



Science and
Technology
Facilities Council

ICAT Update

PaNOSC WP3 Catalogue Integration
Best Practices Meeting
May 2021

Stuart Pullinger
The ICAT Collaboration

ICAT Collaboration

STFC

- ISIS (ExPaNDS)
 - CLF
- Diamond (ExPaNDS)**

- **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 Collaboration

HZB
(ExPaNDS)

CERIC
(ExPaNDS)

ALBA
(ExPaNDS)

ESRF
(PaNOSC)

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



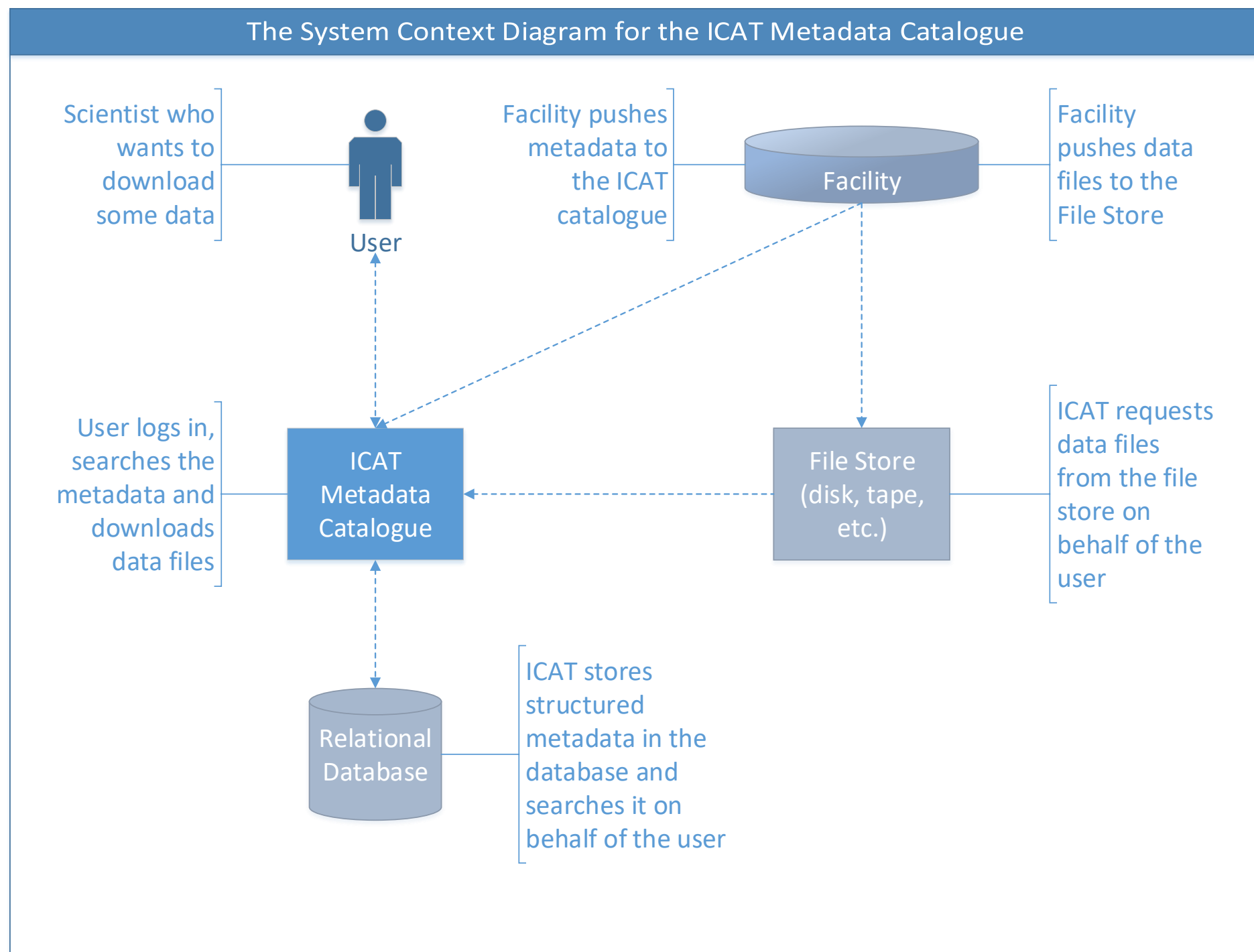
Science and
Technology
Facilities Council

