eosc | FAIRCORE4EOSC
Core Components Supporting a FAIR EOSC

EOSC Research Discovery Graph Service (RDGraph)

CRIS2024 pre-workshop: Hands-on Research Data, 14 May, 2024

Thanasis Vergoulis, OpenAIRE

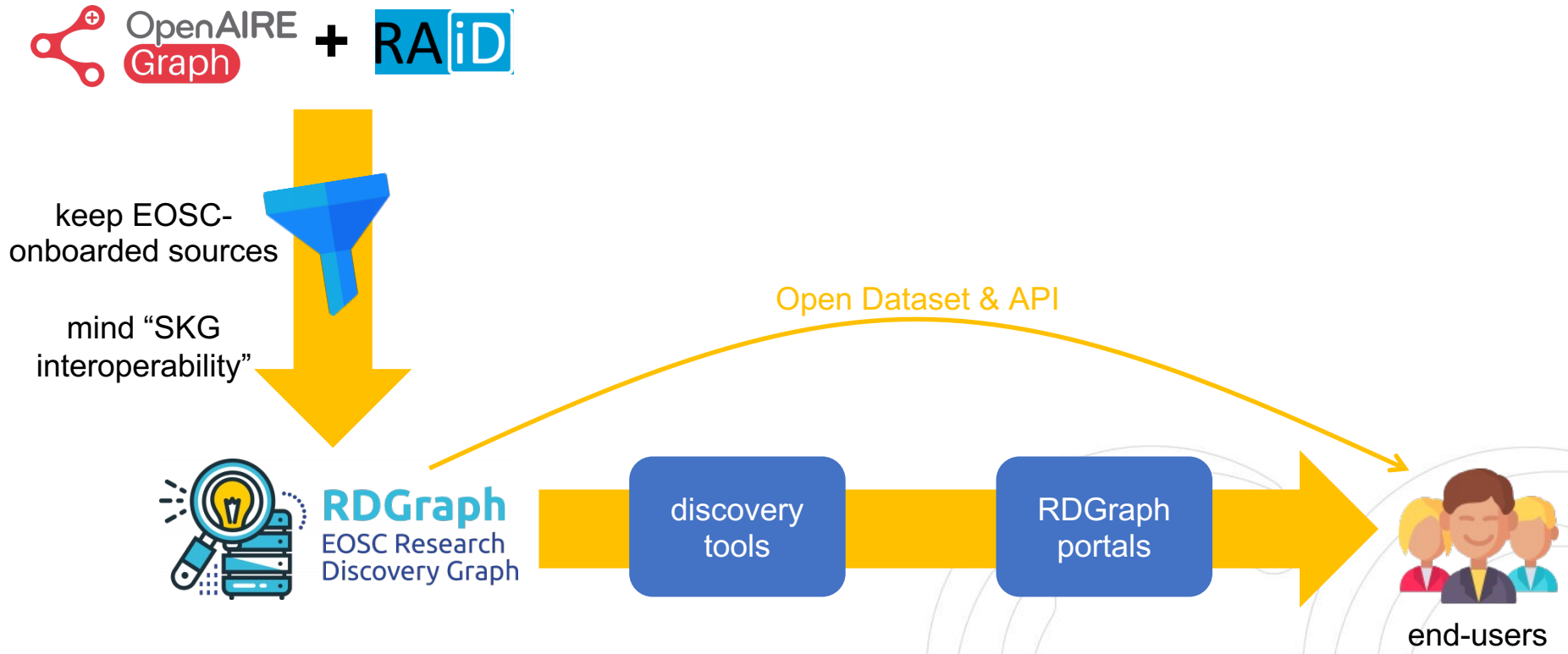
Funded by
