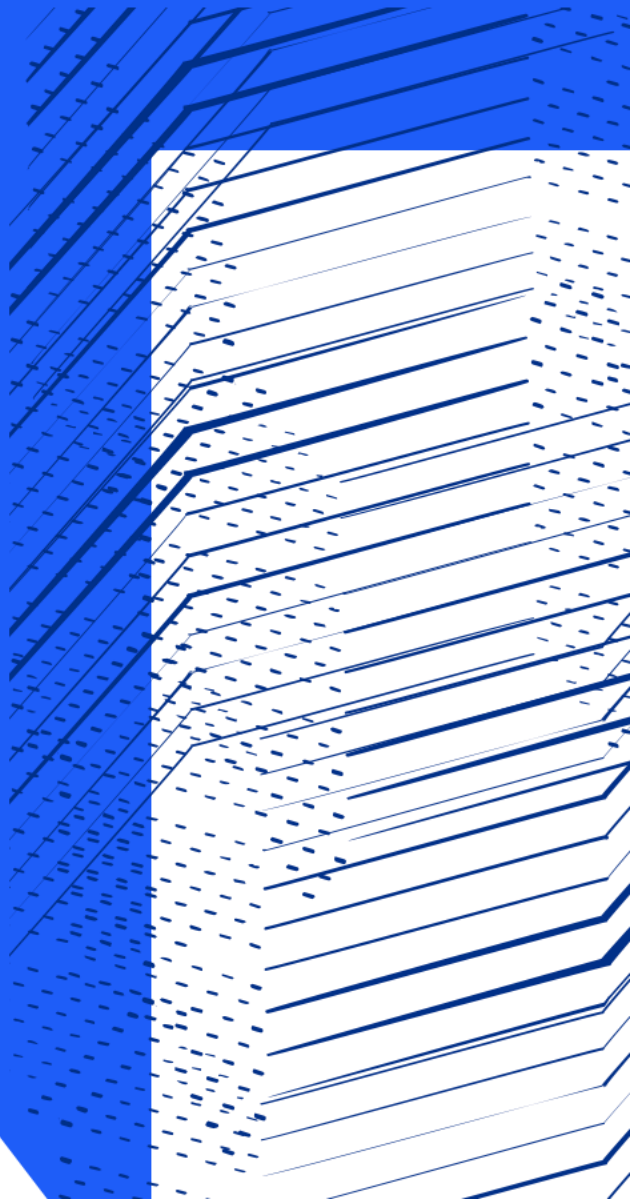
Science and Technology Facilities Council

ICAT Free Text Search

Current functionality, new features, possible approaches, frontend changes



Topics

1 Current state

How does free text search with Lucene work currently in ICAT?

2 User stories and features

What do users need, and how can improvements to the free text search provide it?

3 Engines overview

What alternatives exist, and how do they work?

4 Engines implementation

How exactly will we implement these changes into the stack?

5 Frontend changes

How will new functionality be exposed in DataGateway Search?

Science and Technology Facilities Council

Image © STFC Alan Ford

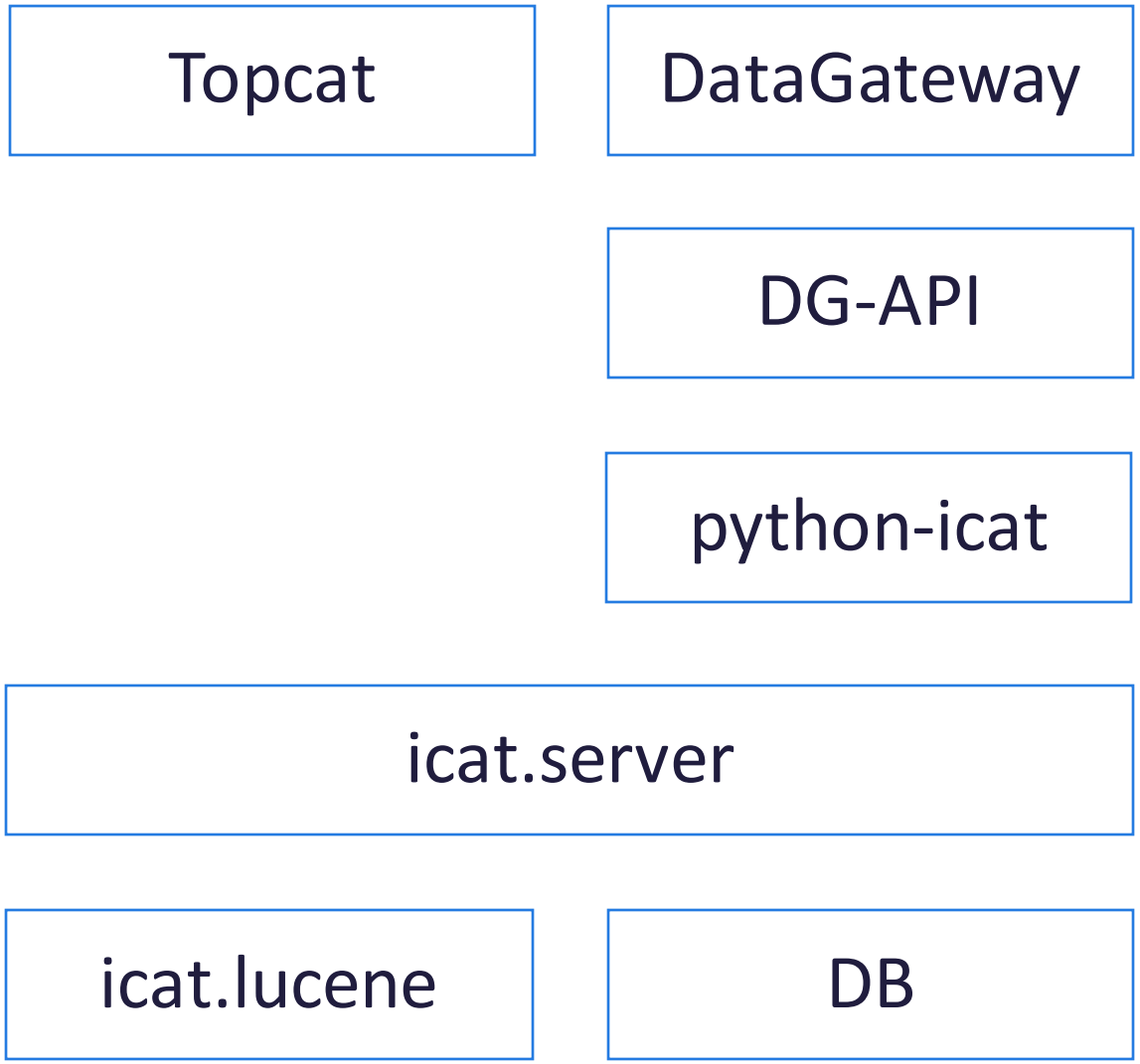
Science and Technology Facilities Council

Current state

How does free text search with Lucene work currently in ICAT?

