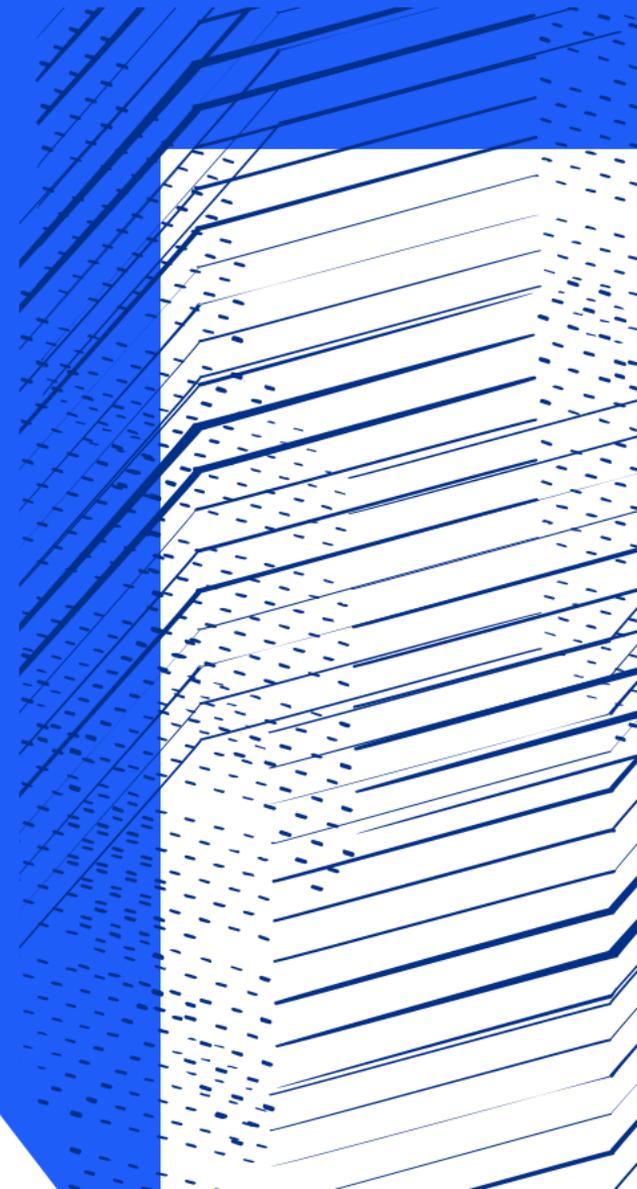




Science and
Technology
Facilities Council

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

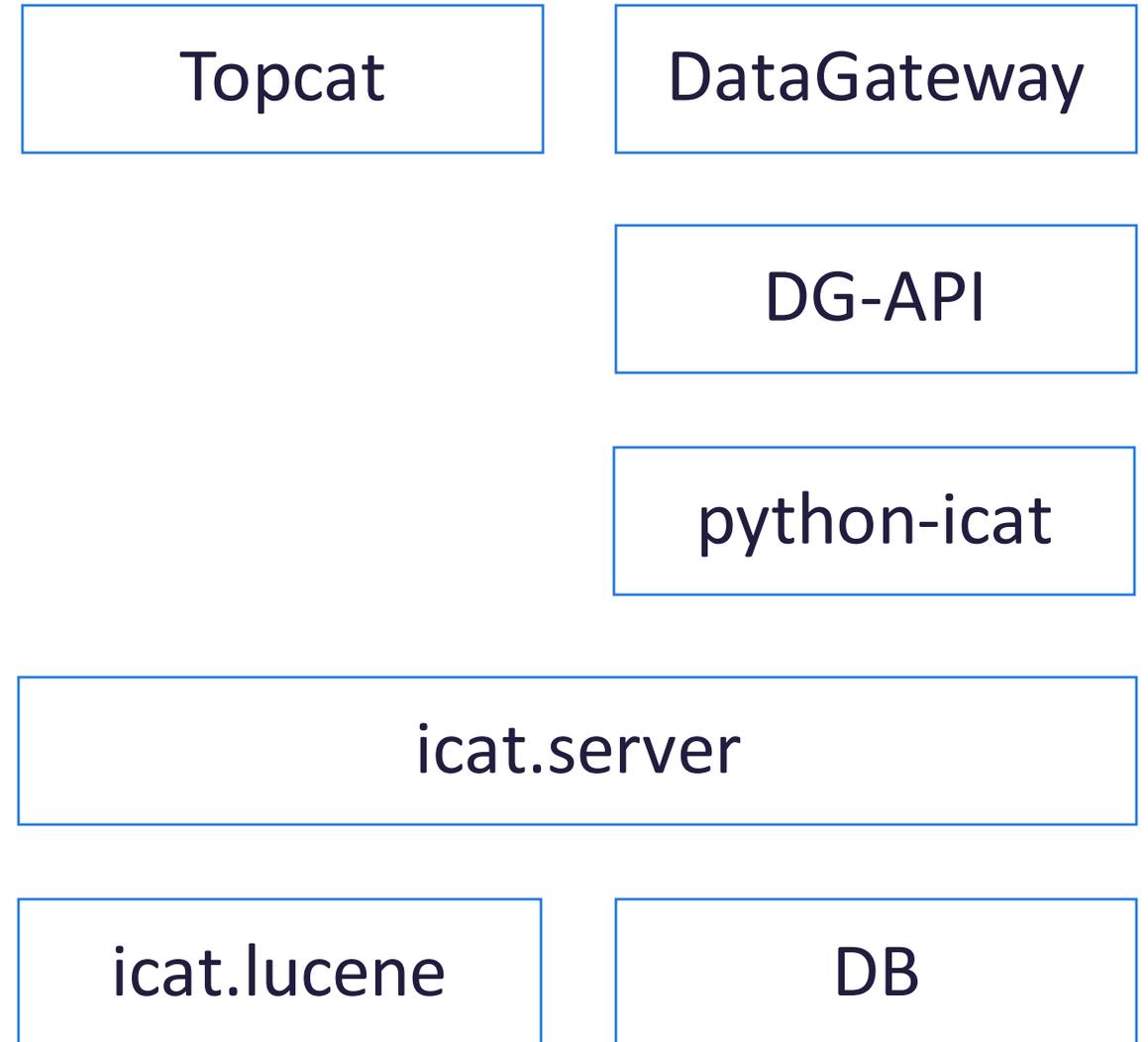
Current state

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



Science and