ICAT Overview

Architecture – Components – Data Model

ICAT

Architecture



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

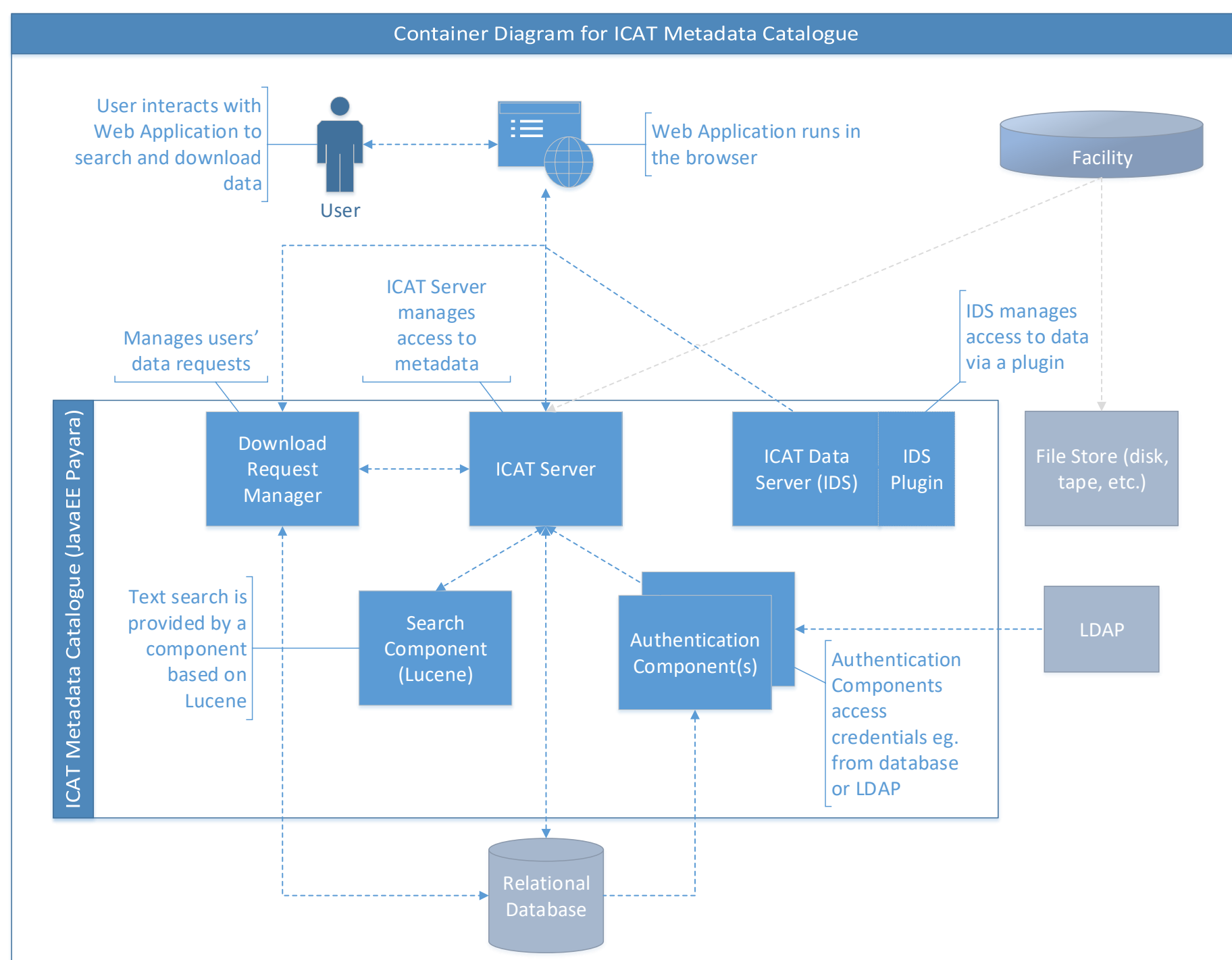
* Java Persistence API



Science and
Technology
Facilities Council

ICAT

Components



ICAT Data Model

Based on Core Scientific Metadata Model:

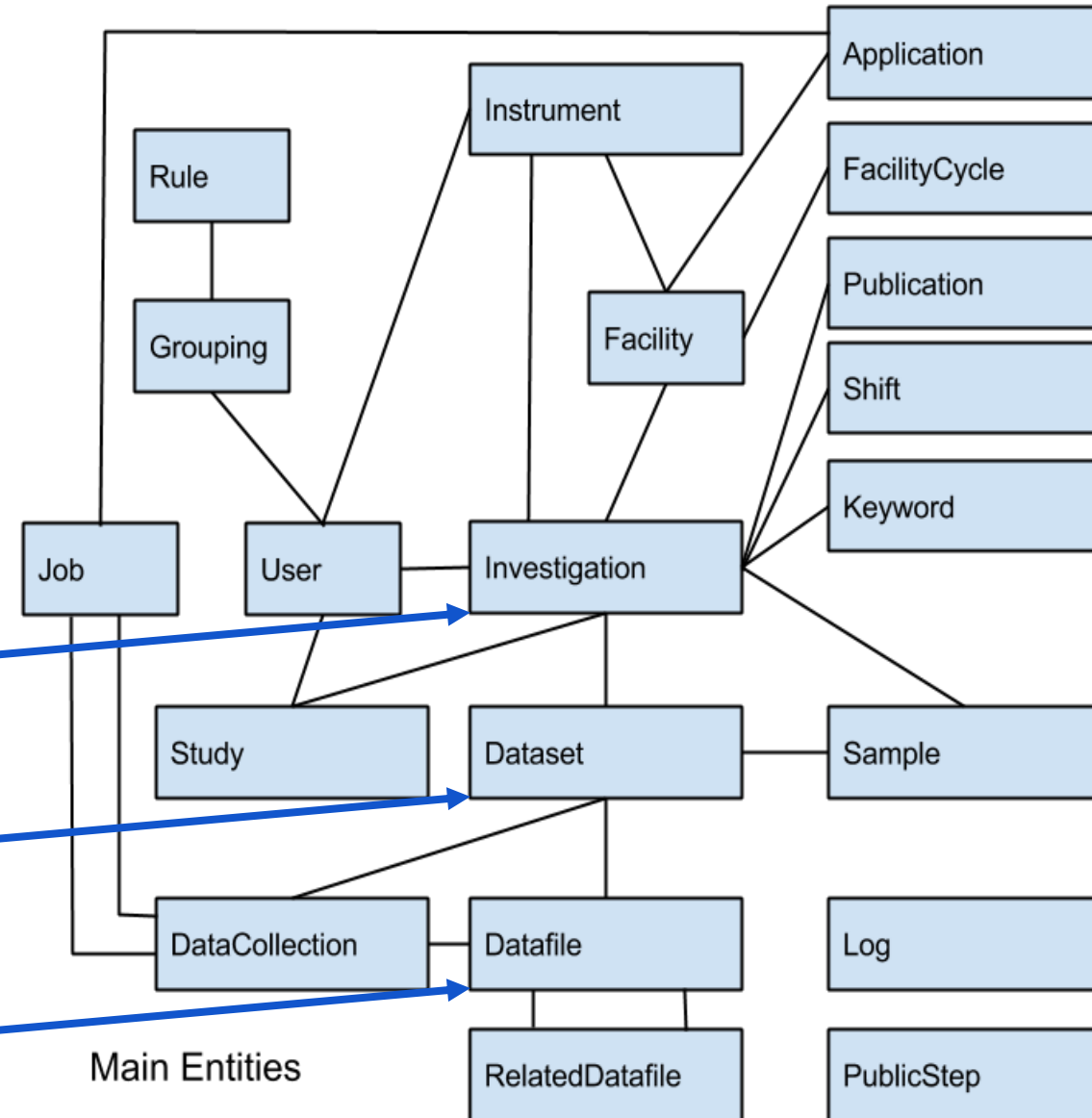
<http://icatproject-contrib.github.io/CSMD/>