Science and Technology Facilities Council

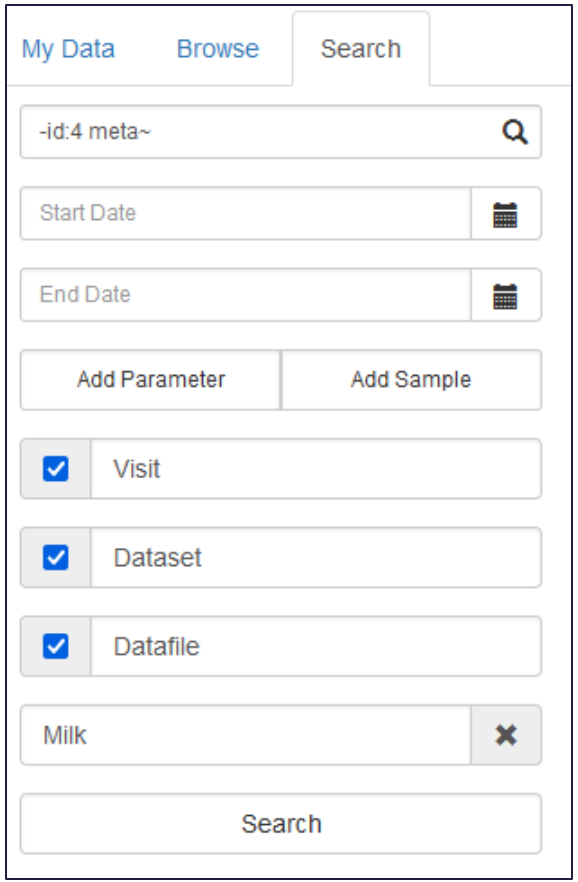
Current Situation

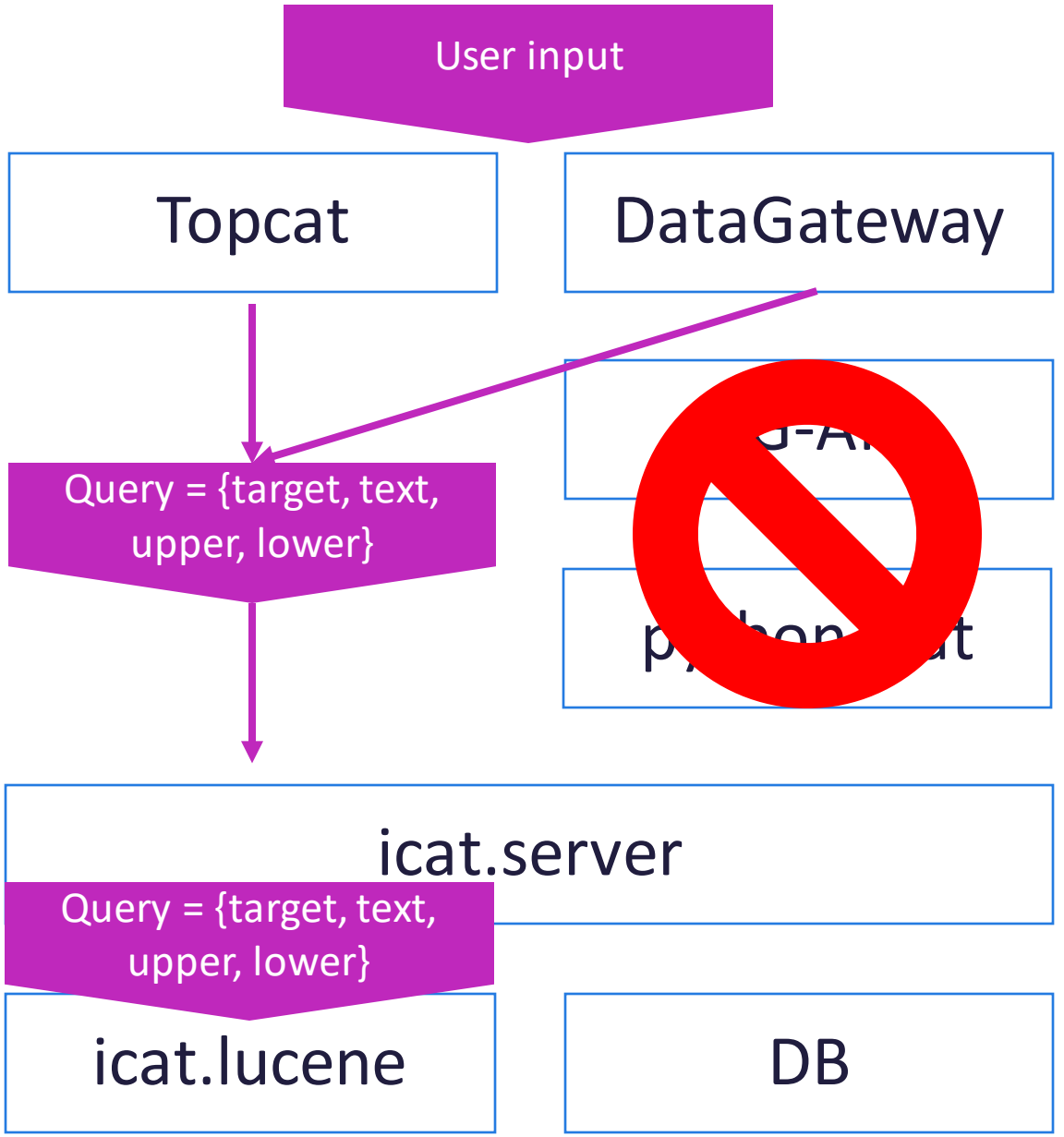


Science and Technology Facilities Council

Current Situation

- User enters free text and/or dates into either application
- Calls to Lucene go direct to the icat.server

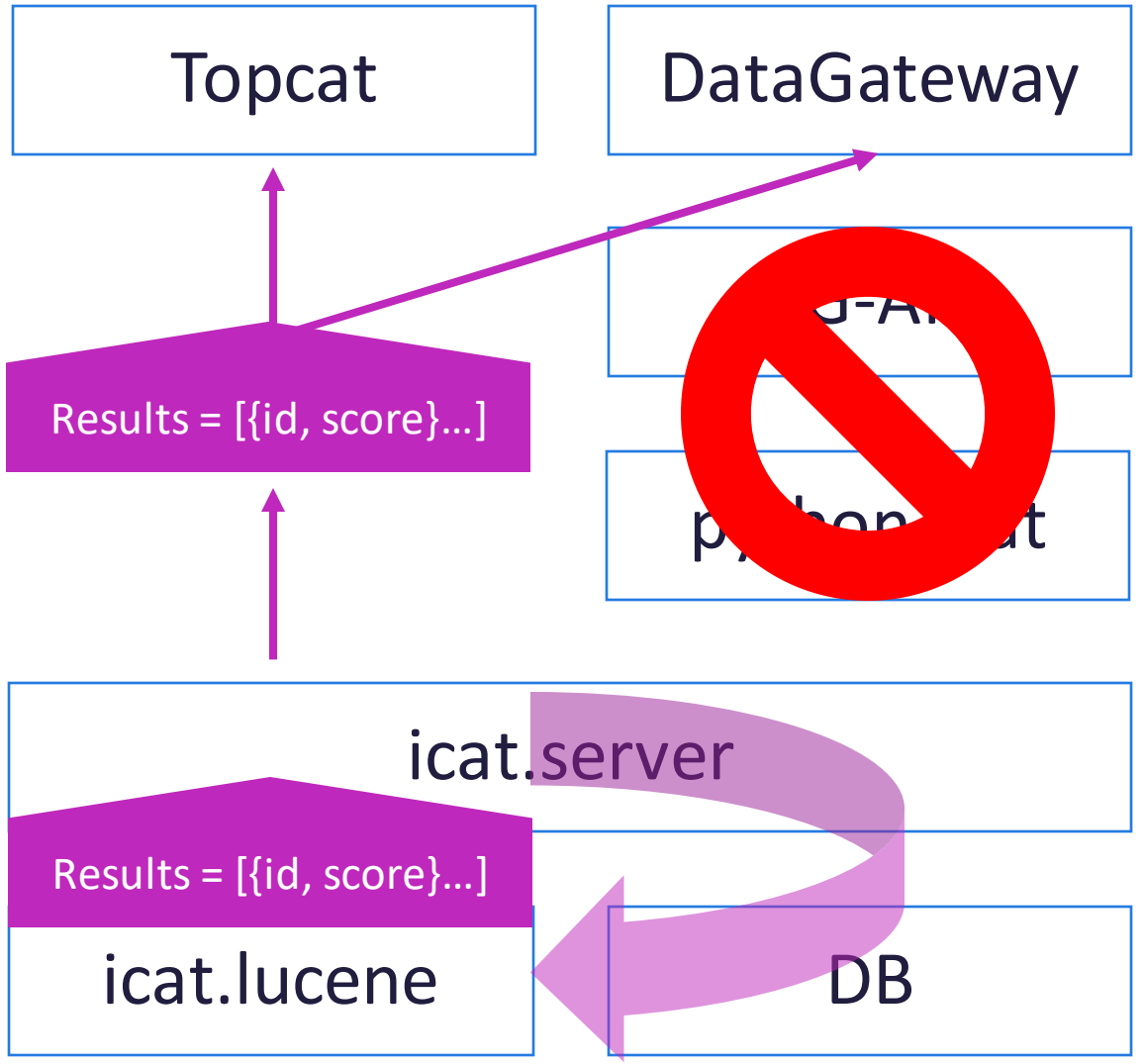




Science and Technology Facilities Council

Current Situation

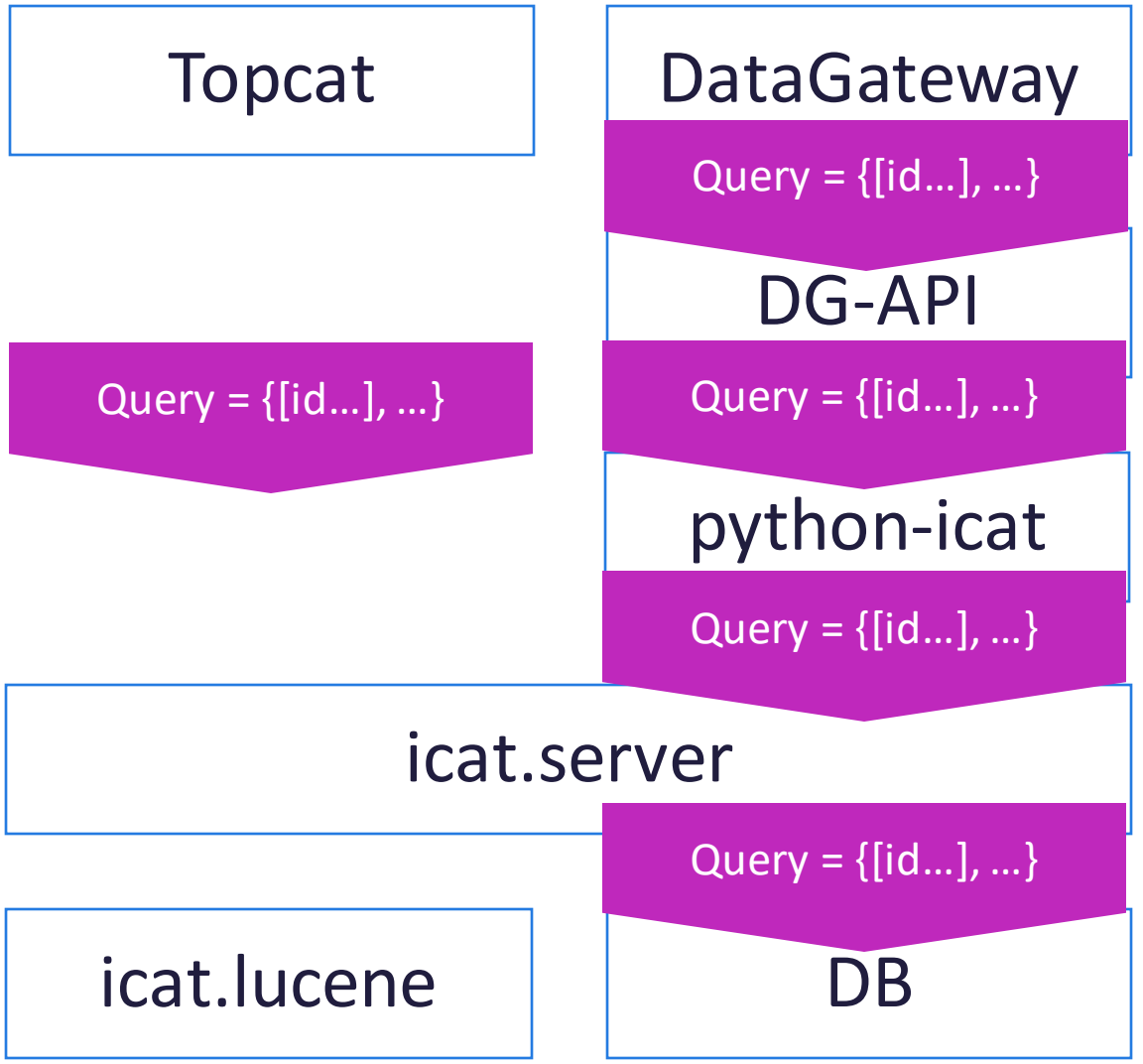
- User enters free text and/or dates into either application
- Calls to Lucene go direct to the icat.server
- icat.server calls icat.lucene
  - icat.lucene has no concept of rules
  - icat.server evaluates rules on IDs
  - Calls icat.lucene again if needed