Technology
Facilities Council

Current Situation

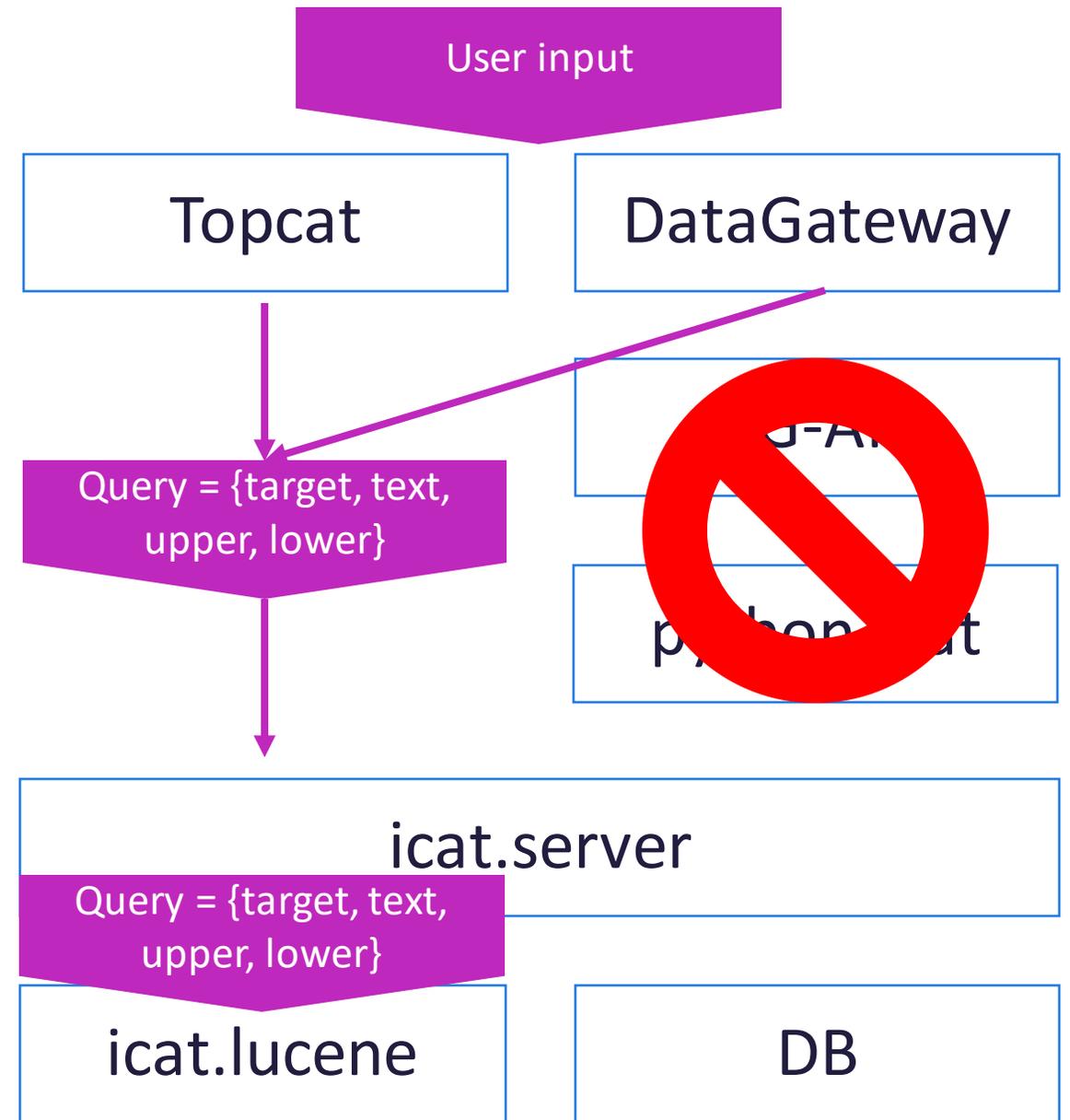


Current Situation

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

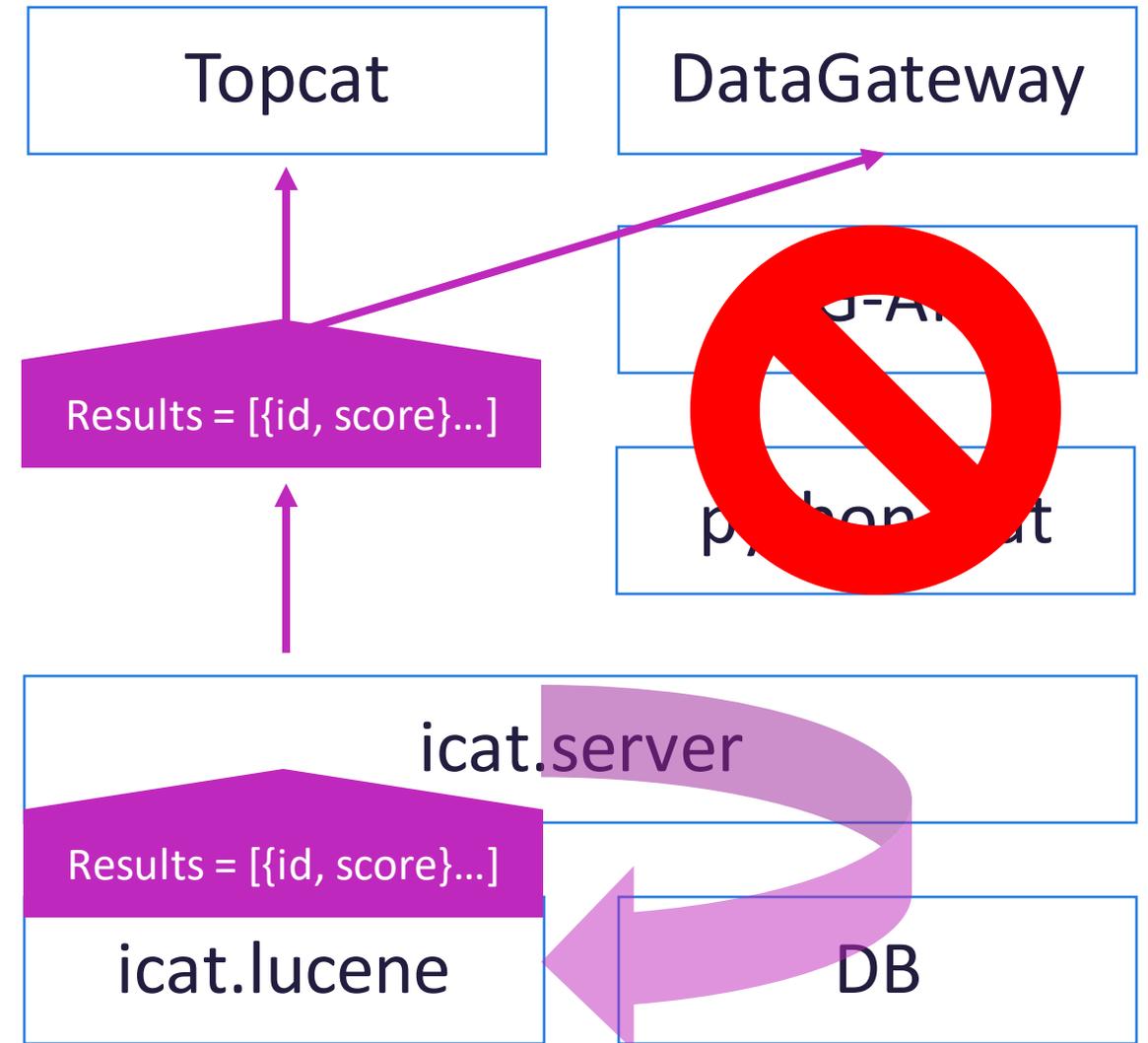
The screenshot shows a search interface with the following elements:

- Navigation tabs: My Data, Browse, Search
- Search input field: -id:4 meta~
- Filter fields: Start Date, End Date
- Action buttons: Add Parameter, Add Sample
- Filter checkboxes: Visit (checked), Dataset (checked), Datafile (checked)
- Text input field: Milk
- Search button: Search



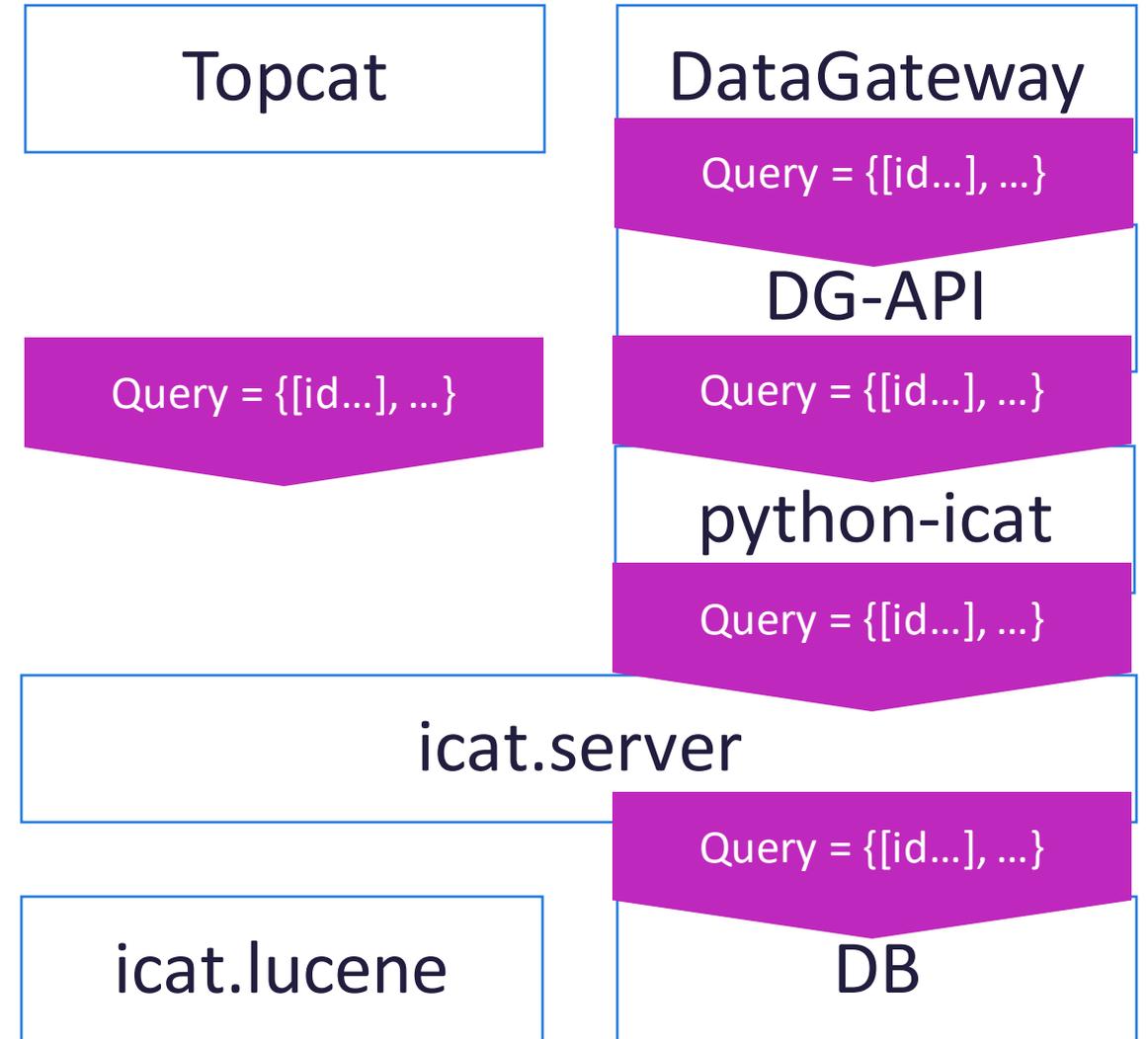
Current Situation

- 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



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



Current Situation

Search Results

Visit Dataset Datafile

Title	Visit Id	Size	Beamline
Containing	Containing..		Containing.
Milk Samples	Proposal 21 - 7	0 B	Instrument 1

