PaN-Finder

Intelligent Search Engine

OSCARS
Open Science Clusters' Action for Research & Society

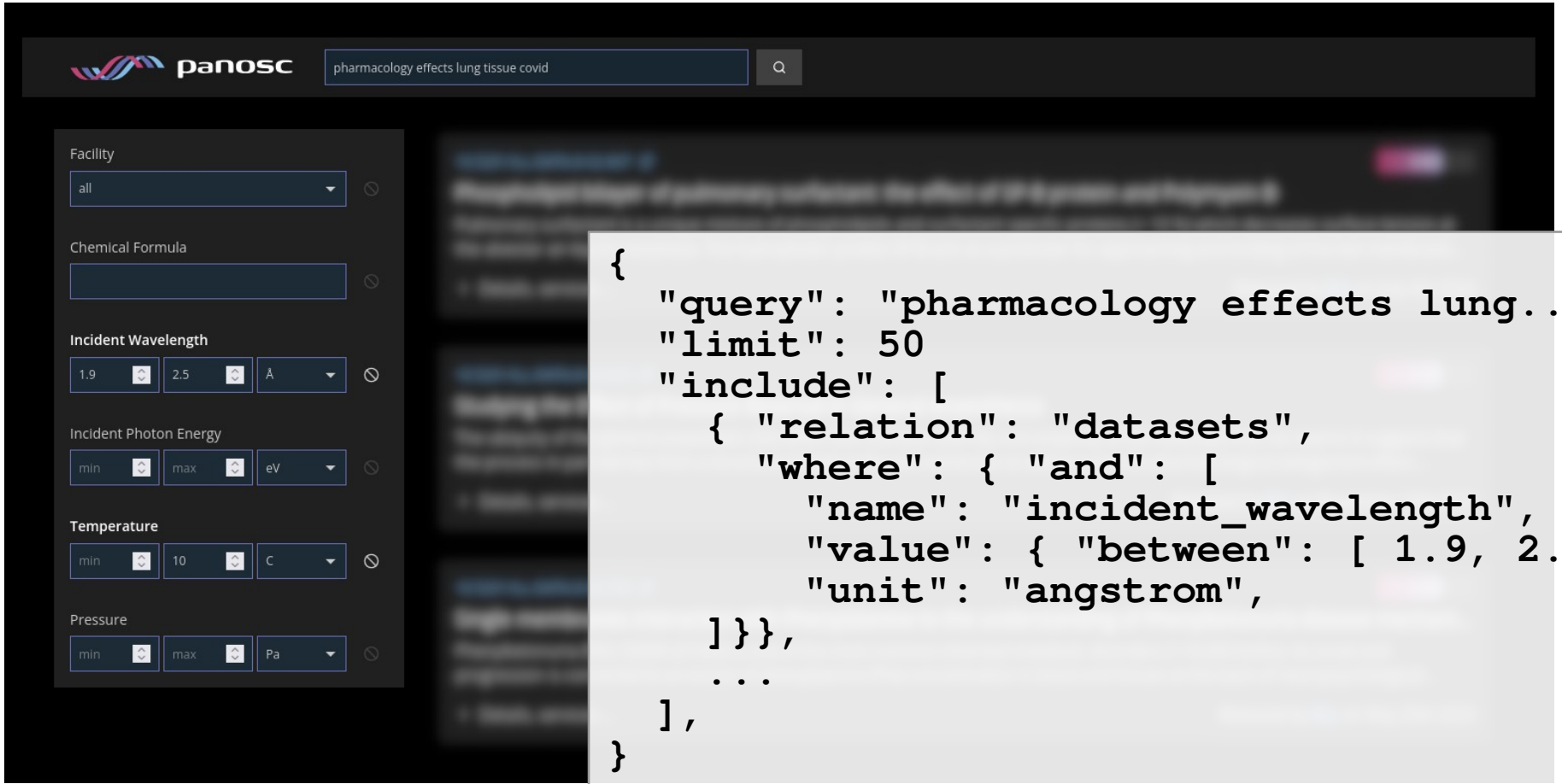
Massimiliano (Max) Novelli
max.novelli@ess.eu