Science and Technology Facilities Council

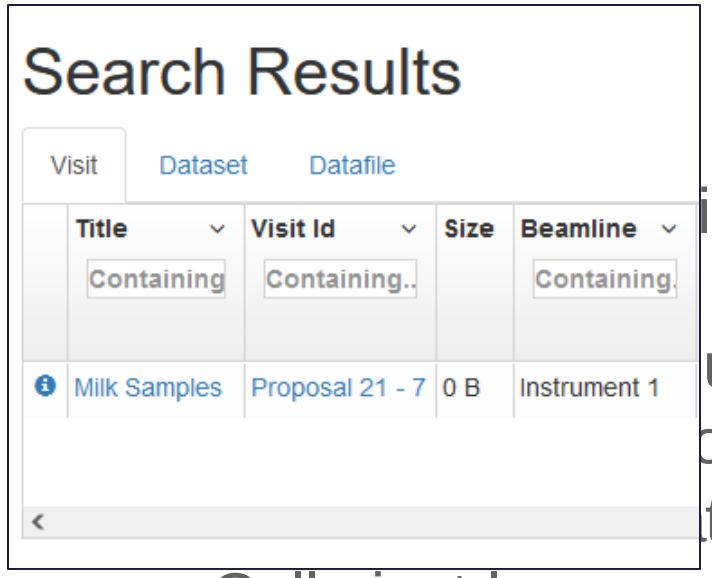
Current Situation

- User enters free text and/or dates into either application
- Calls to Lucene go direct to the icat.server
- icat.server calls icat.lucene
  - icat.lucene has no concept of rules
  - icat.server evaluates rules on IDs
  - Calls icat.lucene again if needed
- IDs and score are used as part of a new query which goes to the DB



Science and Technology Facilities Council

Current Situation

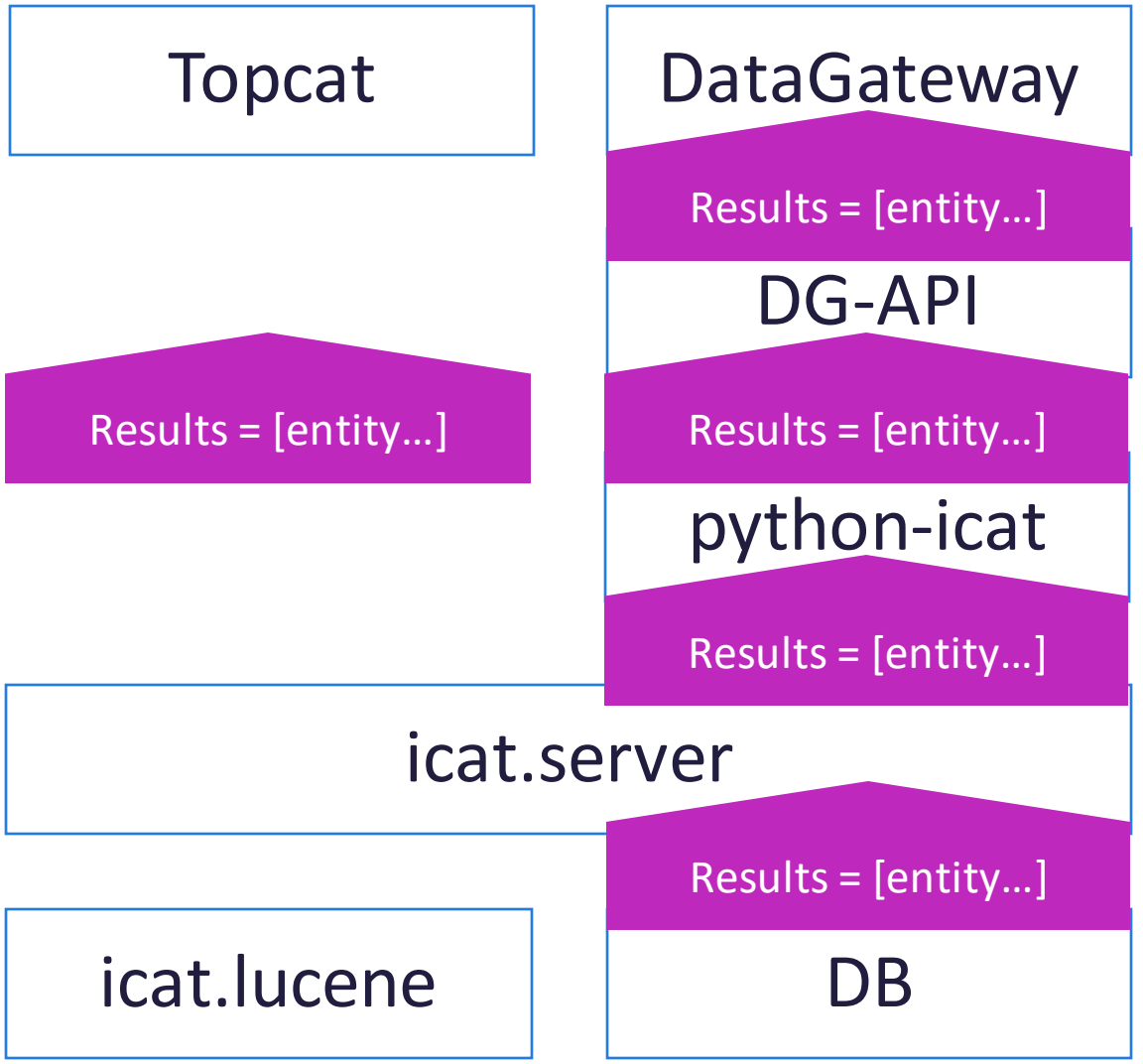


and/or dates into

irect to the

ucene
o concept of rules
tes rules on IDs

  - Calls icat.lucene again if needed
- IDs and score are used as part of a new query which goes to the DB
- Finally, actual entities are returned



Science and Technology Facilities Council

Science and Technology Facilities Council

User stories and features

What do users need, and how can improvements to the free text search provide it?

Based on user stories collected by ExPaNDS for WP2/3, task 3.2, December 2019

Science and Technology Facilities Council

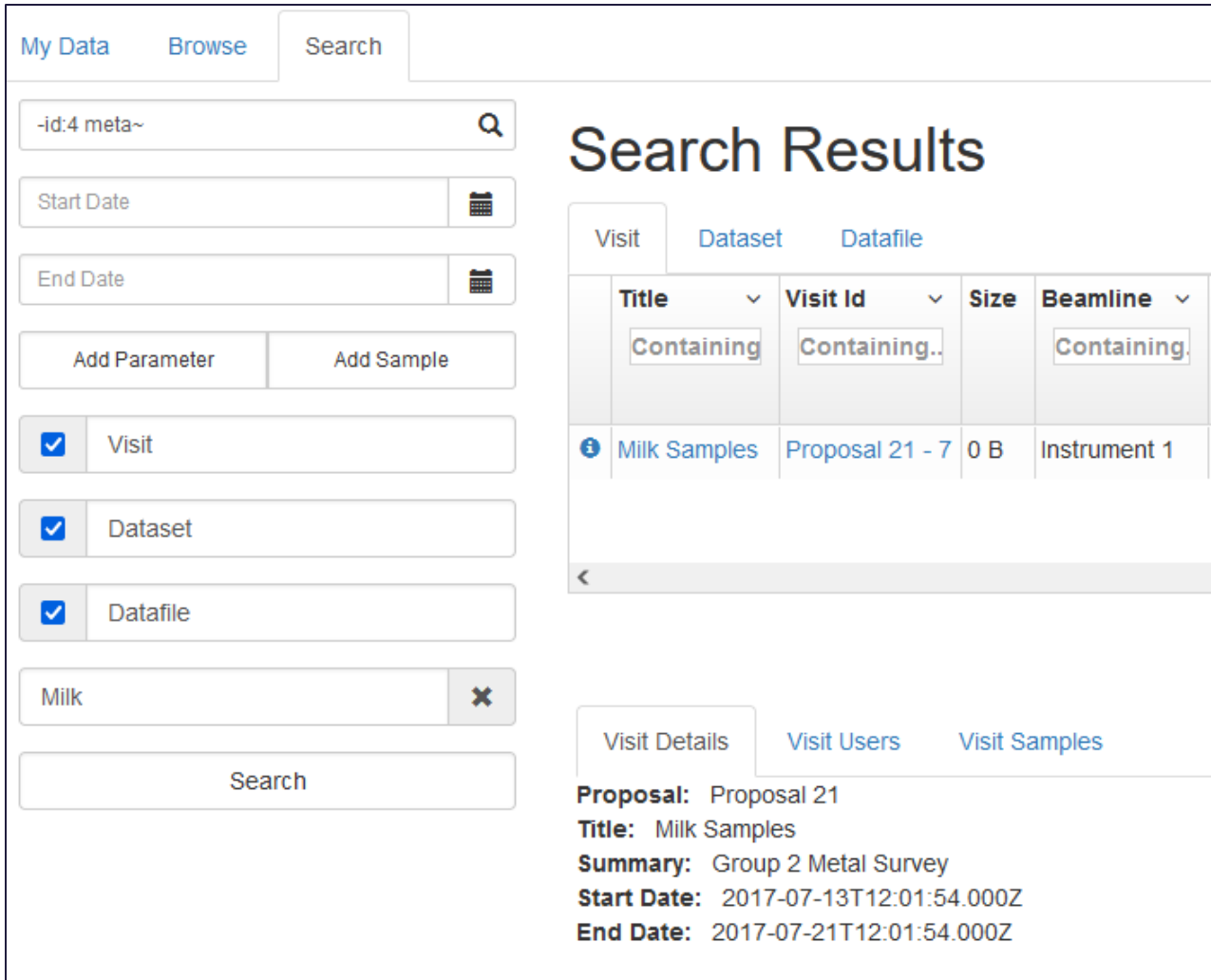
Functionality: Entities

User's search: ExperimentType, PhotonEnergy, SampleTemperature
User's result: Suitable beamlines

- What documents to index?
- What fields to index?