and/or dates into

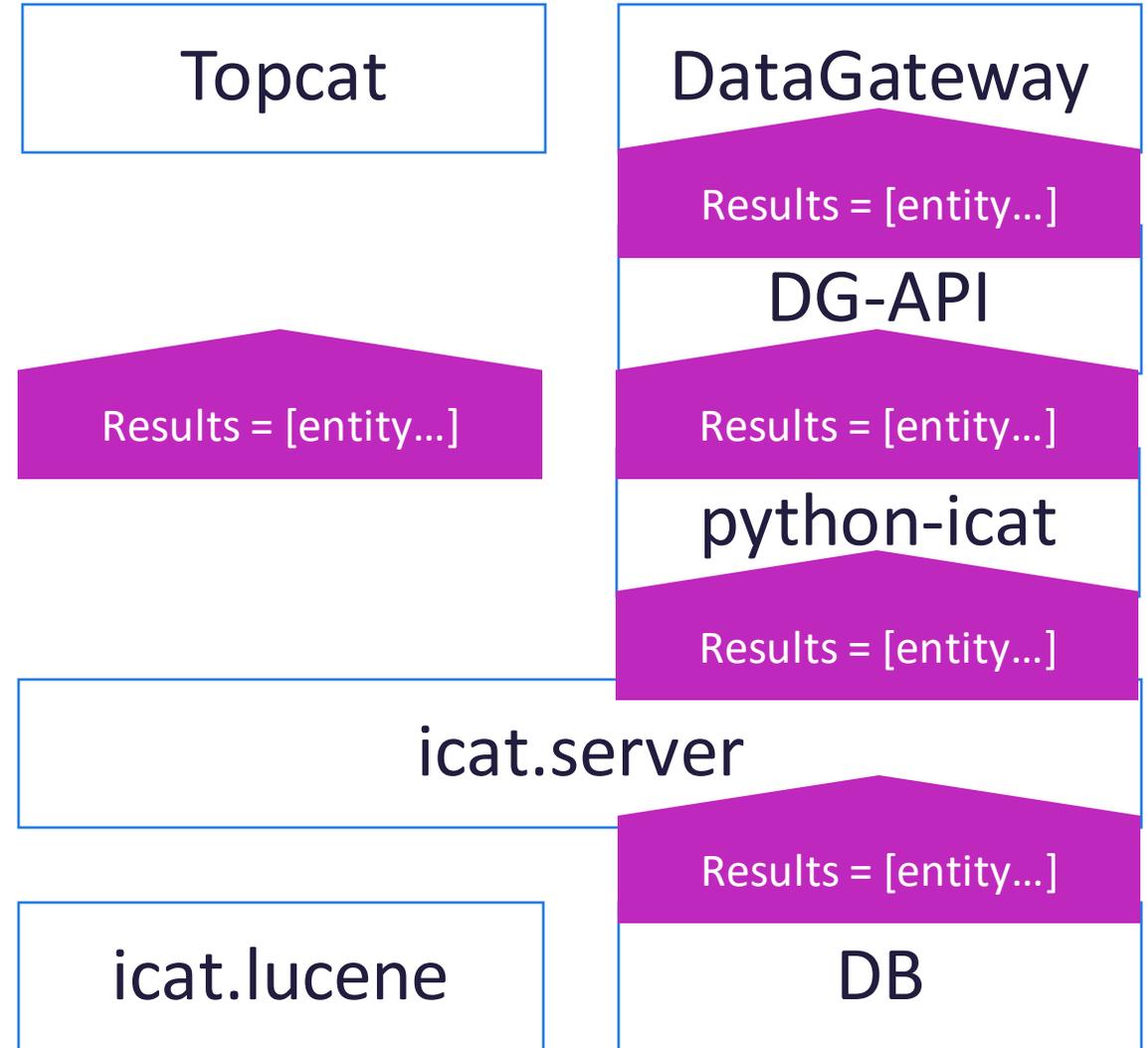
irect to the

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

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?

The screenshot displays a search interface with the following components:

- Navigation:** 'My Data', 'Browse', and 'Search' tabs.
- Search Input:** A search bar containing '-id:4 meta~'.
- Filters:** 'Start Date' and 'End Date' fields with calendar icons.
- Actions:** 'Add Parameter' and 'Add Sample' buttons.
- Entity Selection:** Checkboxes for 'Visit', 'Dataset', and 'Datafile', all of which are checked.
- Search Term:** A text input field containing 'Milk'.
- Search Button:** A large 'Search' button at the bottom.

Search Results:

Results are shown under the 'Visit' tab. The table below summarizes the results:

Title	Visit Id	Size	Beamline
Containing	Containing..		Containing.
Milk Samples	Proposal 21 - 7	0 B	Instrument 1

Below the table, there are tabs for 'Visit Details', 'Visit Users', and 'Visit Samples'. The 'Visit Details' tab is active, showing the following information:

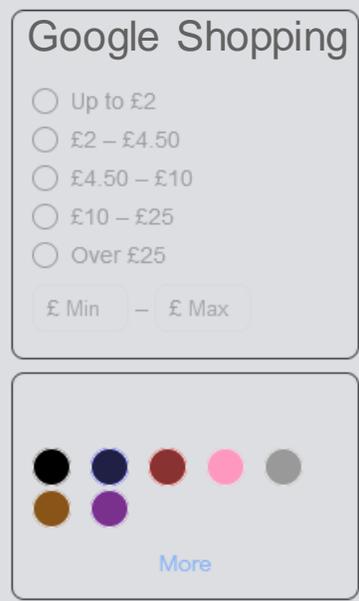
- Proposal:** Proposal 21
- Title:** Milk Samples
- Summary:** Group 2 Metal Survey
- Start Date:** 2017-07-13T12:01:54.000Z
- End Date:** 2017-07-21T12:01:54.000Z

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



DATAFILEPARAMETER
123 ID
ABC CREATE_ID
CREATE_TIME
DATETIME_VALUE
123 ERROR
ABC MOD_ID
MOD_TIME
123 NUMERIC_VALUE
123 RANGEBOTTOM
123 RANGETOP
ABC STRING_VALUE
123 DATAFILE_ID
123 PARAMETER_TYPE_ID

DATASETPARAMETER
123 ID
ABC CREATE_ID
CREATE_TIME
DATETIME_VALUE
123 ERROR
ABC MOD_ID
MOD_TIME
123 NUMERIC_VALUE
123 RANGEBOTTOM
123 RANGETOP
ABC STRING_VALUE
123 DATASET_ID
123 PARAMETER_TYPE_ID

INVESTIGATIONPARAMETER
123 ID
ABC CREATE_ID
CREATE_TIME
DATETIME_VALUE
123 ERROR
ABC MOD_ID
MOD_TIME
123 NUMERIC_VALUE
123 RANGEBOTTOM
123 RANGETOP
ABC STRING_VALUE
123 INVESTIGATION_ID
123 PARAMETER_TYPE_ID

PARAMETER_TYPE
123 ID
123 APPLICABLETODATACOLLECTION
123 APPLICABLETODATAFILE
123 APPLICABLETODATASET
123 APPLICABLETOINVESTIGATION
123 APPLICABLETOSAMPLE
ABC CREATE_ID
CREATE_TIME
ABC DESCRIPTION
123 ENFORCED
123 MAXIMUMNUMERICVALUE
123 MINIMUMNUMERICVALUE
ABC MOD_ID
MOD_TIME
ABC NAME
ABC UNITS
ABC UNITSFULLNAME
123 VALUETYPE
123 VERIFIED
123 FACILITY_ID
ABC PID

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

My Data Browse Search

ionise

Start Date

End Date

Add Parameter Add Sample

Visit

Search Results

Visit Dataset Datafile

Title

Containing...