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

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



DataGateway API

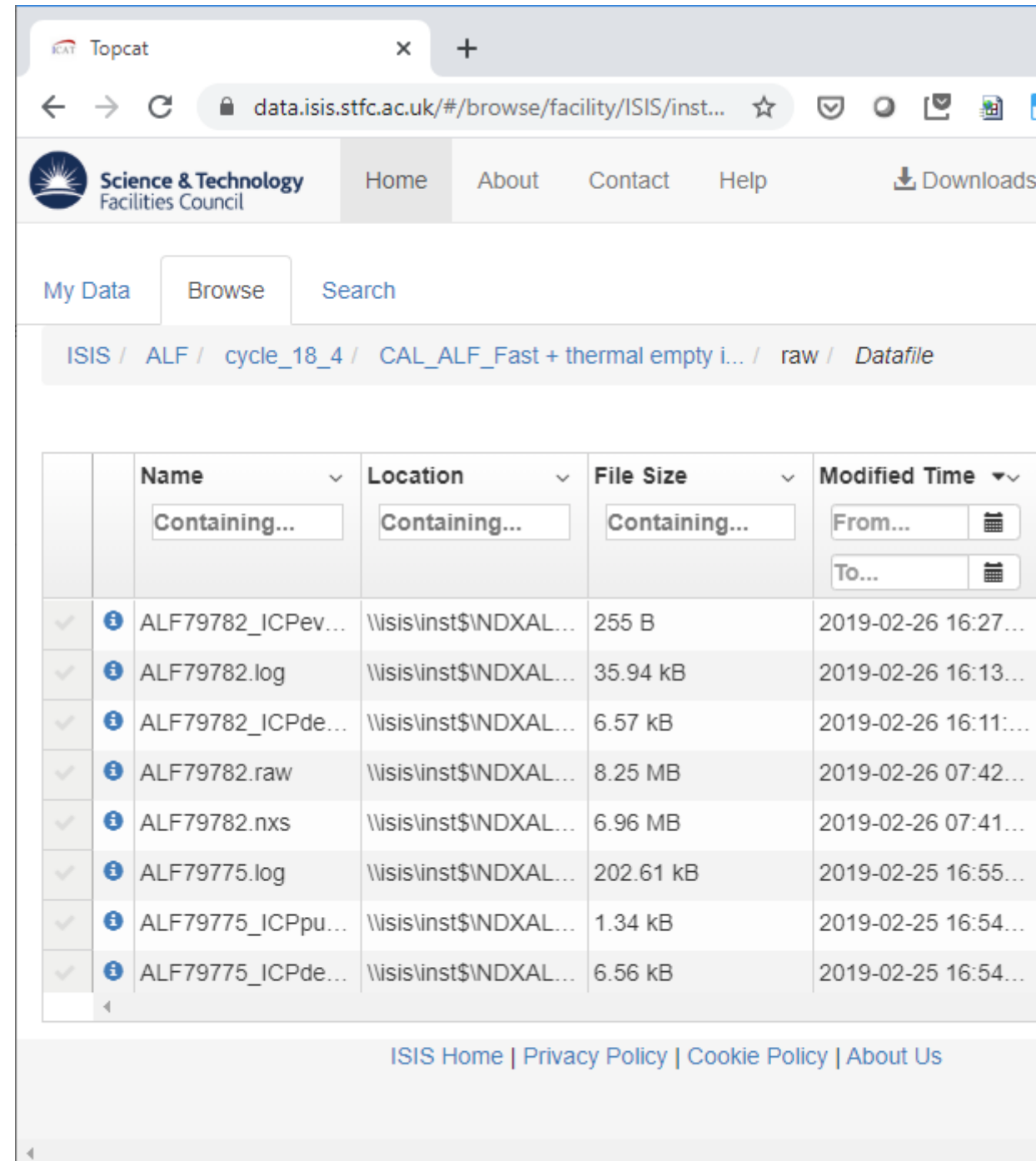
Implementation

- 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

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



The screenshot displays the Topcat web interface for the ISIS facility. The browser address bar shows the URL `data.isis.stfc.ac.uk/#/browse/facility/ISIS/inst...`. The page header includes the Science & Technology Facilities Council logo and navigation links: Home, About, Contact, Help, and Downloads. Below the header, there are tabs for My Data, Browse (selected), and Search. The breadcrumb trail indicates the current location: ISIS / ALF / cycle_18_4 / CAL_ALF_Fast + thermal empty i... / raw / Datafile.

	Name	Location	File Size	Modified Time
	Containing...	Containing...	Containing...	From... To...
✓	ALF79782_ICPev...	\\isis\inst\$INDXAL...	255 B	2019-02-26 16:27...
✓	ALF79782.log	\\isis\inst\$INDXAL...	35.94 kB	2019-02-26 16:13...
✓	ALF79782_ICPde...	\\isis\inst\$INDXAL...	6.57 kB	2019-02-26 16:11...
✓	ALF79782.raw	\\isis\inst\$INDXAL...	8.25 MB	2019-02-26 07:42...