- What documents to search?
- What fields to search?

- What to return?



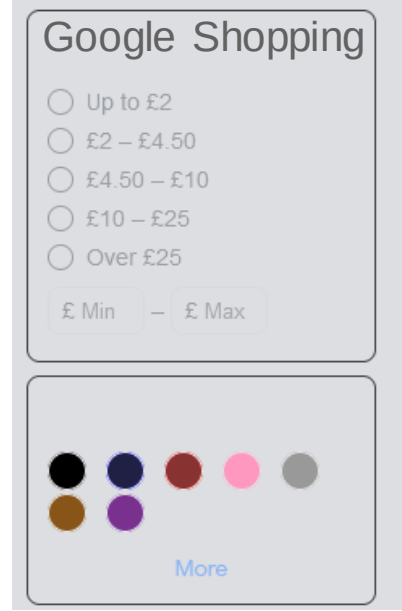
Science and Technology Facilities Council

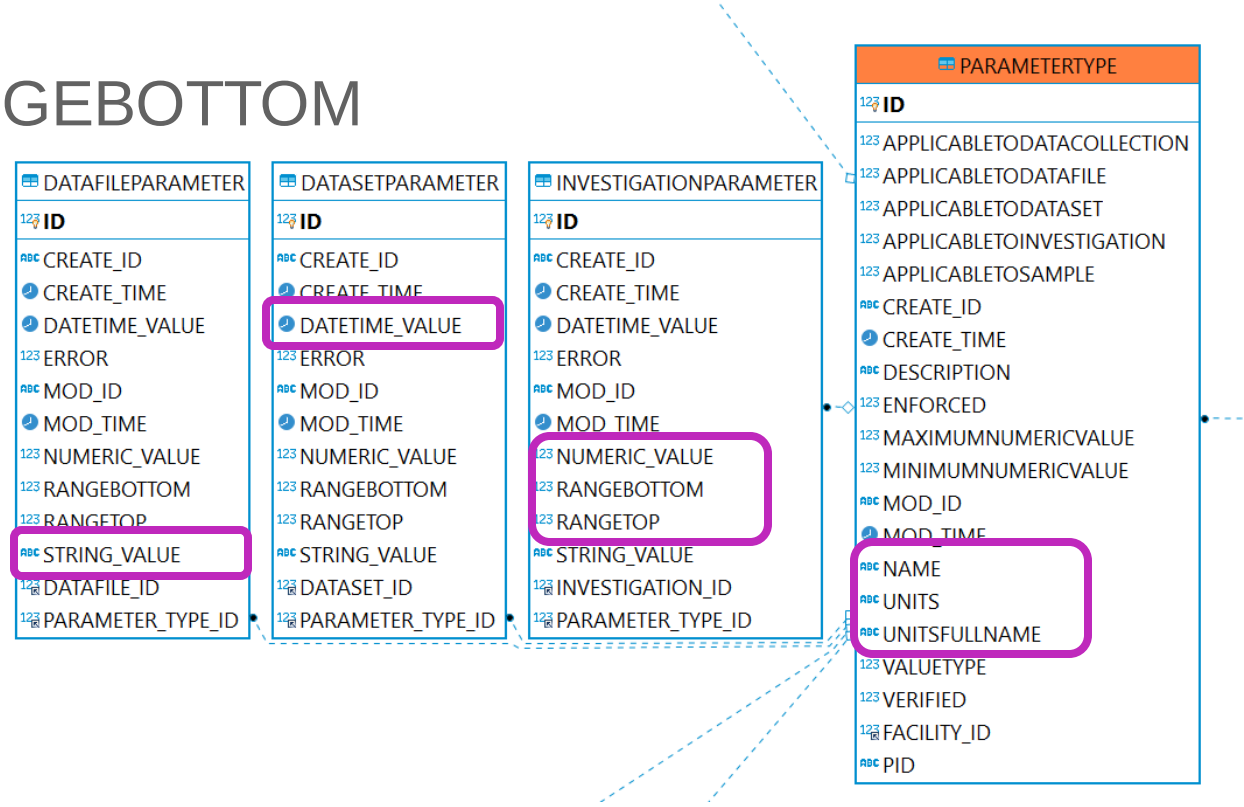
Functionality: Facets

User's search: SampleName: CeO2, DateCollected: 01/18 - 04/19, Temperature: 80degC

- STRING_VALUE
- DATETIME_VALUE
- NUMERIC_VALUE / RANGETOP / RANGEBOTTOM
- NAME
- UNITS
- UNITSFULLNAME

Ideally, turn all this into NAME: VALUE



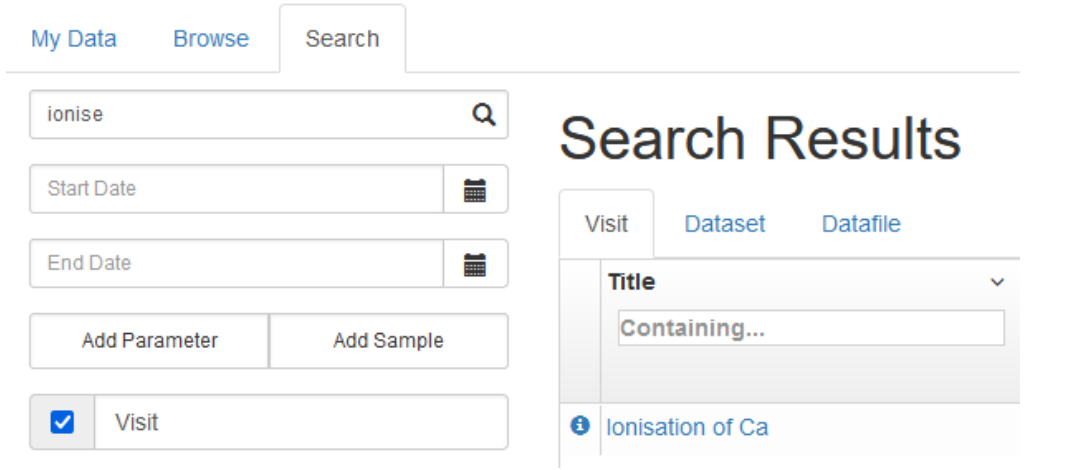


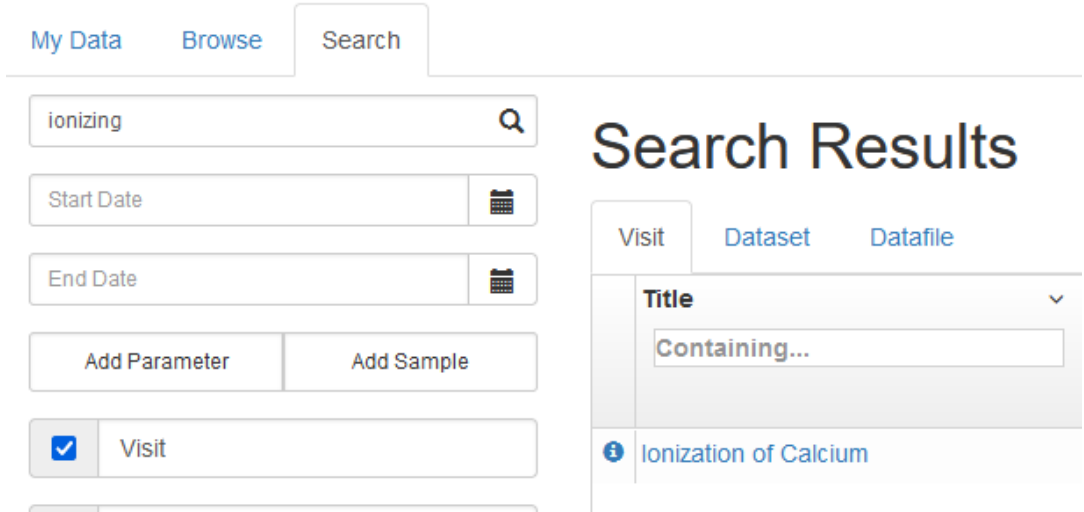
Science and Technology Facilities Council

Functionality: Synonyms

User's search: SampleChemicalFormula: contains Ba AND Fe AND O
SpaceGroup: $Pm\bar{3}m$

- By default, the search has no scientific understanding of terms
- Can provide this with synonyms, to say two terms have equivalent meaning
- Can also modify stop words, e.g. to allow At, As, Be, In, No to be searchable
- Has potential, but would need configuration





