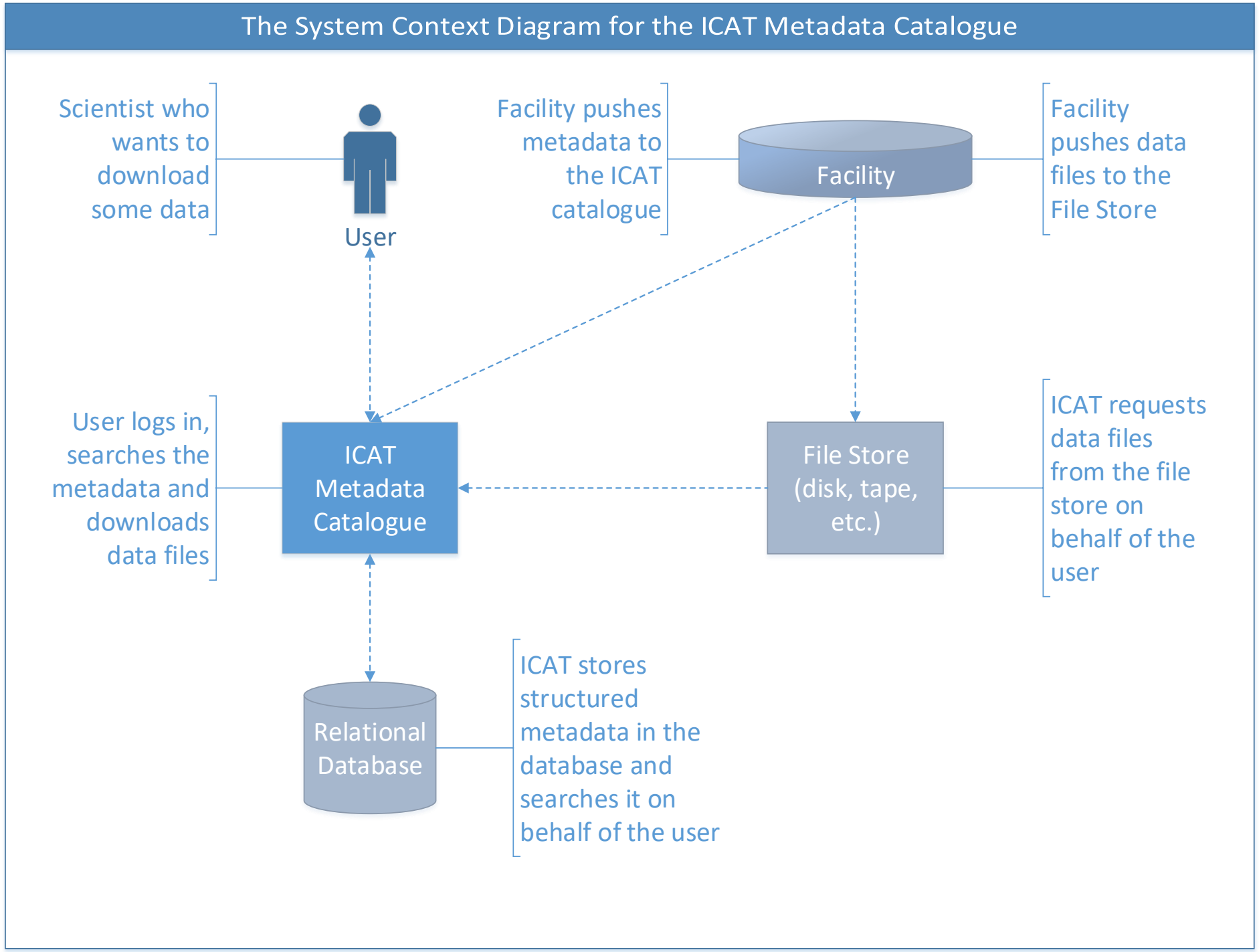
Science and Technology Facilities Council

ICAT Overview

Architecture – Components – Data Model

ICAT

Architecture



Science and Technology Facilities Council

ICAT

Technology

- Written in **Java EE**
- Runs on **Payara**
  - open-source successor to Sun/Oracle Glassfish Application Server
- Search via **Lucene** component