✓	ALF79782.nxs	\\isis\inst\$INDXAL...	6.96 MB	2019-02-26 07:41...
✓	ALF79775.log	\\isis\inst\$INDXAL...	202.61 kB	2019-02-25 16:55...
✓	ALF79775_ICPpu...	\\isis\inst\$INDXAL...	1.34 kB	2019-02-25 16:54...
✓	ALF79775_ICPde...	\\isis\inst\$INDXAL...	6.56 kB	2019-02-25 16:54...

The footer contains links for ISIS Home, Privacy Policy, Cookie Policy, and About Us.

Web Interfaces

DataHub

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

Datahub

My Data 0

Open Data 12

Closed Data 589

My Selection 0

Open Data / 10.15151/ESRF-DC-187128349

Dataset List 6

Search

	Date ▾▴	Sample ▾▴	Dataset ▾▴	Definition ▾▴
<input type="checkbox"/>	🕒 01:09 Sep 4, 2018	FAHD1-CD023796_H11	FAHD1-CD023796_H11_1_2379996	

Summary

Crystallography

Instrument

Files 900

Metadata List

Name

FAHD1-CD023796_H11_1

Definition

Start

1:09:00 AM

Sample

FAHD1-CD023796_H11

Images

900

Transmission

100 %

Prefix

FAHD1-CD023796_H11_1_####.cbf

Resolution

2.17996 Å

Wavelength

0.966 Å

Exposure Time

0.1 s

Flux start

7.02e+10

Flux end

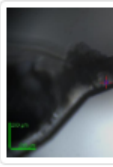
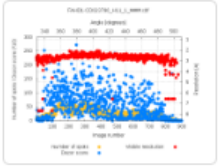
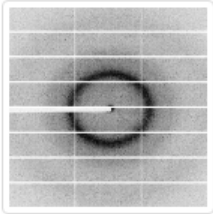
6.99e+10