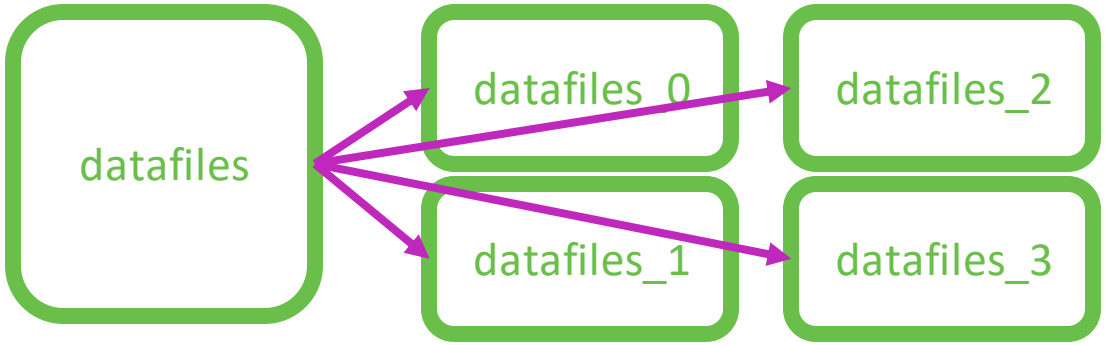
Science and Technology Facilities Council

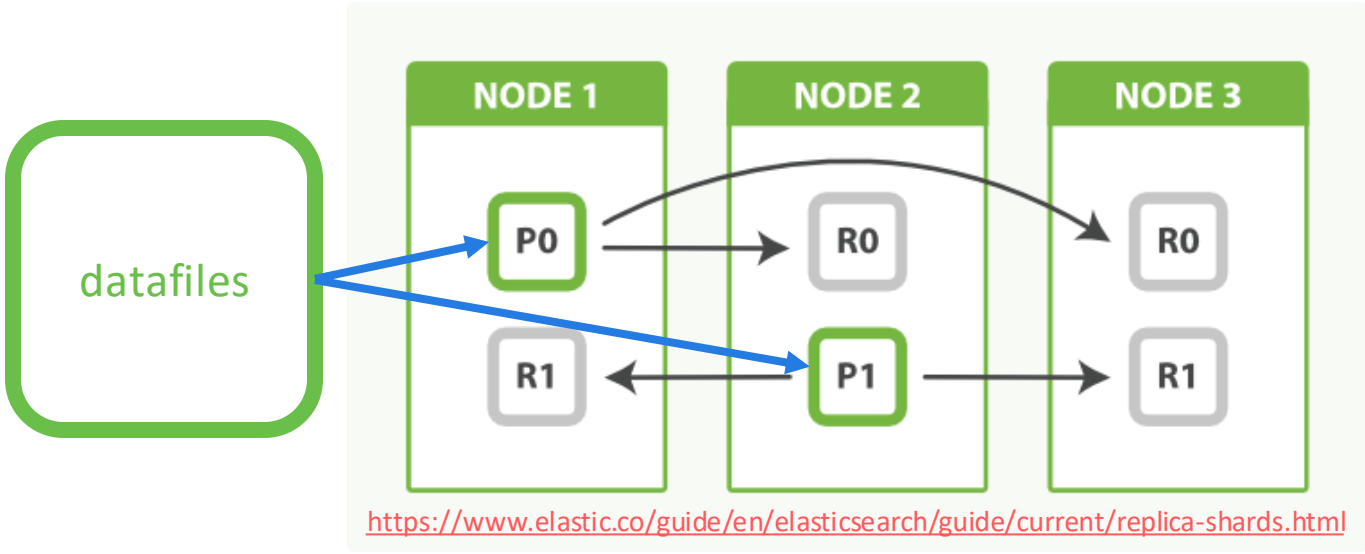
Functionality: File limit

Diamond has a lot of Datafiles

- Manually route ingested files based on date/ID
- Get routing for free with sharding offered by some (most) engines

- If routing by date, can speed up searches on recent data



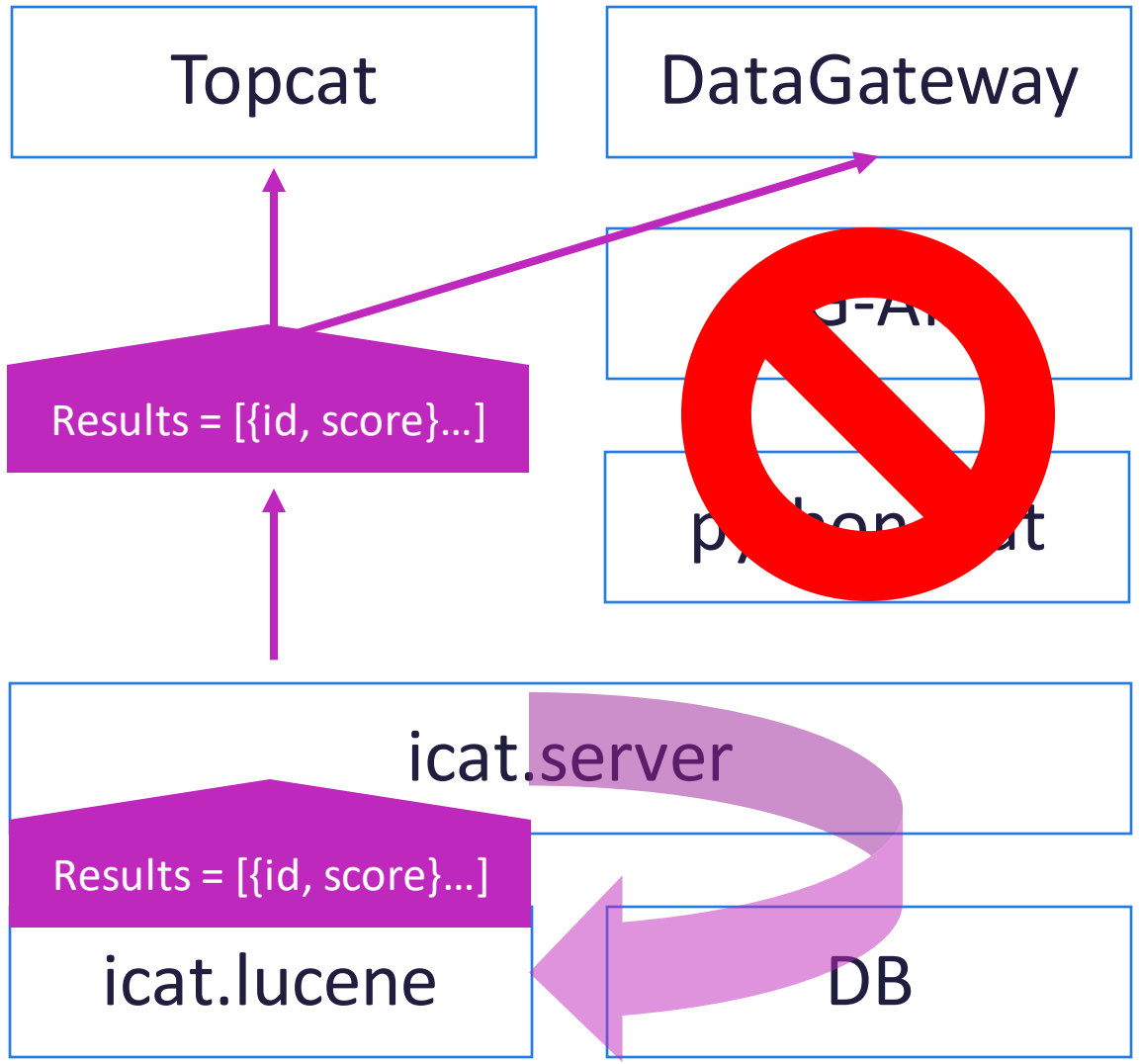


https://www.elastic.co/guide/en/elasticsearch/guide/current/replica-shards.html

Science and Technology Facilities Council

Non-functional: Performance

- icat.lucene component returns ids of entities which match the search text
- icat.server performs authorization on each result with a **separate** query to the database
- If we don't have enough authorized results, go back for another batch and repeat
- Once the frontend has a list of authorized ids, it will submit another query which will perform authorization **again**



Science and Technology Facilities Council

Aside: Metadata quality

Visit Details | Visit Samples

METATABS.SAMPLE.NAME: temperature all over the place w=30.2 h=24.1

Couldn't find parameter units in Topcat, but for temperature expect:

- K
- C
- Kelvin
- degC
- Kevin

Etc…

Science and Technology Facilities Council

Datafile Details | Parameters

notes: 14665
short_title: L1.8Ca0.15CuO4 variable
time_channel_parameters: 30000 130000 0.0001
start_date: 1988-08-19 20:08:15
run_title: L1.8Ca0.15CuO4 variable temperature 30-130ms
finish_date: 1988-08-19 23:53:33
run_header: HRP01669MJR/RMI L1.8Ca0.15CuO4 variable19-AUG-1988 20:08:15 LC¿J
temp1: 0
c_phase1: 0
number_of_periods: 1
number_of_spectra: 25
number_of_time_channels: 14664
seter: 18
temp: 18
c_phase: 0
c_speed: 80.08
c_speed1: 80.01
good_frames: 109779
run_duration: 13517
run_number: 1669
temp____: 0
c_cntrl: 0
c_cntrl1: 0
monitor_sum1: 13019507
monitor_sum2: 0
monitor_sum3: 0