Ionisation of Ca

My Data Browse Search

ionizing

Start Date

End Date

Add Parameter Add Sample

Visit

Search Results

Visit Dataset Datafile

Title

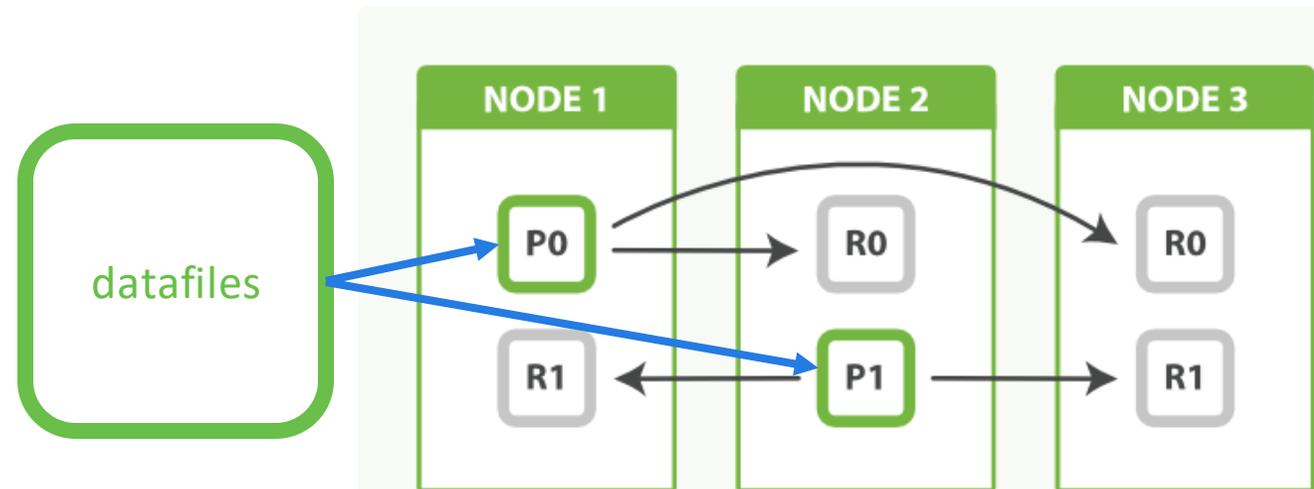
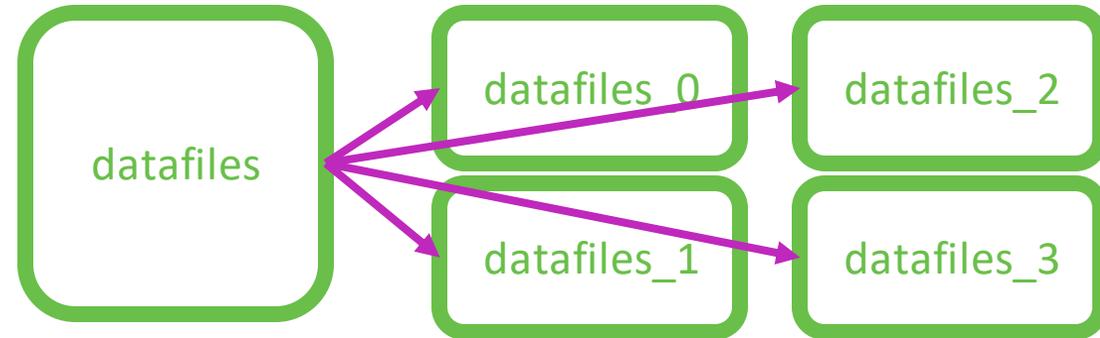
Containing...

Ionization of Calcium

Functionality: File limit

Diamond has a lot of Datafiles

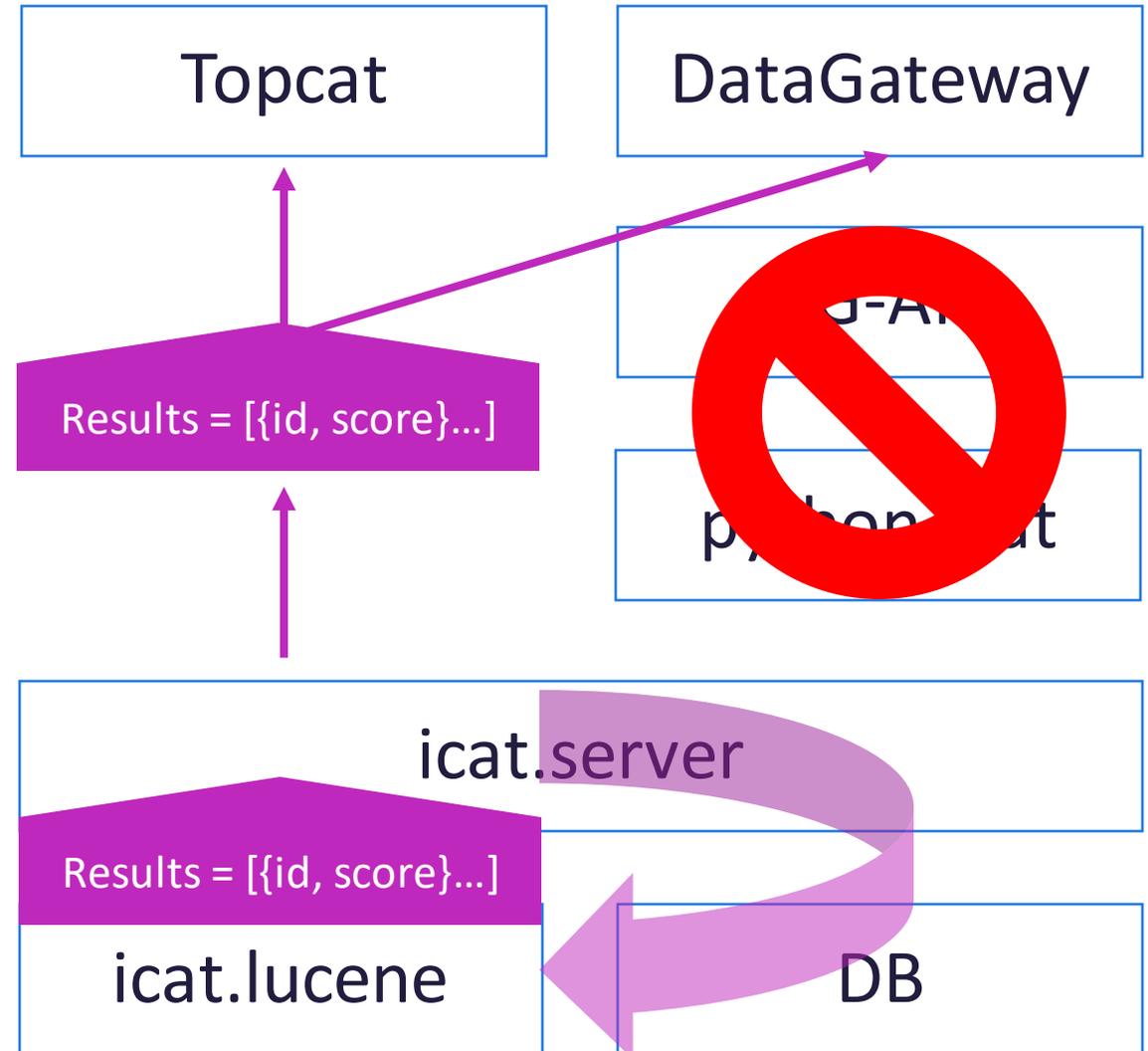
- **Manually route** ingested files based on date/ID
- Get **routing for free** with sharding offered by some (most) engines
- If routing by date, can speed up searches on recent data



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

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**



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...