X Beam

129.161 mm

Y Beam

146.854 mm

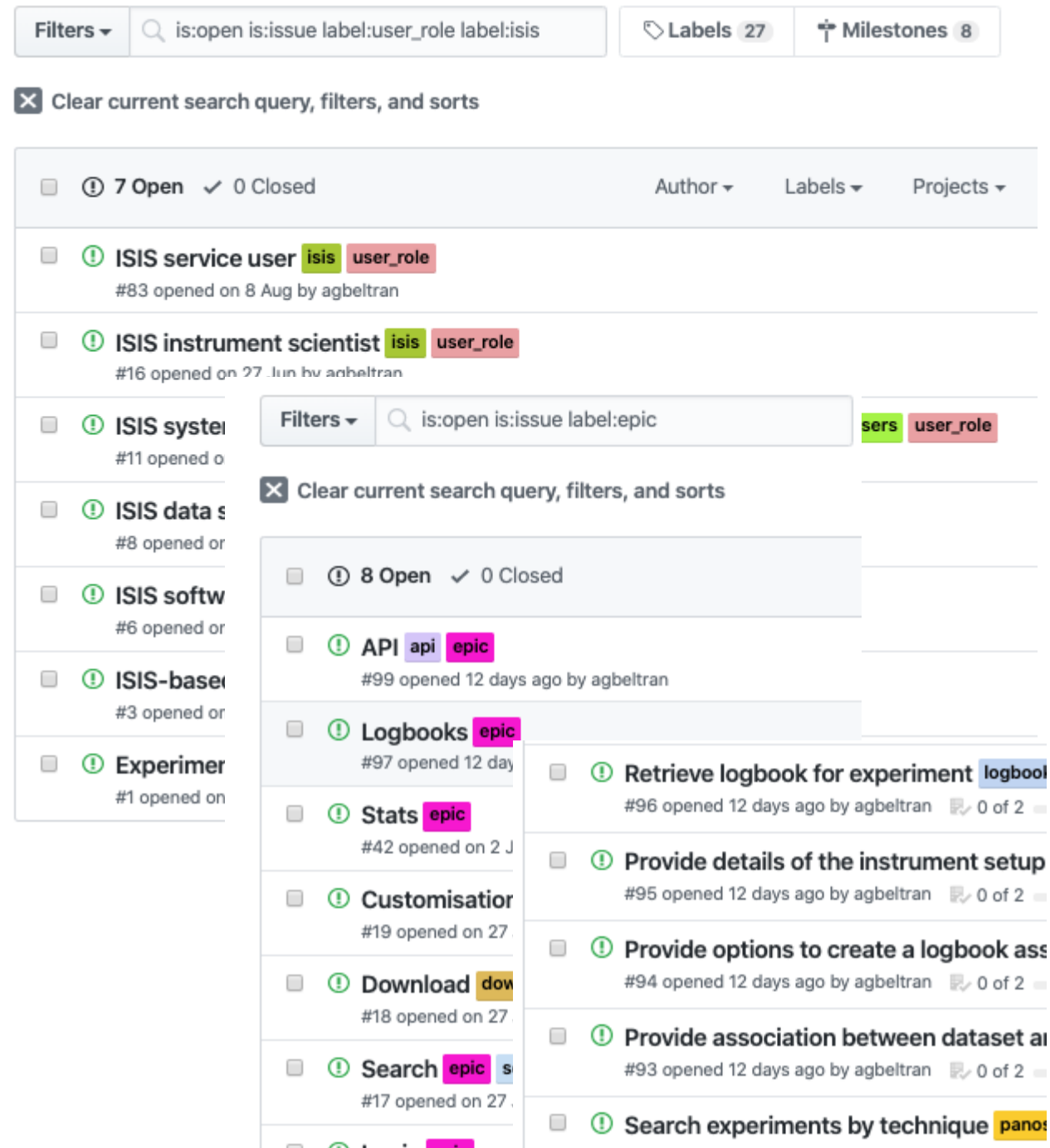


/data/id30a1/inhouse/opid30a1/20180903/RAW_DATA/FAHD1/FAHD1-CD023796_H11

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



The screenshot displays the DataGateway web interface, which is a project management tool. It features a sidebar on the left with a list of issues, each with a checkbox, a status icon (green exclamation mark), a title, and a label. The main area shows a detailed view of a selected issue, including its title, status, and a list of related issues. The interface includes search bars, filters, and a clear button to reset the search query, filters, and sorts. The sidebar issues are:

- ISIS service user (isis, user_role) #83 opened on 8 Aug by agbeltran
- ISIS instrument scientist (isis, user_role) #16 opened on 27 Jun by agbeltran
- ISIS system (isis, user_role) #11 opened on 27 Jun by agbeltran
- ISIS data science (isis, user_role) #8 opened on 27 Jun by agbeltran
- ISIS software (isis, user_role) #6 opened on 27 Jun by agbeltran
- ISIS-based (isis, user_role) #3 opened on 27 Jun by agbeltran
- Experimenter (isis, user_role) #1 opened on 27 Jun by agbeltran