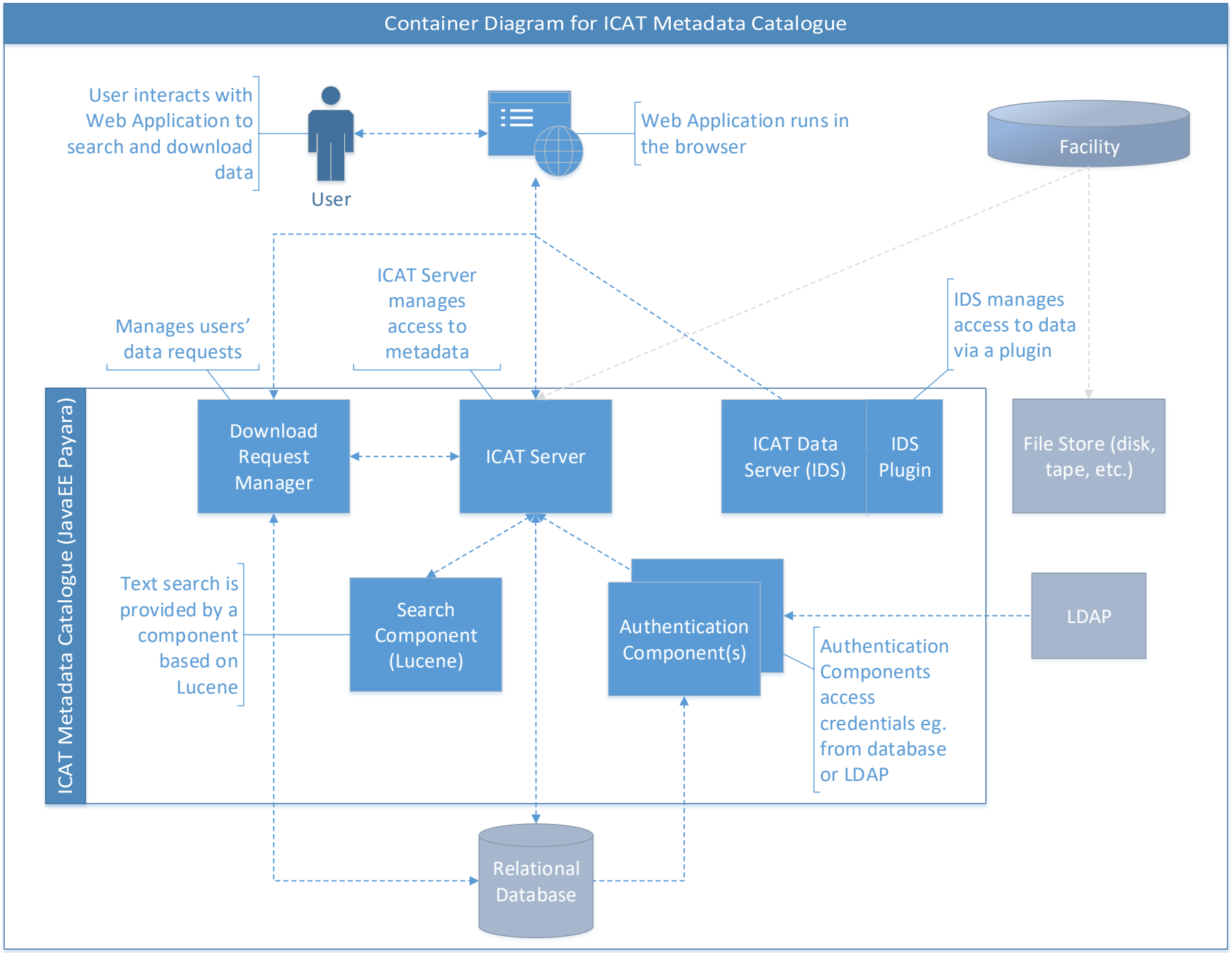
Databases

- **MySQL/MariaDB** & **Oracle** supported
- Any JPA-compatible* database ought to be possible