Senior Data Curation Scientist
European Spallation Source, DMSC
Copenhagen, Denmark

EUROPEAN SPALLATION SOURCE

Metadata Catalogs Satellite Meeting
NoBUGS 2024, ESRF, France, 2024/09/27

PaNOSC Data Portal




```
{
  "query": "pharmacology effects lung...",
  "limit": 50
  "include": [
    { "relation": "datasets",
      "where": { "and": [
        "name": "incident_wavelength",
        "value": { "between": [ 1.9, 2.5 ] },
        "unit": "angstrom",
    ]}},
    ...
  ],
}
```

Ideas

Query in a natural language

Find the ***50*** most relevant ***datasets*** related to ***pharmacology effects*** on ***lung tissues*** infected with ***Covid***, where the ***incident wavelength*** was between ***1.9 and 2.5 Angstrom*** and the ***sample*** was kept at a ***temperature lower than 10 celsius***?

-- Please --

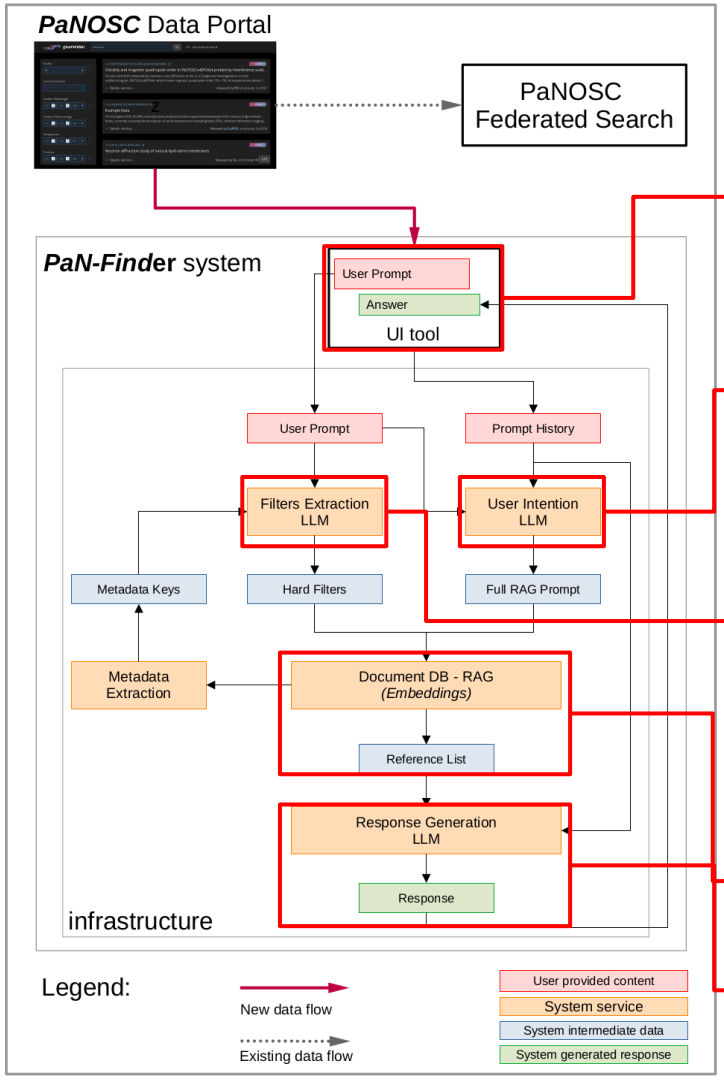
Tools

ML, AI, LLM, Embeddings, NLP, and much more...

What else?