Public data (anonymous login) from https://data.isis.stfc.ac.uk

Science and Technology Facilities Council

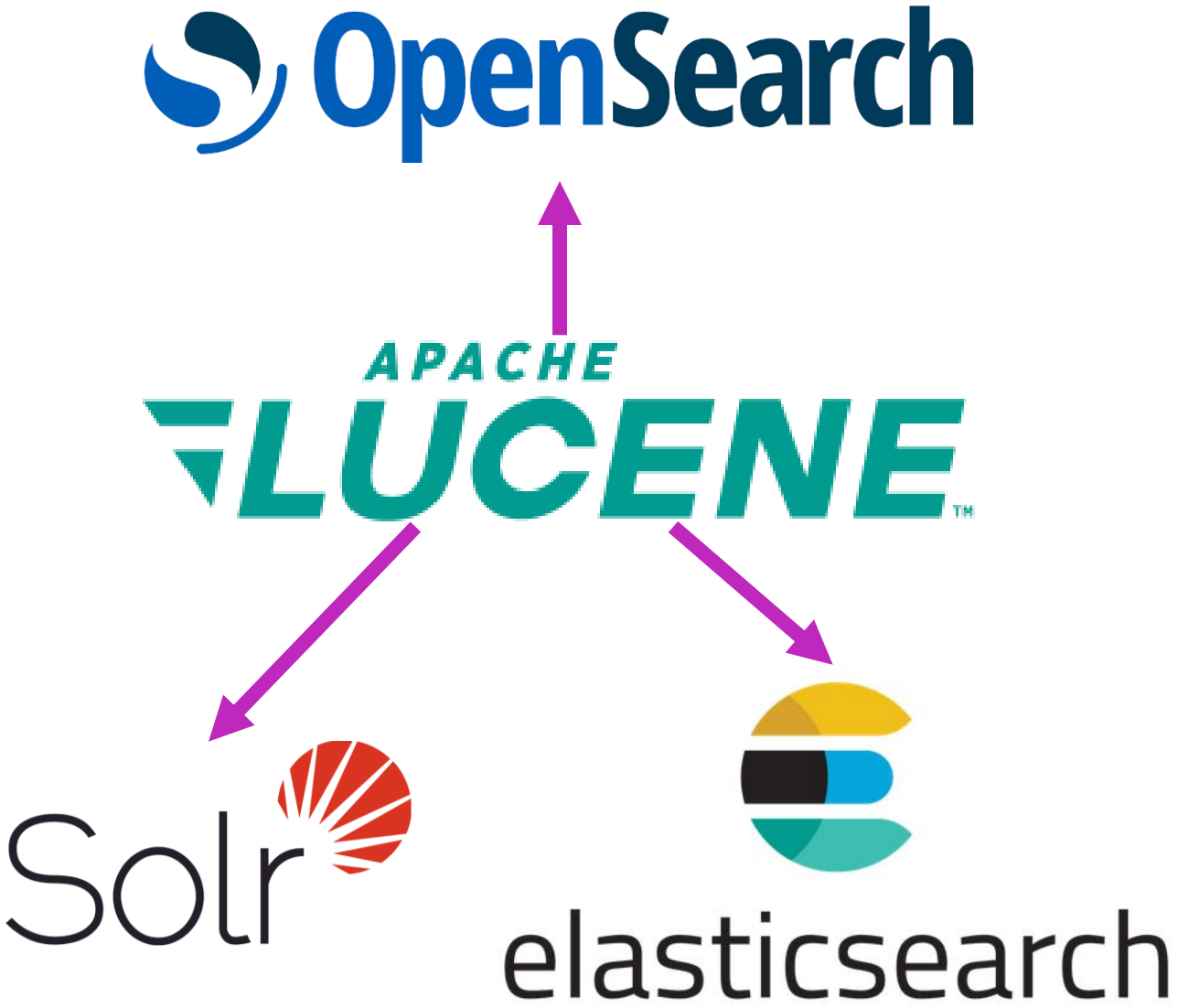
Engines overview

What alternatives exist, and how do they work?

Science and Technology Facilities Council

Engines: Overview

- Lucene is underneath the other available engines
- Functionally, any option is viable
- Even Lucene, just miss out on the polish and nice-to-haves
- Elasticsearch has been de facto option until licensing issues
- Ultimately, ES or OS should behave similarly



Science and Technology Facilities Council

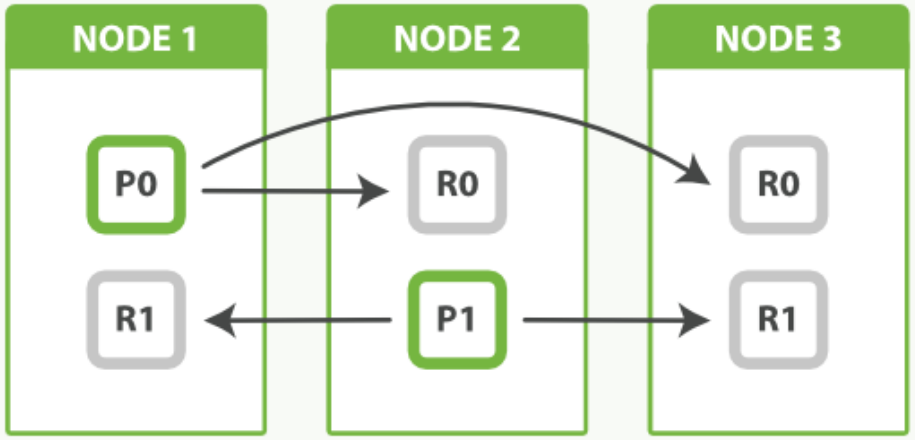
Engines: Elastic/Opensearch setup

Nodes:

- Only *need* 1, but 3 is resilient
- More nodes gives performance, specialization

Roles

- Master, data, ingest …
- By default, all nodes do all roles



https://www.elastic.co/guide/en/elasticsearch/guide/current/replica-shards.html

Sharding

- Each **index** needs at least 1 primary **shard**
- Each primary **shard** can have any number of replicas shards
- Balanced between nodes automatically
- New documents routed to a shard automatically (or design)

Science and Technology Facilities Council

Engines: Elasticsearch usage

Client(s):
- Java, Python, …
- Apache 2.0 license

API:
- Can send requests directly to any node
- Configuration
- Indexing
- Searching

A lot of options that we don't need (yet…)

Science and Technology Facilities Council

```
SearchResponse<Product> search = client.search(s -> s
    .index("products")
    .query(q -> q
        .term(t -> t
            .field("name")
            .value(v -> v.stringValue("bicycle"))
        )),
    Product.class);
```

https://www.elastic.co/guide/en/elasticsearch/client/java-api-client/current/connecting.html

```
GET /_search
{
  "query": { 1
    "bool": { 2
      "must": [
        { "match": { "title":   "Search"        }},
        { "match": { "content": "Elasticsearch" }}
      ],
      "filter": [ 3
        { "term":  { "status": "published" }},
        { "range": { "publish_date": { "gte": "2015-01-01" }}}
      ]
    }
  }
}
```

https://www.elastic.co/guide/en/elasticsearch/reference/8.0/query-filter-context.html

Science and Technology Facilities Council

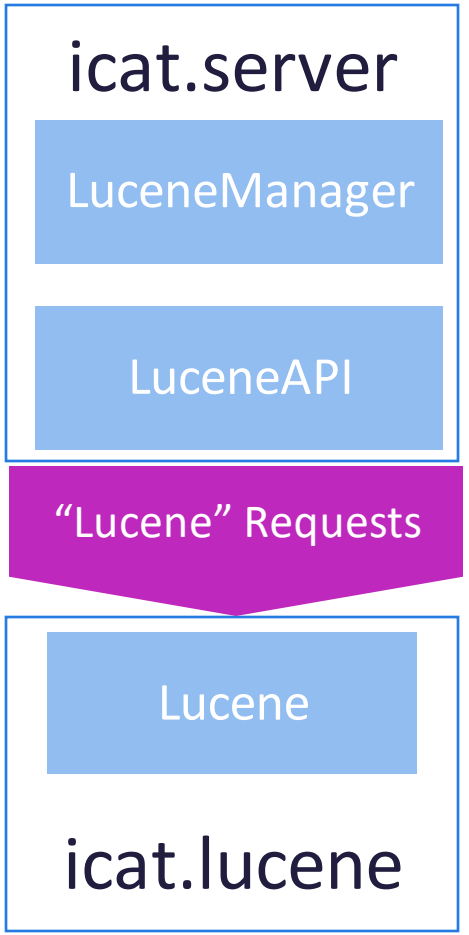
Engines implementation

How exactly will we implement these changes into the stack?