* Java Persistence API

ICAT

Components



Science and Technology Facilities Council

ICAT Data Model

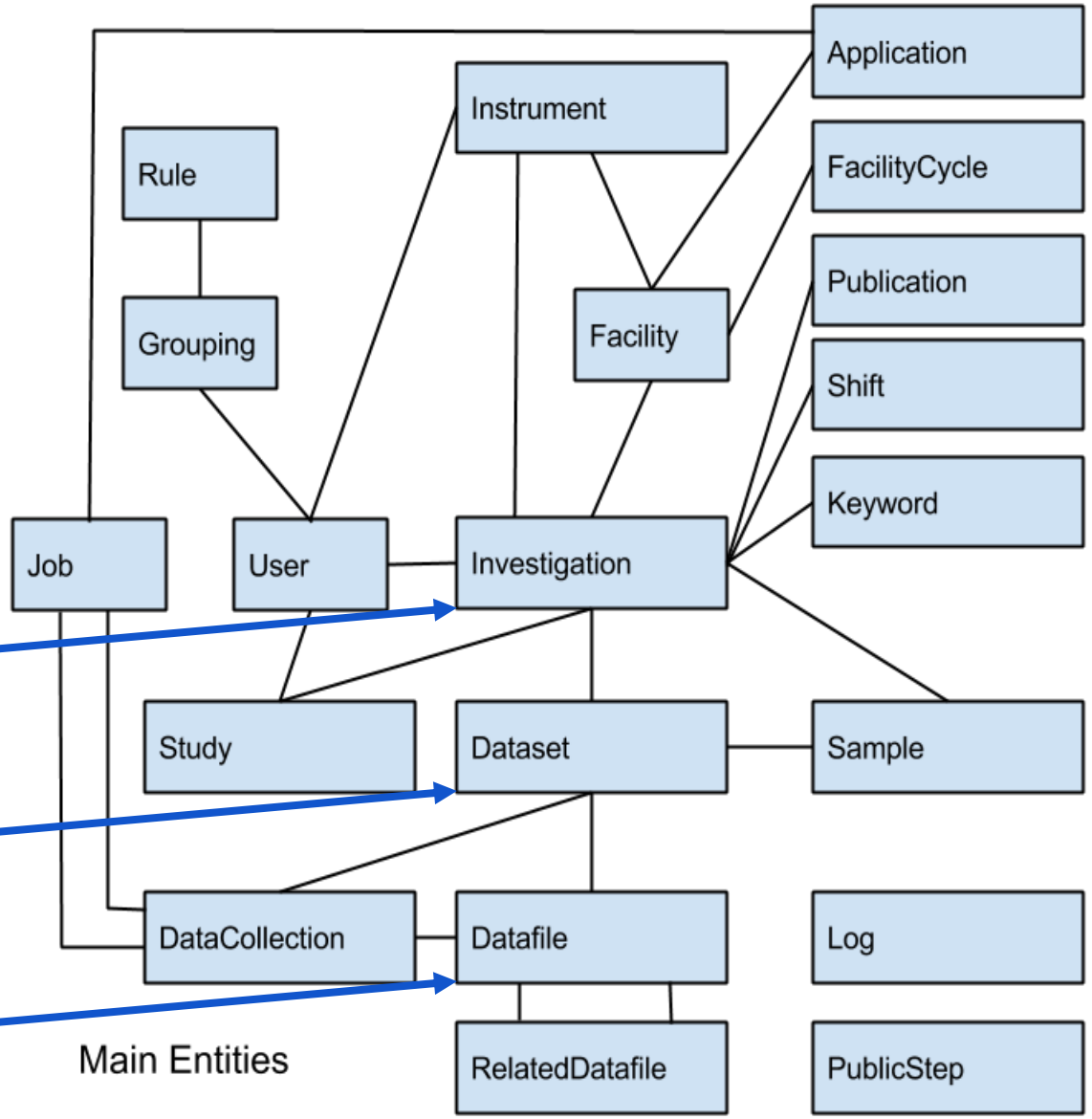
Based on Core Scientific Metadata Model:
http://icatproject-contrib.github.io/CSMD/

Core Entities:

Investigation

Dataset

Datafile



Science and Technology Facilities Council

Science and Technology Facilities Council

Schema Additions

Techniques – Data Publications – And more

Techniques

- **PaNOSC Data Model includes ‘Technique’ entity**
  - “common name of scientific method used”
  - PID
  - Name
- ‘Missing Piece’ of PaNOSC model not represented in ICAT model
- Pull Request ready for a future release

Data Publications

- UNDER DISCUSSION
- to be able to store and manage information about data publications within ICAT
- Adds entities for
  - DataPublication
  - FundingReference
    - funderName
    - funderIdentifier
    - awardNumber
    - awardTitle
  - + others

Science and Technology Facilities Council

Other Schema Proposals

UNDER DISCUSSION

- Additional information about users and their affiliation
  - To keep track of affiliation and name at the time of the publication
  - In case it changes over time
  - or maybe separate UserParameters
- Adding images to represent Datasets
  - to display in a GUI
- Roles for InstrumentScientists
  - scientists might play a different role in a beamline: beamline manager, staff, Phd, technician, collaborator, etc...
- And many more…

Science and Technology Facilities Council

APIs and DataGateway

DataGateway API – PaNOSC Search API – DataGateway

Current APIs

APIs – Metadata Access	Ingest	Retrieve
SOAP – **Java** client + **Python** library	x	x
“REST” Http API for querying with JPQL (JSON response)	x	x
ICAT+ Rest API for ESRF DataHub Web Interface		x
OAI-PMH		x
DataGateway API – Rest API for DataGateway Web Interface • basis for a future PaNOSC/ExPaNDS API	Maybe?	x

Science and Technology Facilities Council

DataGateway API

Implementation

- A more RESTful API for the ICAT metadata catalogue for use by DataGateway.
- One endpoint per entity-type (eg. datafile, dataset, investigation)
  - **/datafiles/<id>** GET, PATCH, DELETE - Returns/edits/deletes an entity based on id
- Filters are provided in the query string – based on Loopback (SciCAT)
  - **/datafiles?where={“id”:{“eq”: 1}}**
  - Also limit, order, skip, include (ie. nested objects), distinct filters
- Have not hard-coded a hierarchy – difficult to hardcode for every facility
- Authentication information in HTTP Authorization header