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

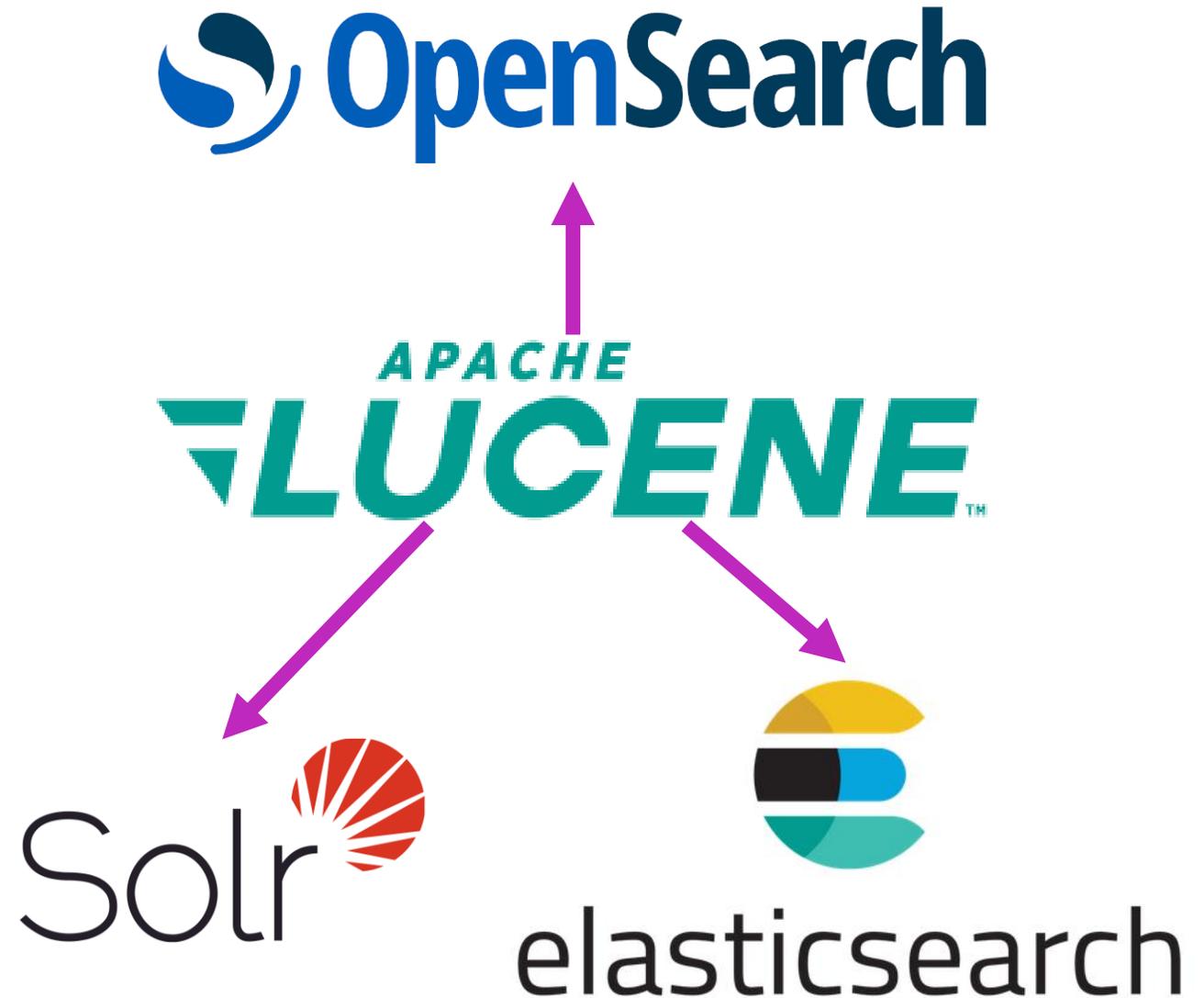
What alternatives exist, and how do they work?



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



Engines: Elastic/Opensearch setup

Nodes:

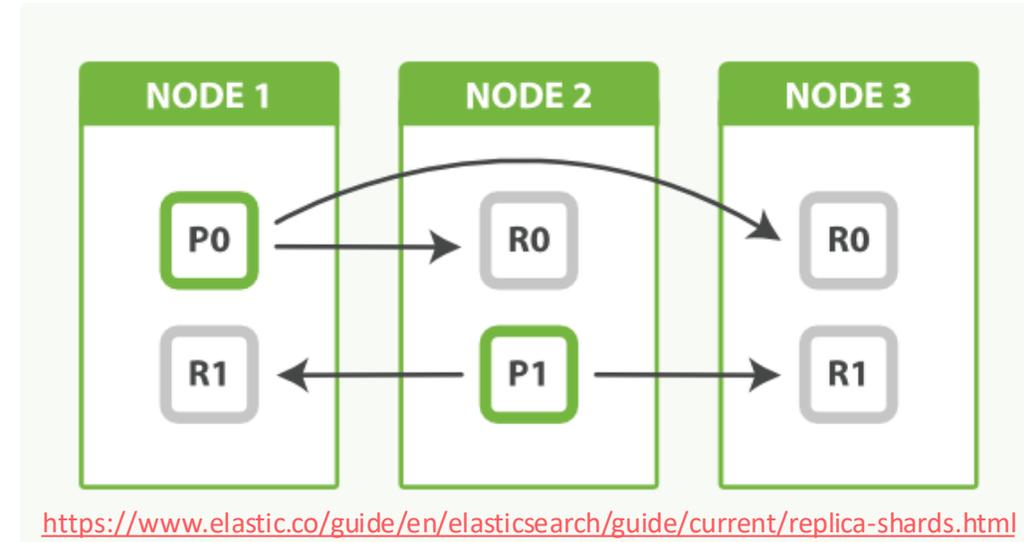
- Only *need* 1, but 3 is resilient
- More nodes gives performance, specialization

Roles

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

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)