The main area shows a detailed view of a selected issue, including its title, status, and a list of related issues. The interface includes search bars, filters, and a clear button to reset the search query, filters, and sorts. The main area issues are:

- API (api, epic) #99 opened 12 days ago by agbeltran
- Logbooks (epic) #97 opened 12 days ago by agbeltran
- Stats (epic) #42 opened on 27 Jun by agbeltran
- Customisation (epic) #19 opened on 27 Jun by agbeltran
- Download (dow) #18 opened on 27 Jun by agbeltran
- Search (epic, s) #17 opened on 27 Jun by agbeltran

The interface also includes a sidebar with filters and a clear button to reset the search query, filters, and sorts. The sidebar filters are:

- Filters
- is:open is:issue label:user_role label:isis
- Labels 27
- Milestones 8

The main area also includes a sidebar with filters and a clear button to reset the search query, filters, and sorts. The main area filters are:

- Filters
- is:open is:issue label:epic
- sers user_role

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

Data Browser

 DataGateway Browse DataGateway My Data

Data discovery & access

 DataGateway Download DataGateway SearchToggle 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](#)**T** Type: Small-angle scat...**🔗** URL: <http://www.isis.s...>

More Information





Results: 32

Toggle Cards ☐

	Title	investigations.vis	RB Number	DOI	Size	Instrument	Start Date	End Date
<input type="checkbox"/>	<u>Include.</u>	<u>Include.</u>	<u>Include.</u>	<u>Include.</u>		<u>Include.</u>	<u>From...</u> (y) <u>To...</u> (y)	<u>From...</u> (y) <u>To...</u> (y)
<input type="checkbox"/>	Exchange kinetics...	1_SANS2D	1510186	10.5286/ISIS.E.RB...	0 B	SANS2D	2015-07-02 09:00:...	2015-07-05 08:59:...
<input type="checkbox"/>	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:...
<input type="checkbox"/>	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:...
<input type="checkbox"/>	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

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 accomplishment that has unlocked many exciting applications, but to

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

	T Name		Location		Size		Modified Time		Actions
<input type="checkbox"/>	Include...		Include...		Include...		From... (yyyy-MM-dd) 		
							To... (yyyy-MM-dd) 		
<input type="checkbox"/>	^ 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

<input type="checkbox"/>	✓	SANS2D00030574.log	\\isis\inst\$\NDXSANS2D\Instrument\da...	43.3 KB	2015-07-14 09:34:54.825000+01:00	
<input type="checkbox"/>	✓	SANS2D00030574.RAW	\\isis\inst\$\NDXSANS2D\Instrument\da...	0 B	2015-07-14 09:34:54.827000+01:00	
<input type="checkbox"/>	✓	SANS2D00030574_ICPdebug.txt	\\isis\inst\$\NDXSANS2D\Instrument\da...	11.06 KB	2015-07-14 09:34:54.829000+01:00	



	Name	Location	Size	Modified Time	Actions
<input type="checkbox"/>	<input type="text" value="Include..."/>	<input type="text" value="Include..."/>	<input type="text" value="Include..."/>	<div>From... (yyyy-MM-dd) </div> <div>To... (yyyy-MM-dd) </div>	

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**

Example Mapping - Diamond

Investigation

ICAT DB	OAI Dublin Core	OAI Datacite
ID	<identifier>	<identifier>
create_id		
create_time	<timestamp>	<timestamp>
DOI		<identifier identifierType="DOI">
endDate	<dc:date>Created:	<date dateType="Collected">
mod_id		
mod_time	? <timestamp>	? <timestamp>
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?