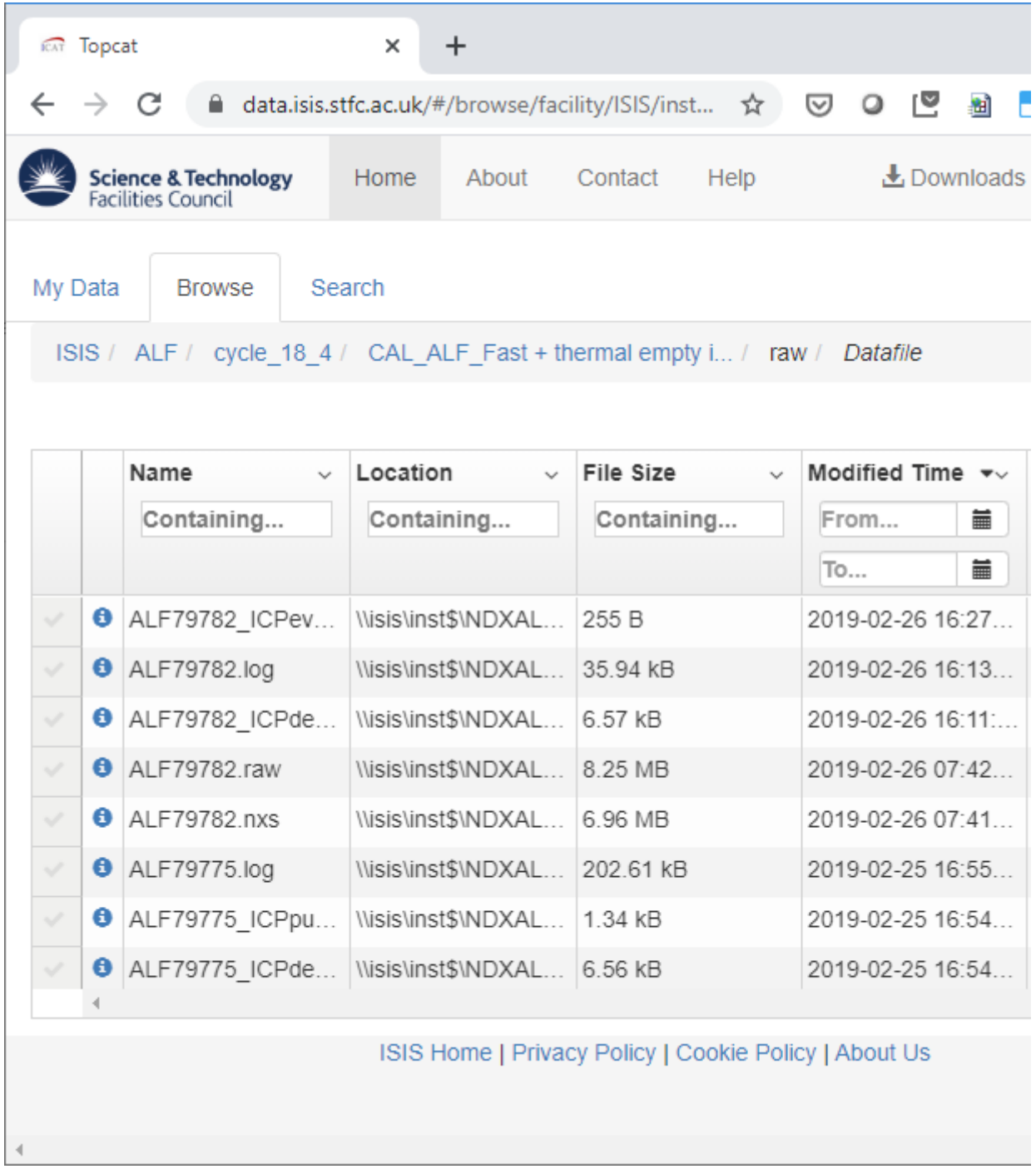
Science and Technology Facilities Council

More detail here

Web Interfaces

Topcat

- In production for 3+ years
- Data hierarchy presented as sequence of table views
- Free text search
- Download carts
  - Direct & deferred (tape)
  - Send to Globus, HPC, etc.
- Written in AngularJS – EOL June 2021