Recommending system helping user to discover what is available

Infrastructure



Interactive Prompt UI

User Intention

Find the ***50*** most relevant ***datasets*** related to ***pharmacology effects*** on ***lung tissues*** infected with ***Covid***, where the ***incident wavelength*** was between ***1.9 and 2.5 Angstrom*** and the ***sample*** was kept at a ***temperature lower than 10 celsius***

Hard filters

```
{
  "number of documents": 50
  "where": { "and": [
   {
    "name": "incident_wavelength",
    "value": { "between": [ 1.9, 2.5 ] },
    "unit": "angstrom",
   },
   ...
  ]}
}
```

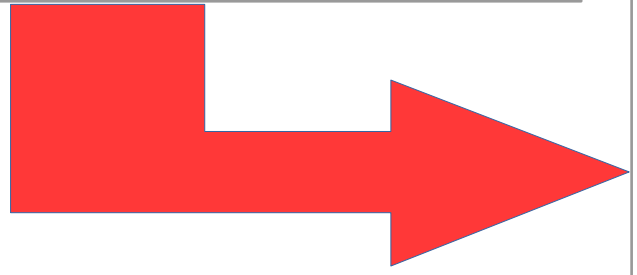
Results provider

Answer formulation

System answers

User query

Find the ***50*** most relevant ***datasets*** related to ***pharmacology effects*** on ***lung tissues*** infected with ***Covid***, where the ***incident wavelength*** was between ***1.9 and 2.5 Angstrom*** and the ***sample*** was kept at a ***temperature lower than 10 celsius***



System answer

I found a list of **50 datasets** related to the **pharmacology effects** on **lung tissues** infected with with **COVID**, matching the conditions:

- **Incident wavelength between 1.9 and 2.5 Angstrom**
- **Sample temperature lower than 10 Celsius**.

Here are datasets including the relevancy score:

- Dataset 1, DOI 1
- …
- Dataset 50, DOI 50

I also found relevant information in the following resources:

- Resource 1 (URL 1)
- Resource 2 (URL 2)

Paradigm Shift

PaNOSC Data Portal

https://data.panosc.eu

Federated Search

RI Infrastructure

PaNOSC Local API

Data Catalogue

Decentralized

OSCARS PaN-Finder

Information Collection

Training

Deployment Production

RI Infrastructure

PaNOSC Local API

Public API

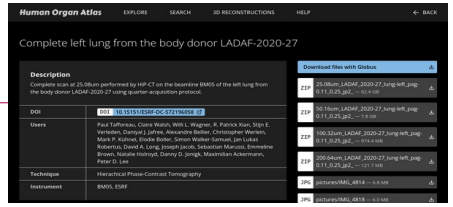
Data Catalogue

Centralized

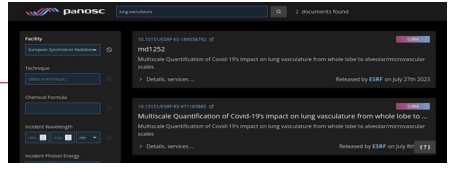
First Task

Information Collection

Local Storage

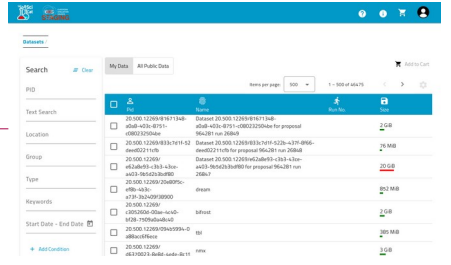


Thematic portals



Cross domain portals

Publications



Facilities portals

First Challenges

Which data sources should we use?

- PaNOSC data portal
- Institution portal
- Other sources
- Institution insiders

Should we use one or many? How many?

Which information should we include?

How can we be sure that we have enough?

Sustainability

How can we find resources to:

- maintain and upgrade infrastructure
- keep the body of knowledge updated
- explore new solutions: LLM, AI, MLP, ...
- coordinate effort between all stakeholders

Benefits

Collaboration between iCAT and SciCat

- Unified exchange format
- Dedicated exchange endpoints
- Unified data discovery method
- Existence validation method
- What else?

Thanks

Questions? Ideas? Feedback?

Massimiliano (Max) Novelli
max.novelli@ess.eu
Senior Data Curation Scientist
European Spallation Source, DMSC
Copenhagen, Denmark