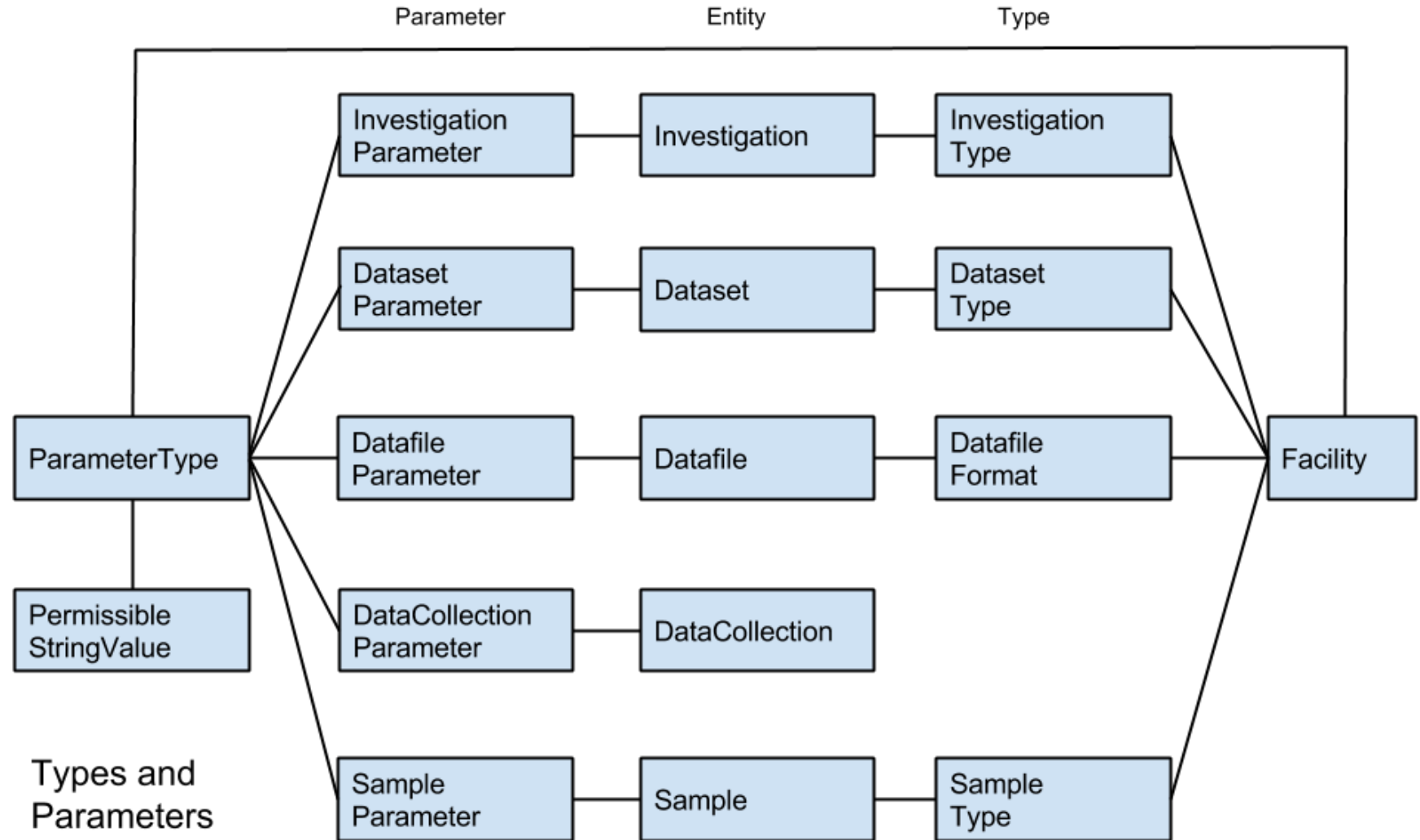


Science and
Technology
Facilities Council

The background features a large blue rectangle on the right side, with a blue triangle pointing downwards at its bottom-left corner. To the left of this rectangle, there are several thin, parallel blue lines of varying lengths, some pointing right and some pointing left, creating a sense of motion or a stylized 'Z' shape.

Questions?

Data Model - Parameters



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.

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