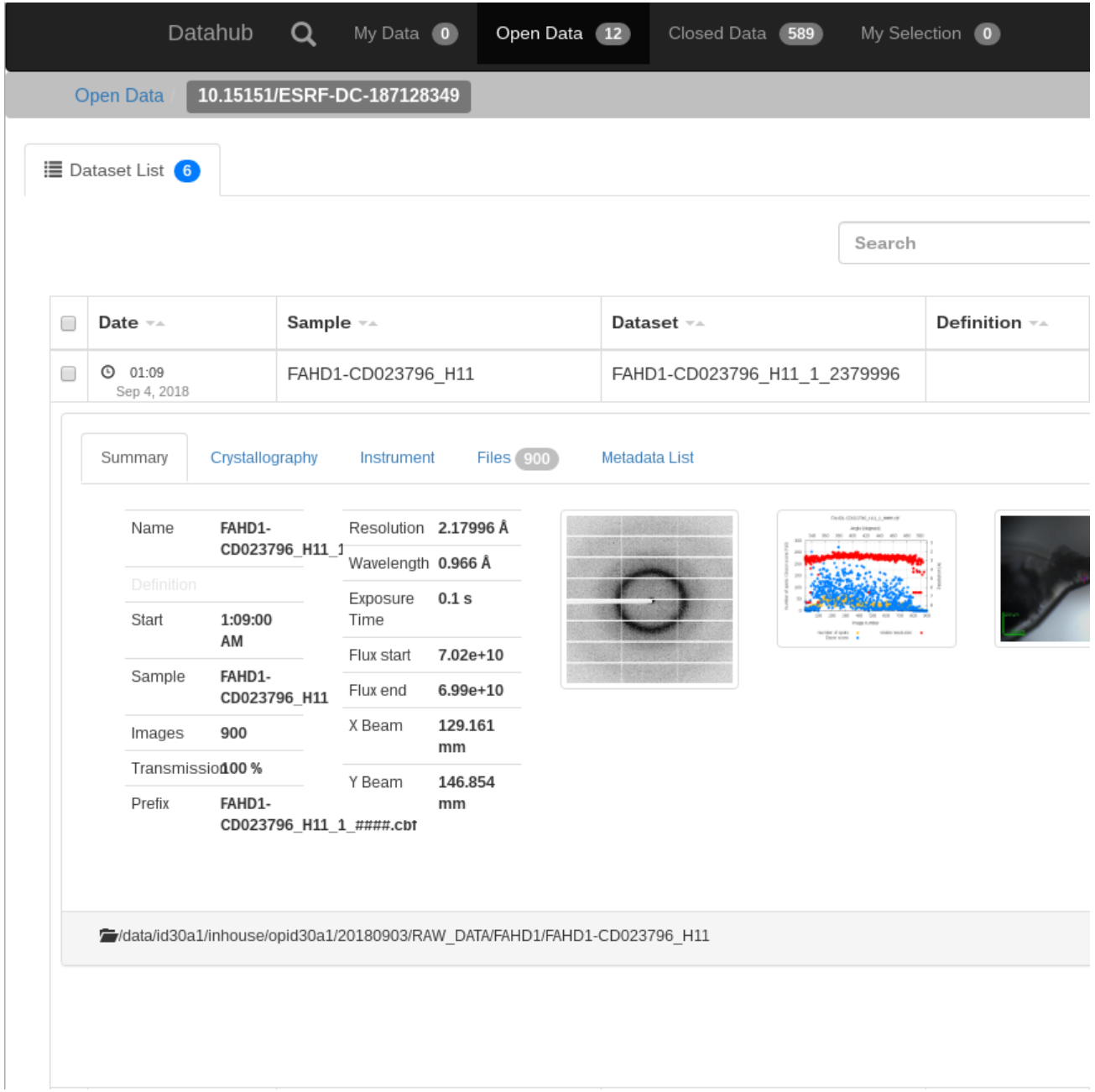


Science and Technology Facilities Council

Web Interfaces

DataHub

- Developed at ESRF using React
- Uses ICAT+ API
- Richer metadata
- Previews/thumbnails
- Access to experiment log books