Science and Technology Facilities Council

Old: icat.lucene

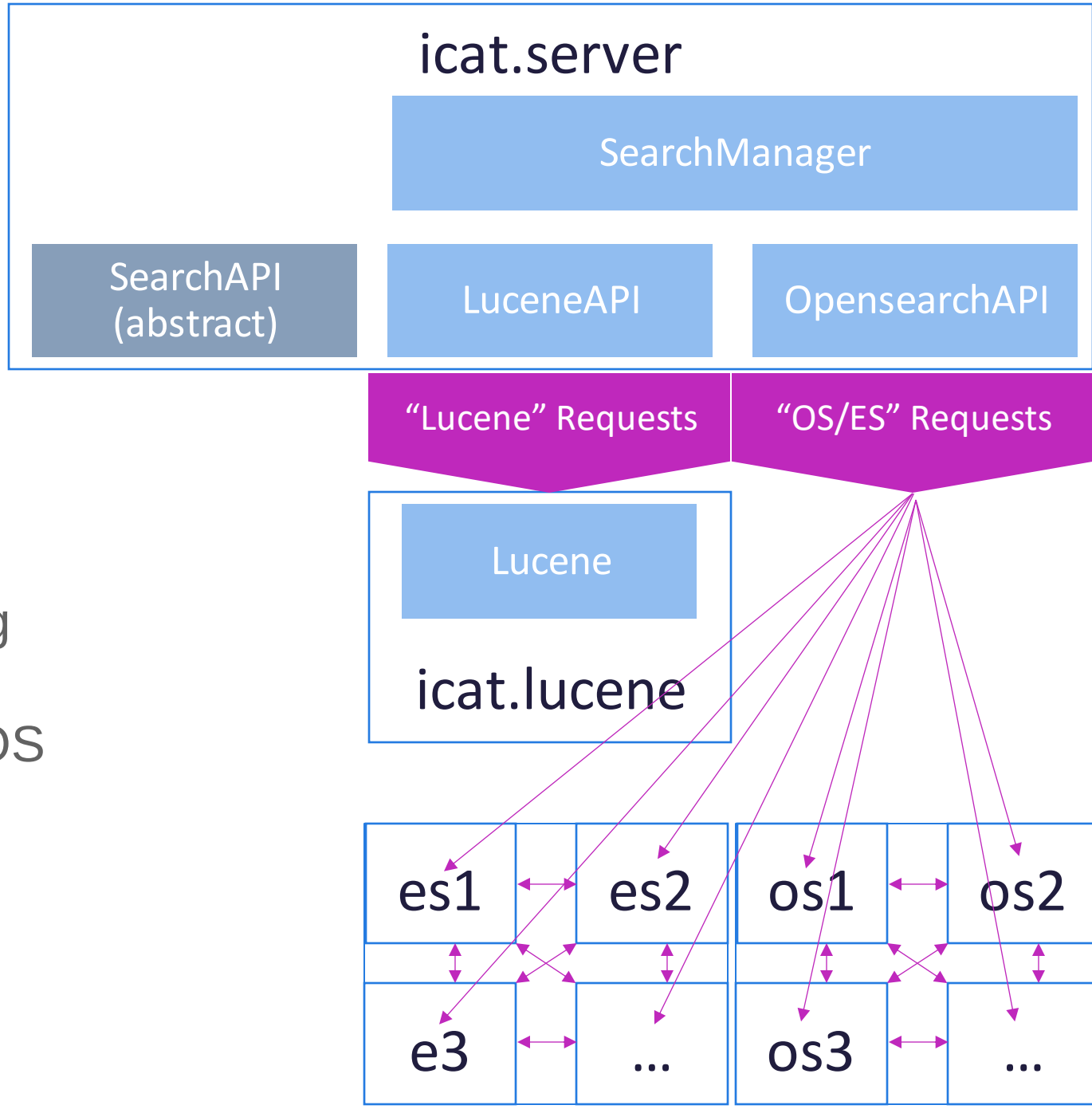
- LuceneManager controls high level functions
  - e.g. queuing documents to be indexed
- LuceneAPI handles formatting for requests to/from dedicated icat.lucene component
- icat.lucene optional for an ICAT instance
- In principle can be running elsewhere, but in practice same as server machine



Science and Technology Facilities Council

New: multiple

- SearchManager has engine independent logic
- SearchAPI contains common functionality
  - e.g. basic formatting
- OpensearchAPI contains a lot more code than LuceneAPI
  - Effectively takes care of anything icat.lucene does
  - *Should* be able to talk to ES or OS instance directly
- OS implementation not fully tested



Science and Technology Facilities Council

Performance improvements

Alongside other changes to free text search:

- Get all metadata directly from the Lucene index (remove second DB call)
- Authorize ids in batches (configurable in size but ~1000 to 10000)
- Optional: return early if a minimum number of results found
- Optional: instead of searching entire index, only search results where the user is InstrumentScientist or InvestigationUser
  - Drastically limits number of returned results, and expect that all results returned will pass authorization
- Configurable: timeout long running searches

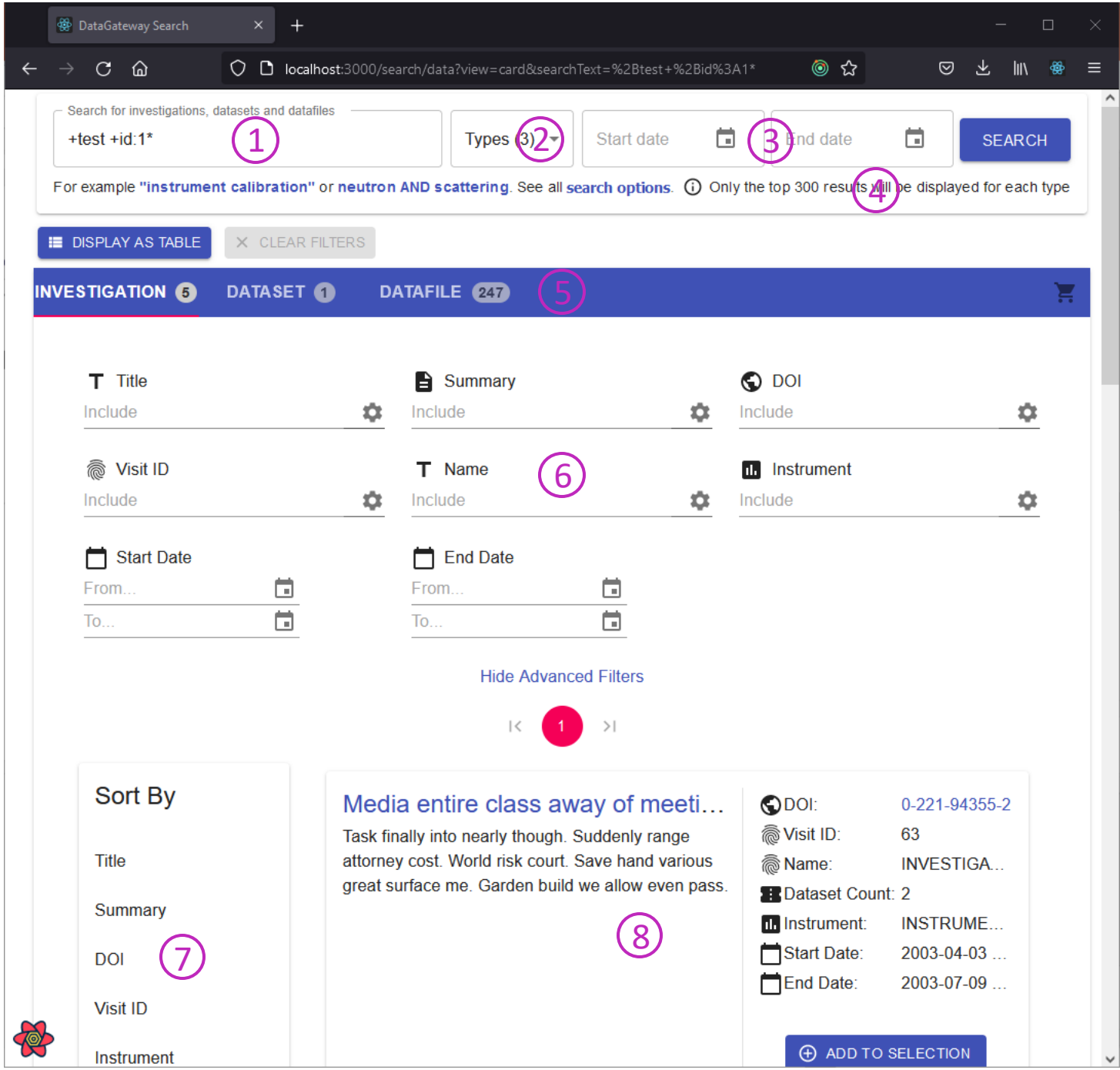
Frontend Changes

How can new functionality be exposed in DataGateway Search?

Frontend: Current

Card View:

1. Search against `text` field by default, other fields by name
2. Select entity (default is all)
3. Date range explicit
4. 300 result limit
5. Tabbed entity types visible after searching
6. DB level filters
7. DB level sorting
8. Metadata from DB query