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...)

```
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": { ❶
    "bool": { ❷
      "must": [
        { "match": { "title": "Search" }},
        { "match": { "content": "Elasticsearch" }}
      ],
      "filter": [ ❸
        { "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

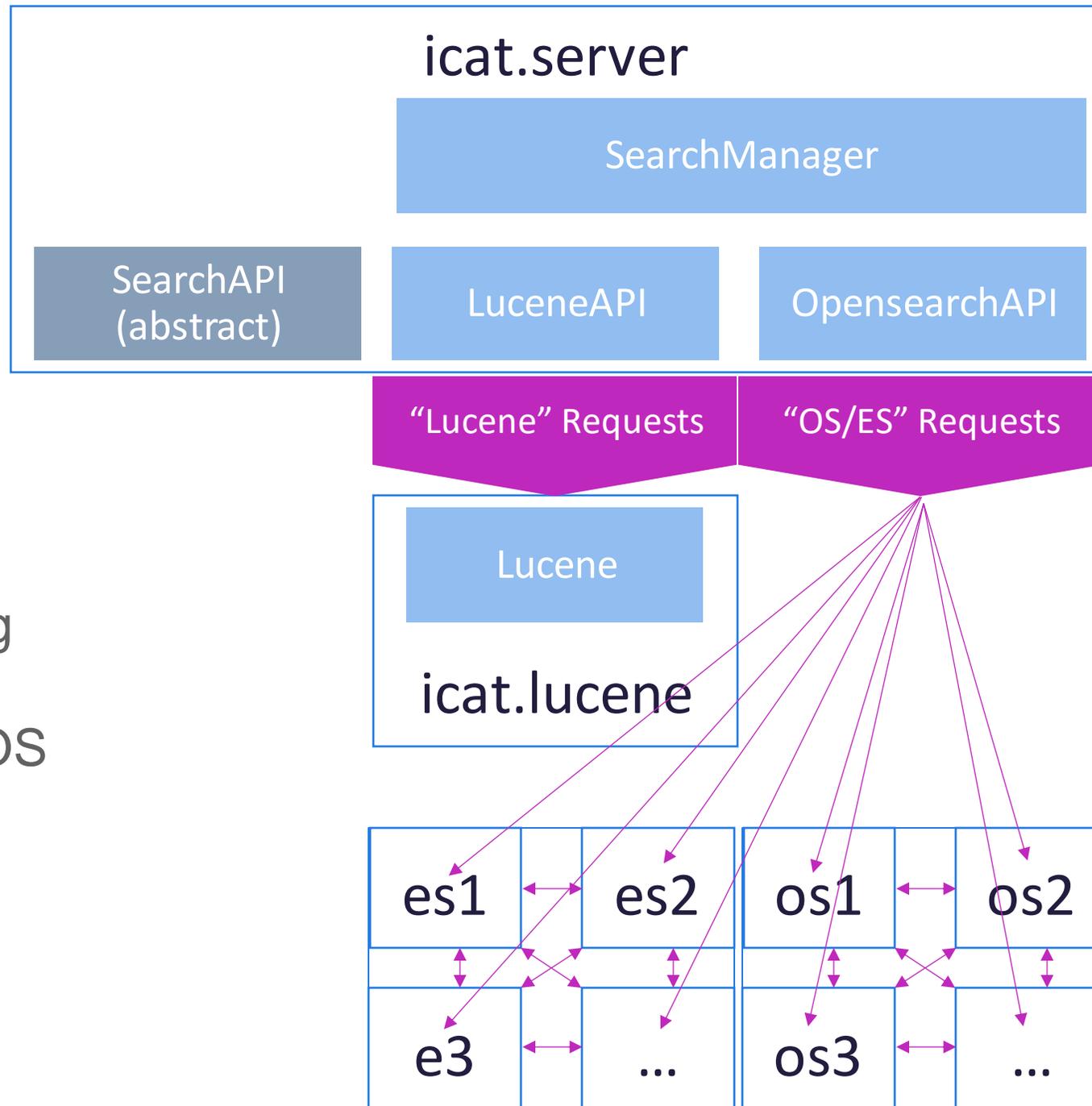
Old: icat.lucene

- 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



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



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



Science and
Technology
Facilities Council

Frontend Changes

How can new functionality be exposed in
DataGateway Search?

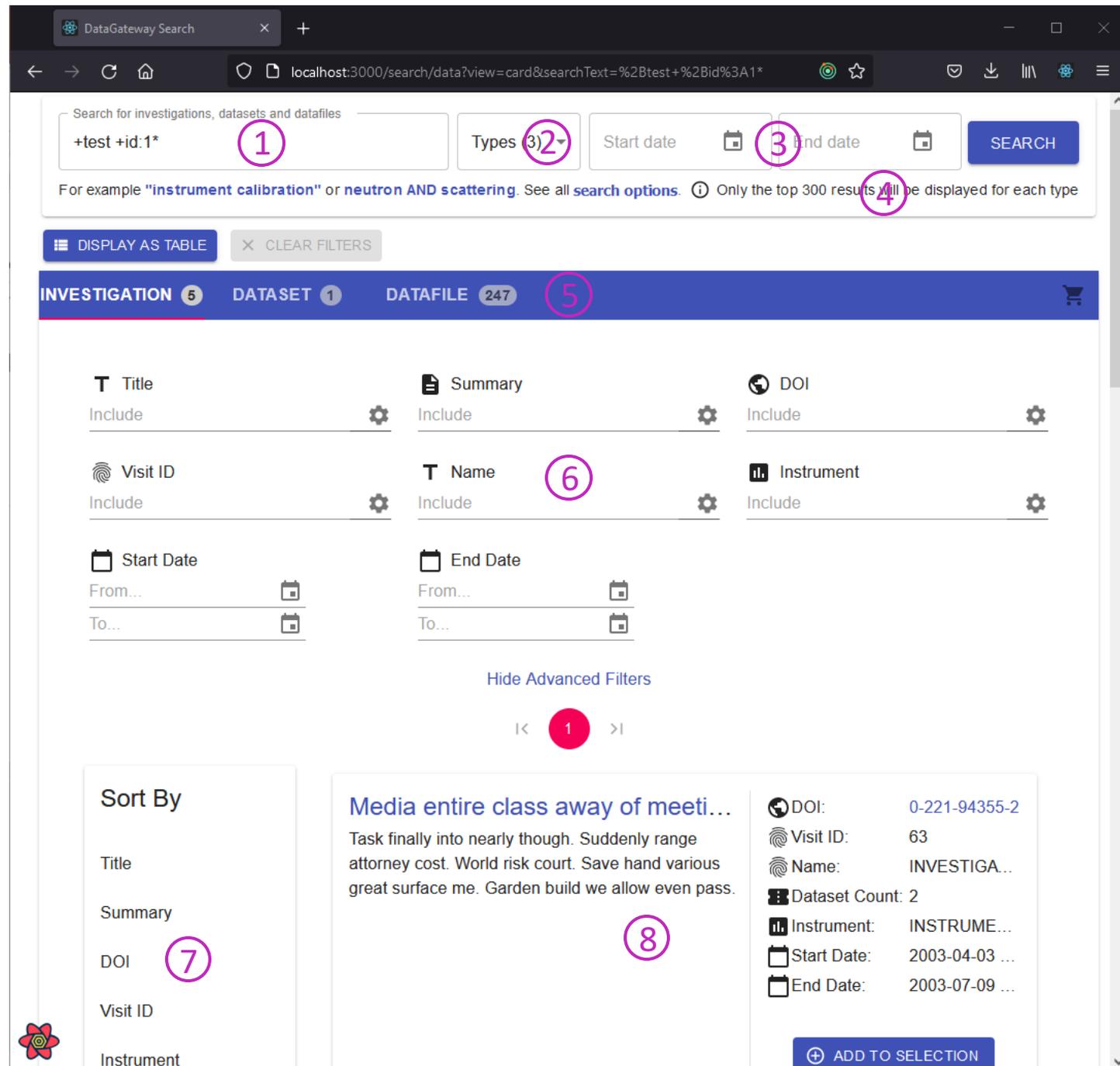


Science and
Technology
Facilities Council

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



The screenshot displays the DataGateway Search interface. The search bar contains the query "+test +id:1*" (1). The search filters include "Types" (2), "Start date" (3), and "End date" (3). A note indicates that only the top 300 results will be displayed for each type (4). The interface shows tabs for "INVESTIGATION" (5), "DATASET" (1), and "DATAFILE" (247). The search results are displayed in a card view, showing fields like Title, Summary, DOI, Visit ID, Name (6), Instrument, Start Date, and End Date. A "Sort By" section (7) is visible, and a "Media entire class away of meeti..." result is shown (8). The interface also includes a "Hide Advanced Filters" button and a "1" page indicator.



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

The screenshot shows the DataGateway Search interface. The search bar contains the query "+test +id:1*" (1). The "Types" dropdown is set to "all" (2). The "Start date" and "End date" fields are empty (3). A "SEARCH" button is present. Below the search bar, a message states "Only the top 300 results will be displayed for each type" (4). The interface has tabs for "INVESTIGATION" (5), "DATASET" (1), and "DATAFILE" (247). The table below shows search results with columns: Title, Visit ID, Name, DOI, Dataset Count, Instrument, Start Date, and End Date. The table is filtered to show 5 results. The "Dataset Count" column is circled (6), and the "Dataset Count" value "2" is circled (7). The "DOI" value "0-7851-97..." is circled (8).

Title	Visit ID	Name	DOI	Dataset Cot	Instrument	Start Date	End Date
Media entire ...	63	INVESTIG...	0-221-943...	2	Paper ene...	2003-04-03	2003-07-09
Tell others st...	36	INVESTIG...	0-234-111...	2	Season id...	2011-08-05	2011-11-19
Television co...	69	INVESTIG...	0-7851-97...	2	Look weig...	2002-06-04	2002-09-14
Method final ...	82	INVESTIG...	0-402-417...	2	Success p...	2004-06-04	2004-09-14
Too mouth m...	25	INVESTIG...	0-321-497...	2	Look weig...	2005-06-04	2005-09-14



Frontend: "New"

localhost:3000/search/data?searchText=carbon&dataset=false&datafile=false&restrict=false&filters=%7B%7D

Search for investigations, datasets and datafiles
carbon

Types (1) Start date End date Sort by Score My data SEARCH