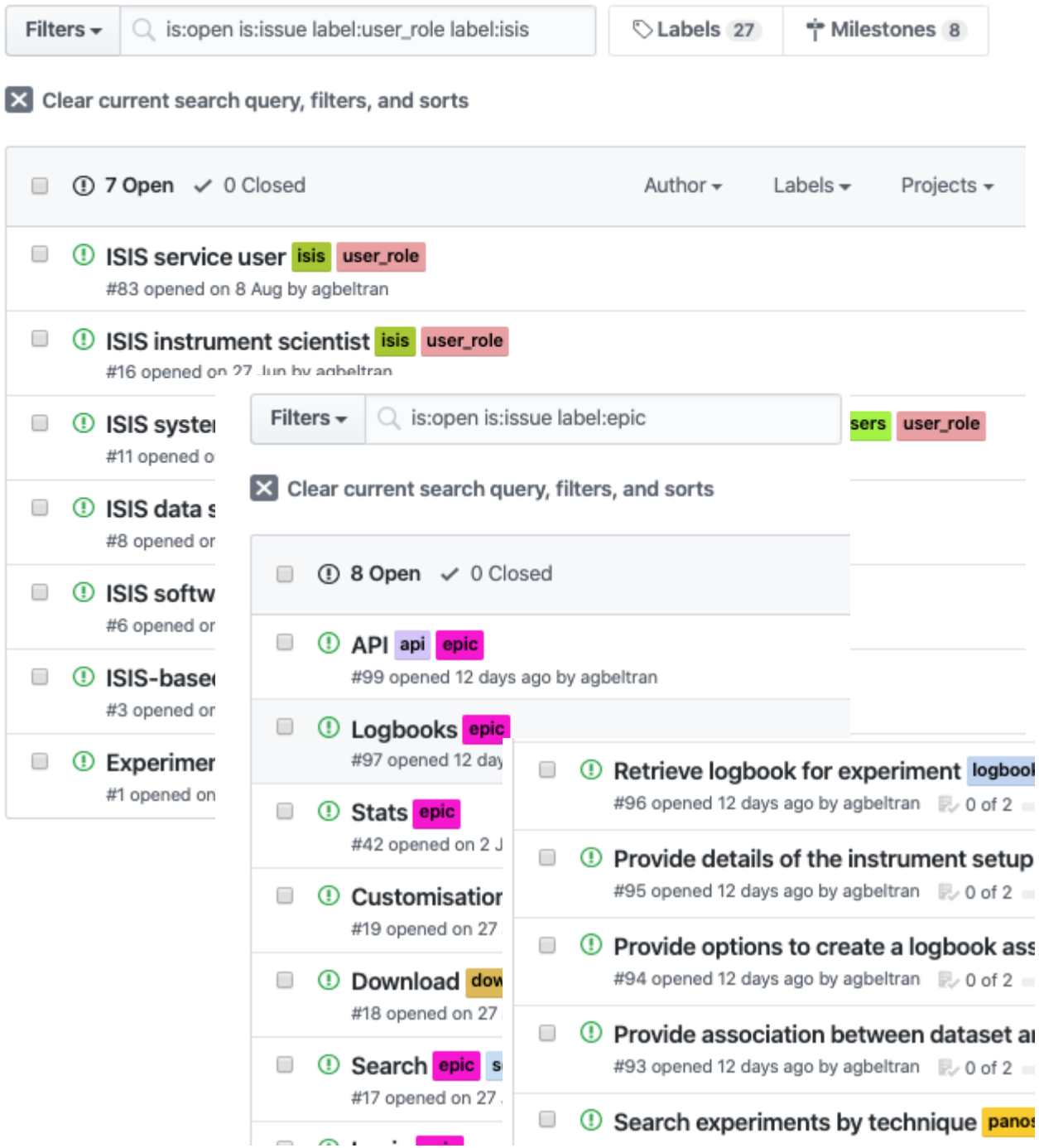


Science and Technology Facilities Council

Web Interfaces

DataGateway

- New interface written in React
- Internal user testing phase
- Aiming for same functionality as Topcat
- Plugs into SciGateway
  - Micro-frontend architecture
  - Enables integration & code-sharing with other STFC projects eg. next eCat, DAaaS



Science and Technology Facilities Council

DataGateway

An interface reflecting the users data journeys

- Considering both proposal users and open data users
- From data creation to data publication
- Data provenance: Associate instrument setup with raw & processed data
- DOI creation & workflows for data publication
- Data discovery & data access
- Rich metadata (moving to FAIR data)
- Specialised data catalogue - information about the data hierarchies in each facility
- Data visualisation/processing/publication + impact analysis
- Different at each ICAT instance

Science and Technology Facilities Council

More detail here

SciGateway CONTACT HELP SIGN IN

Home > Instruments

Results: 43

Data Browser

DataGateway Browse

DataGateway My Data

Data discovery & access

DataGateway Download

DataGateway Search

Toggle Cards

Show Advanced Filters

1 2 3 4 5

Max Results 10

Sort By

Name

Description

Type

URL

SANS2D

Time-of-flight Small-Angle Neutron Scattering instrument. Sans2d can be used to examine size, shape, internal structure and spatial arrangement in nanomaterials, 'soft matter', and colloidal systems, including those of biological origin, on length scales of between* 0.25-300 nm. SANS does not locate individual atoms but rather looks at the larger structures they form. This gives important insights into many everyday materials and biological systems.

Show less

Type: Small-angle scat...

URL: http://www.isis.s...

More Information

SciGateway CONTACT HELP SIGN IN

Home > Instruments > SANS2D > cycle_15_1 > *Investigations*

Results: 32

Toggle Cards

	Title	investigations.vis	RB Number	DOI	Size	Instrument	Start Date	End Date
	Include.	Include.	Include.	Include.		Include.	From... (y To... (y	From... (y To... (y
	Exchange kinetics...	1_SANS2D	1510186	10.5286/ISIS.E.RB...	0 B	SANS2D	2015-07-02 09:00:...	2015-07-05 08:59:...
	Controlling the len...	1_SANS2D	1510236	10.5286/ISIS.E.RB...	1.71 GB	SANS2D	2015-06-29 08:35:...	2015-07-02 08:59:...
	Origin of co-nons...	1_SANS2D	1510077	10.5286/ISIS.E.RB...	955.71 MB	SANS2D	2015-06-09 09:00:...	2015-06-13 07:17:...
	A SANS study of n...	1_SANS2D	1510287	10.5286/ISIS.E.RB...	0 B	SANS2D	2015-06-25 09:00:...	2015-06-26 08:59:...

INVESTIGATION DETAILS | INVESTIGATION USERS | INVESTIGATION SAMPLES | PUBLICATIONS

1510287

TITLE

A SANS study of nano-gels with a temperature tunable bio-interface

INVESTIGATIONS.DETAILS.VISITID

1_SANS2D

Home > Instruments > SANS2D > cycle_15_1 > *Investigations*

Results: 32

1 2 3 4

Max Results
10

Sort By

Title

Summary

investigations.visitId

RB Number

DOI

Instrument

Start Date

End Date

Exchange kinetics of complex coacervates core mic...