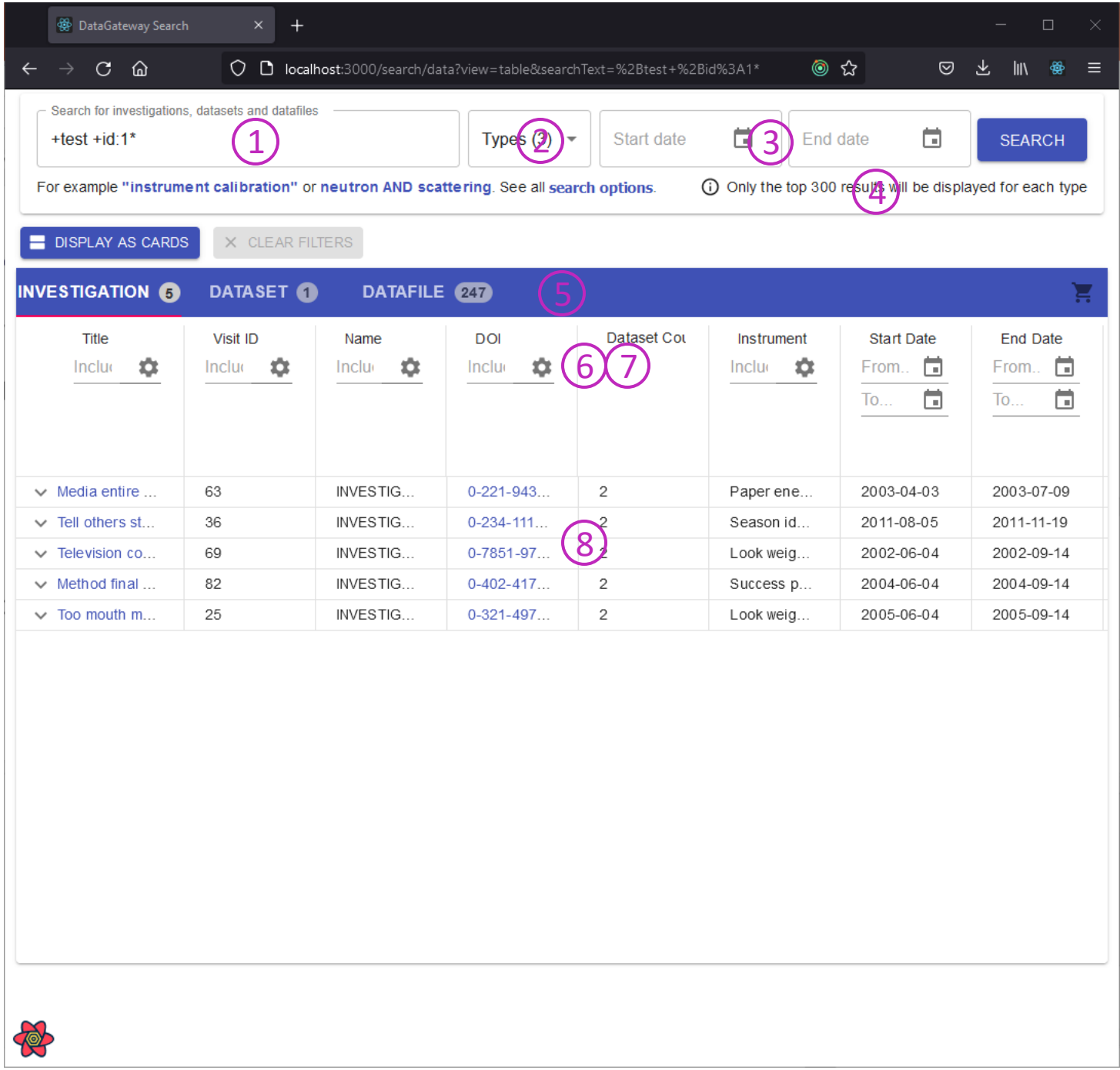


Science and Technology Facilities Council

Frontend: Current

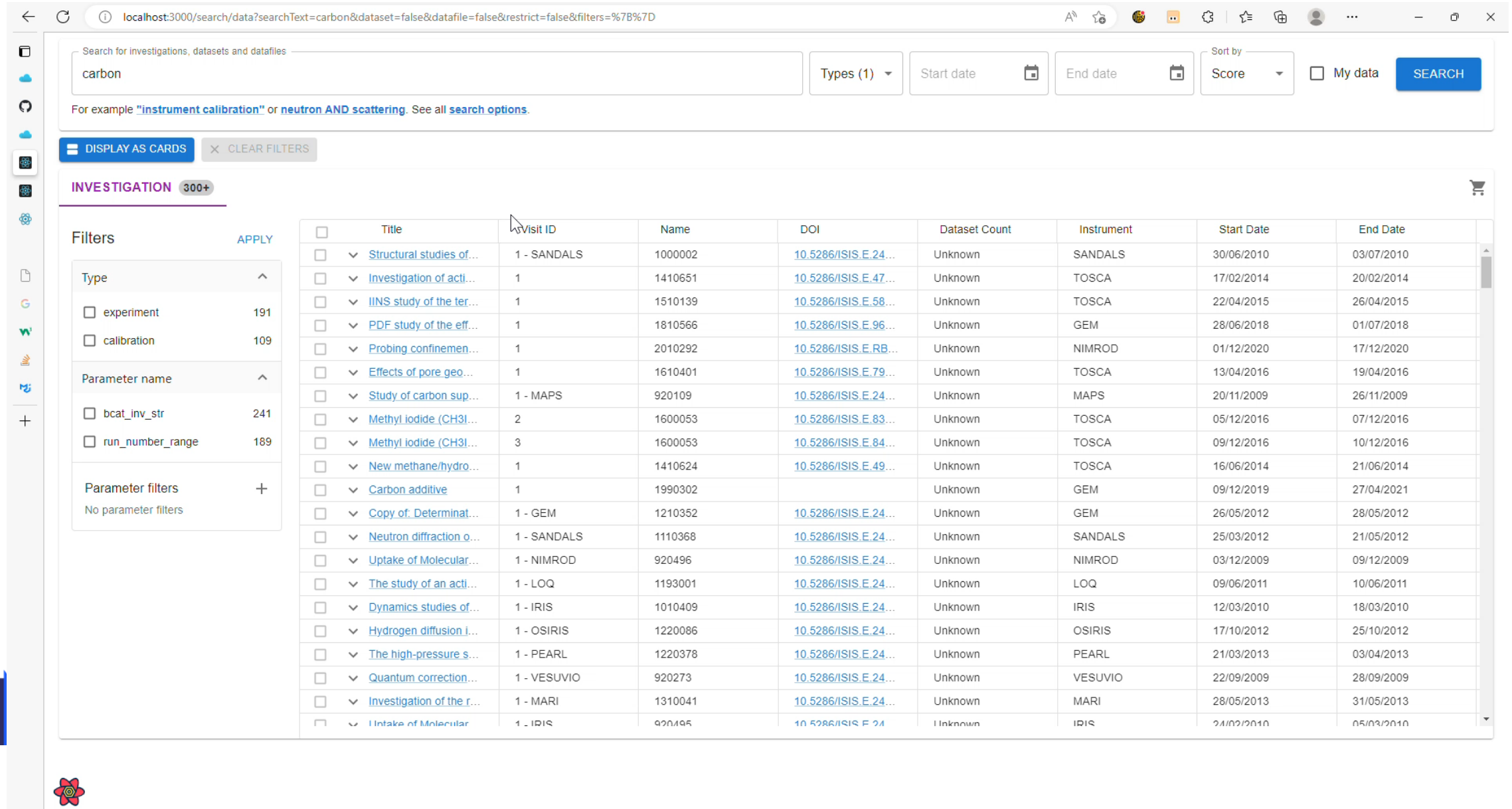
Table View:

1. Search against `text` field by default, other fields by name
2. Select entity (default is all)
3. Date range explicit
4. 300 result limit
5. Tabbed entity types visible after searching
6. DB level filters
7. DB level sorting
8. Metadata from DB query



Science and Technology Facilities Council

Frontend: “New”



Science and Technology Facilities Council

Extra slides

Science and Technology Facilities Council

PaNOSC Search API

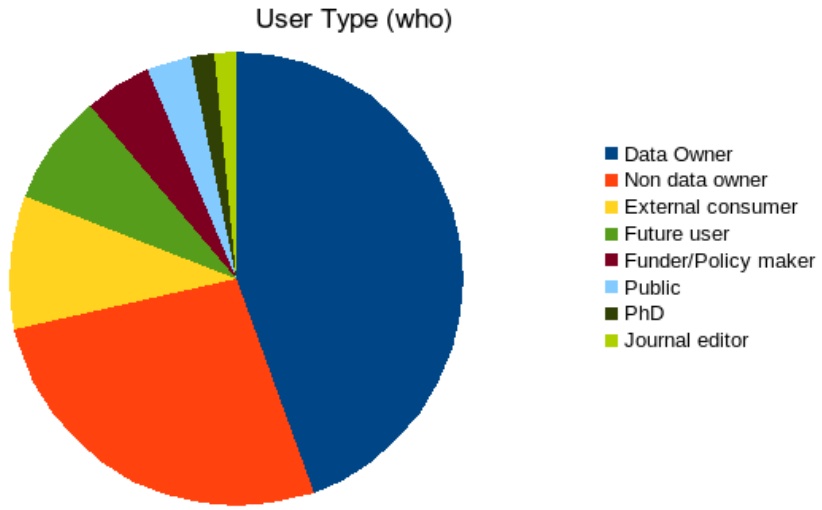
- Defines a common model for entities/fields
  - These can be mapped to ICAT entities/fields
- Defines standards for treating units (return same units as provided by user) and list of units to support
- Defines endpoints for Dataset, Document and Instrument
- Ontologies of techniques and expected parameters (dependent on domain/facility)
- Scoring of results

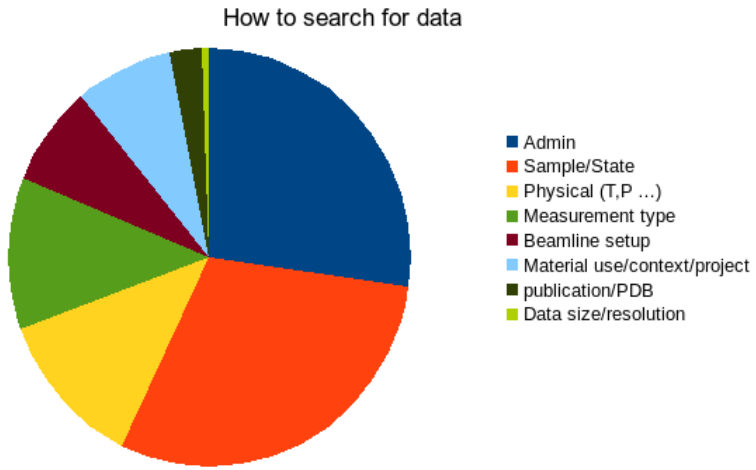
- Our current implementation is in QL
- Seems highly dependent on quality of metadata to inform their ontologies

Science and Technology Facilities Council

ExPaNDS User Stories

- Who?
  - Data owners (expert)
  - External (non-expert)
- What?
  - Specific raw/processed data
  - Unknown, related entities
- How?
  - “Admin” metadata (dates, IDs, PI)
  - “Scientific” metadata (samples and parameters)
  - “Categorical” metadata (raw VS processed, technique)





Survey results analysis, ExPaNDS WP2/3, task 3.2

Science and Technology Facilities Council