For example "instrument calibration" or neutron AND scattering. See all search options.

DISPLAY AS CARDS CLEAR FILTERS

INVESTIGATION 300+

Filters APPLY

Type ^
 experiment 191
 calibration 109

Parameter name ^
 bcat_inv_str 241
 run_number_range 189

Parameter filters +
No parameter filters

<input type="checkbox"/>	Title	Visit ID	Name	DOI	Dataset Count	Instrument	Start Date	End Date
<input type="checkbox"/>	Structural studies of...	1 - SANDALS	1000002	10.5286/ISIS.E.24...	Unknown	SANDALS	30/06/2010	03/07/2010
<input type="checkbox"/>	Investigation of acti...	1	1410651	10.5286/ISIS.E.47...	Unknown	TOSCA	17/02/2014	20/02/2014
<input type="checkbox"/>	IINS study of the ter...	1	1510139	10.5286/ISIS.E.58...	Unknown	TOSCA	22/04/2015	26/04/2015
<input type="checkbox"/>	PDF study of the eff...	1	1810566	10.5286/ISIS.E.96...	Unknown	GEM	28/06/2018	01/07/2018
<input type="checkbox"/>	Probing confinemen...	1	2010292	10.5286/ISIS.E.RB...	Unknown	NIMROD	01/12/2020	17/12/2020
<input type="checkbox"/>	Effects of pore geo...	1	1610401	10.5286/ISIS.E.79...	Unknown	TOSCA	13/04/2016	19/04/2016
<input type="checkbox"/>	Study of carbon sup...	1 - MAPS	920109	10.5286/ISIS.E.24...	Unknown	MAPS	20/11/2009	26/11/2009
<input type="checkbox"/>	Methyl iodide (CH3I...	2	1600053	10.5286/ISIS.E.83...	Unknown	TOSCA	05/12/2016	07/12/2016
<input type="checkbox"/>	Methyl iodide (CH3I...	3	1600053	10.5286/ISIS.E.84...	Unknown	TOSCA	09/12/2016	10/12/2016
<input type="checkbox"/>	New methane/hydro...	1	1410624	10.5286/ISIS.E.49...	Unknown	TOSCA	16/06/2014	21/06/2014
<input type="checkbox"/>	Carbon additive	1	1990302		Unknown	GEM	09/12/2019	27/04/2021
<input type="checkbox"/>	Copy of Determinat...	1 - GEM	1210352	10.5286/ISIS.E.24...	Unknown	GEM	26/05/2012	28/05/2012
<input type="checkbox"/>	Neutron diffraction o...	1 - SANDALS	1110368	10.5286/ISIS.E.24...	Unknown	SANDALS	25/03/2012	21/05/2012
<input type="checkbox"/>	Uptake of Molecular...	1 - NIMROD	920496	10.5286/ISIS.E.24...	Unknown	NIMROD	03/12/2009	09/12/2009
<input type="checkbox"/>	The study of an acti...	1 - LOQ	1193001	10.5286/ISIS.E.24...	Unknown	LOQ	09/06/2011	10/06/2011
<input type="checkbox"/>	Dynamics studies of...	1 - IRIS	1010409	10.5286/ISIS.E.24...	Unknown	IRIS	12/03/2010	18/03/2010
<input type="checkbox"/>	Hydrogen diffusion i...	1 - OSIRIS	1220086	10.5286/ISIS.E.24...	Unknown	OSIRIS	17/10/2012	25/10/2012
<input type="checkbox"/>	The high-pressure s...	1 - PEARL	1220378	10.5286/ISIS.E.24...	Unknown	PEARL	21/03/2013	03/04/2013
<input type="checkbox"/>	Quantum correction...	1 - VESUVIO	920273	10.5286/ISIS.E.24...	Unknown	VESUVIO	22/09/2009	28/09/2009
<input type="checkbox"/>	Investigation of the r...	1 - MARI	1310041	10.5286/ISIS.E.24...	Unknown	MARI	28/05/2013	31/05/2013
<input type="checkbox"/>	Uptake of Molecular...	1 - IRIS	920496	10.5286/ISIS.E.24...	Unknown	IRIS	03/12/2009	09/12/2009





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

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)