Coacervates have been studied for a long time for their unusual properties: they are dense networks of charged polymers that are soluble in water. Instead of flocculating when exposed to high salt concentrations, they dissolve. They are fully permeable to water, yet they are capable of strongly binding heavy metals

Show more

investigations.visitId: 1_SANS2D
RB Number: 1510186
DOI: 10.5286/ISIS.E.R...
Size: 0 B
Instrument: SANS2D
Start Date: 2015-07-02 09:0...
End Date: 2015-07-05 08:5...

ADD TO CART

More Information

Controlling the length of polymeric supramolecular n...

The self-assembly of (cyclic peptide)-polymer conjugates provide a versatile and functional platform for the precise design of polymeric supramolecular nanotubes. Moving the platform into aqueous media has been a recent

investigations.visitId: 1_SANS2D
RB Number: 1510236
DOI: 10.5286/ISIS.E.R...
Size: 1.71 GB

SciGateway CONTACT HELP SIGN IN

Home > Instruments > SANS2D > cycle_15_1 > Polymer-Lipid Nanodisc Self-assem... > DS- > *Datafiles*

Results: 2148

Name	Location	Size	Modified Time	Actions
Include...	Include...	Include...	From... (yyyy-MM-dd) To... (yyyy-MM-dd)	
SANS2D00030574.nxs	\\isis\inst$\NDXSANS2D\Instrument\da...	68.69 MB	2015-07-14 09:34:54.713000+01:00	

DATAFILE DETAILS DATAFILE PARAMETERS

SANS2D00030574.nxs

DESCRIPTION

Blend 1.9 standard behind SF setup SANS

LOCATION

\\isis\inst$\NDXSANS2D\Instrument\data\cycle_15_1\SANS2D00030574.nxs

SANS2D00030574.log	\\isis\inst$\NDXSANS2D\Instrument\da...	43.3 KB	2015-07-14 09:34:54.825000+01:00
SANS2D00030574.RAW	\\isis\inst$\NDXSANS2D\Instrument\da...	0 B	2015-07-14 09:34:54.827000+01:00
SANS2D00030574_ICPdebug.txt	\\isis\inst$\NDXSANS2D\Instrument\da...	11.06 KB	2015-07-14 09:34:54.829000+01:00

ISIS Home | Privacy Policy | About Us

Science and Technology Facilities Council

SciGateway CONTACT HELP SIGN IN

Home > Instruments > SANS2D > cycle_15_1 > Polymer-Lipid Nanodisc Self-assem... > DS- > *Datafiles*

Results: 2148

Name	Location	Size	Modified Time	Actions
Include...	Include...	Include...	From... (yyyy-MM-dd) To... (yyyy-MM-dd)	

DATAFILE DETAILS | DATAFILE PARAMETERS

BCAT_INV_STR

Edler,Arnold,Terry,Tognoloni,Banuelos

SOURCE_FRAMES

9044

RUN_DURATION

907

START_DATE

2015-07-14 09:18:02+0000

DATA_STD_COUNTS

0.455312

HDF_VERSION

Science and Technology Facilities Council

Mapping Facility Entities

From ICAT Schema to OAI-PMH

Data Model

Implementation

- Each facility maps relevant entities from data model onto locally relevant concepts
- **Different at each ICAT instance**

OAI-PMH

- Need to map from ICAT to Dublin Core and Datacite
- **Different at each ICAT instance**

Science and Technology Facilities Council

More detail here

Example Mapping - Diamond

Investigation

ICAT DB	OAI Dublin Core	OAI Datacite
ID	<identifier>	<identifier>
create_id		
create_time	<datestamp>	<datestamp>
DOI		<identifier identifierType="DOI">
endDate	<dc:date>Created:	<date dateType="Collected">
mod_id		
mod_time	? <datestamp>	? <datestamp>
name		
releaseDate	<dc:date>Available:	<publicationYear>
startDate	<dc:date>Created:	<date dateType="Collected">
summary	<dc:description>	<description descriptionType="Abstract">
title	dc:title>	<title>
visit_id		
facility_id		
type_id		

Science and Technology Facilities Council

Thanks to Oliver Copping at Diamond

Science and Technology Facilities Council

Future Plans

OpenID Connect – Improve Search – Cloud

Auth

Authentication

- Simple (username/password), Database, LDAP supported
- Implemented with a **plugin** system
- Most facilities create a plugin for their user-office system

Authorisation

- Rules stored in ICAT database **not** in an external system
- Simple table permissions or more complex SQL ‘where’ clauses
- **Groups**, **Roles** eg. InstrumentScientist
- Generally fixed at deployment time

OpenID Connect

Developed at HZB

- enables users to log in to ICAT via an external OpenID Connect (OIDC) identity provider
  - such as Keycloak.
- doesn't check the user's credentials by itself
- leaves this part to the identity provider (IdP) and relies on a so-called token to actually authenticate the user

Free Text Search

ICAT Lucene

- Current implementation uses Lucene directly
  - limited to 2^{32} datafiles
- Diamond (ExPaNDS) has exceeded this
  - Only Dataset, Visit, Proposal text search until it can be rewritten
- Hope to develop a new search component
- Will look at ElasticSearch

Cloud

Docker

- Some work done already
  - HZB
  - Ceric
- Would be great if each component had its own dockerfile

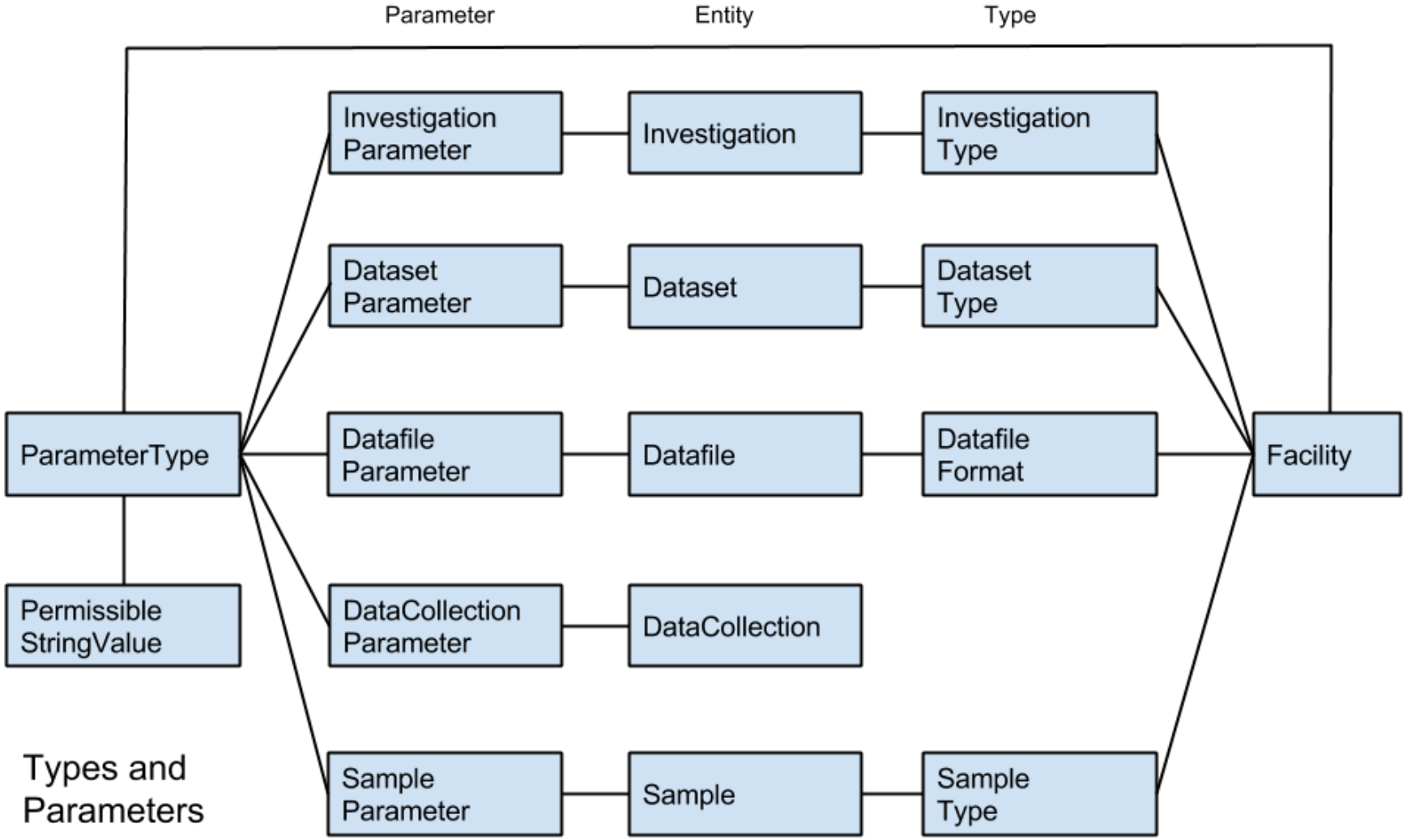
Kubernetes?

- What would a cloud-native ICAT look like?
- What are the risks?
- What are the benefits?

Science and Technology Facilities Council

Questions?

Data Model - Parameters



Types and Parameters

Science and Technology Facilities Council

ICAT Evaluation Conclusions

- For “simple” queries returning small amounts of unordered data, performance improves slightly as more data is added. This is not fully understood but may be due to Oracle caching or improving execution plans.
- For longer running queries requiring ordering of a large number of rows, performance does not degrade significantly as more data is added.
- In both cases the change was only a few percent per year of data added, which over the next 5-10 years should not be a concern.
- The unexplained rises, falls and differences between tests do not appear to be significant and are most likely due to other load on the VM cluster and/or the network at the time the test was run.

From a presentation by Kevin Phipps - STFC

Governance & Resources

Project Structure

- Steering Committee
- Mailing list
- Open Source on Github – http://github.com/icatproject
  - Apache 2 Licence
  - Github issues and pull requests
  - Additional contributions at http://github.com/icatproject-contrib
- Monthly meetings via video conference
- Face to Face meetings – Grenoble,10th/11th March 2020