the European Union

RDGraph: motivation

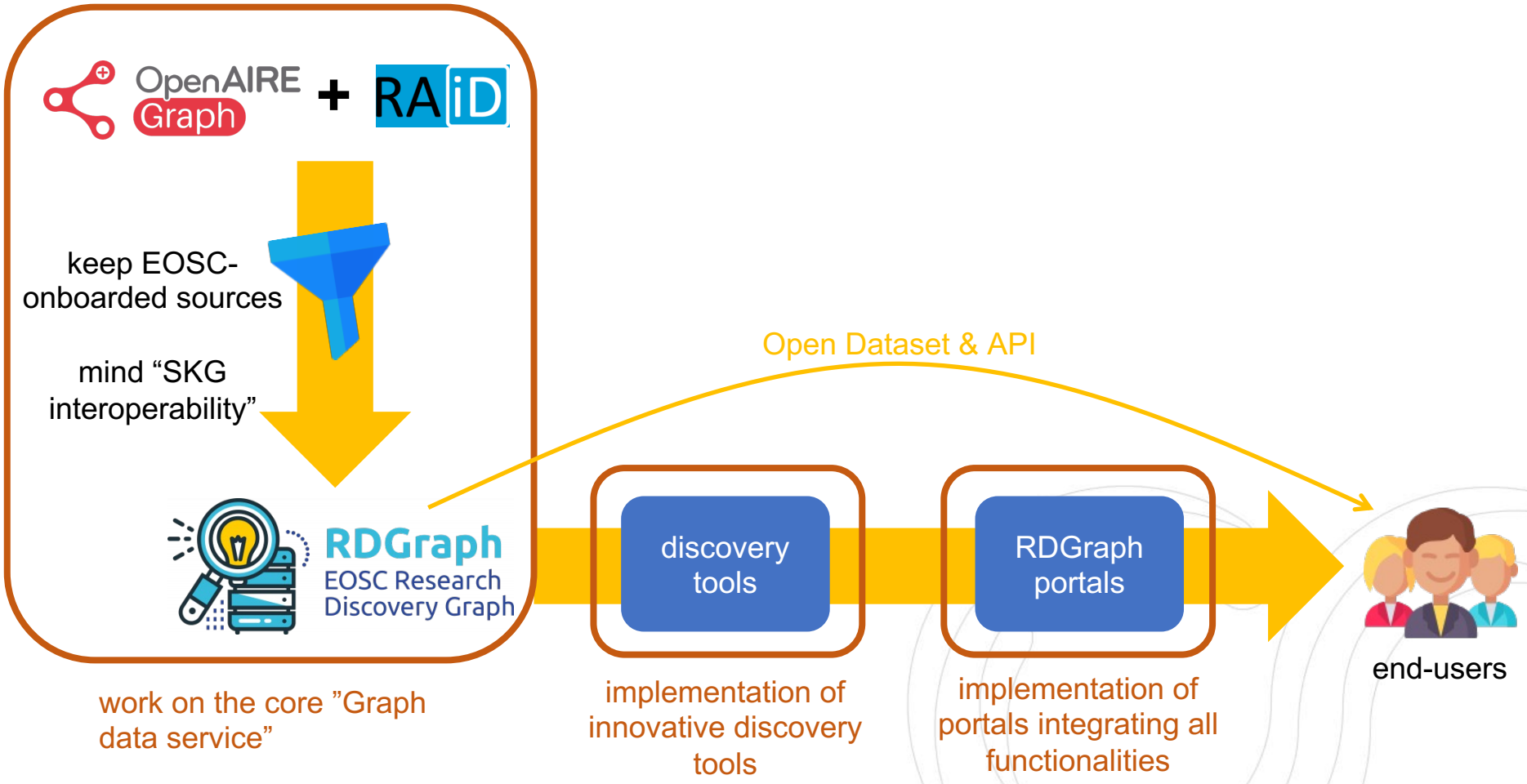
exponential growth of scientific output

difficult to identify really valuable research

RDGraph: the concept



RDGraph: the concept



Graph data service

What has been achieved

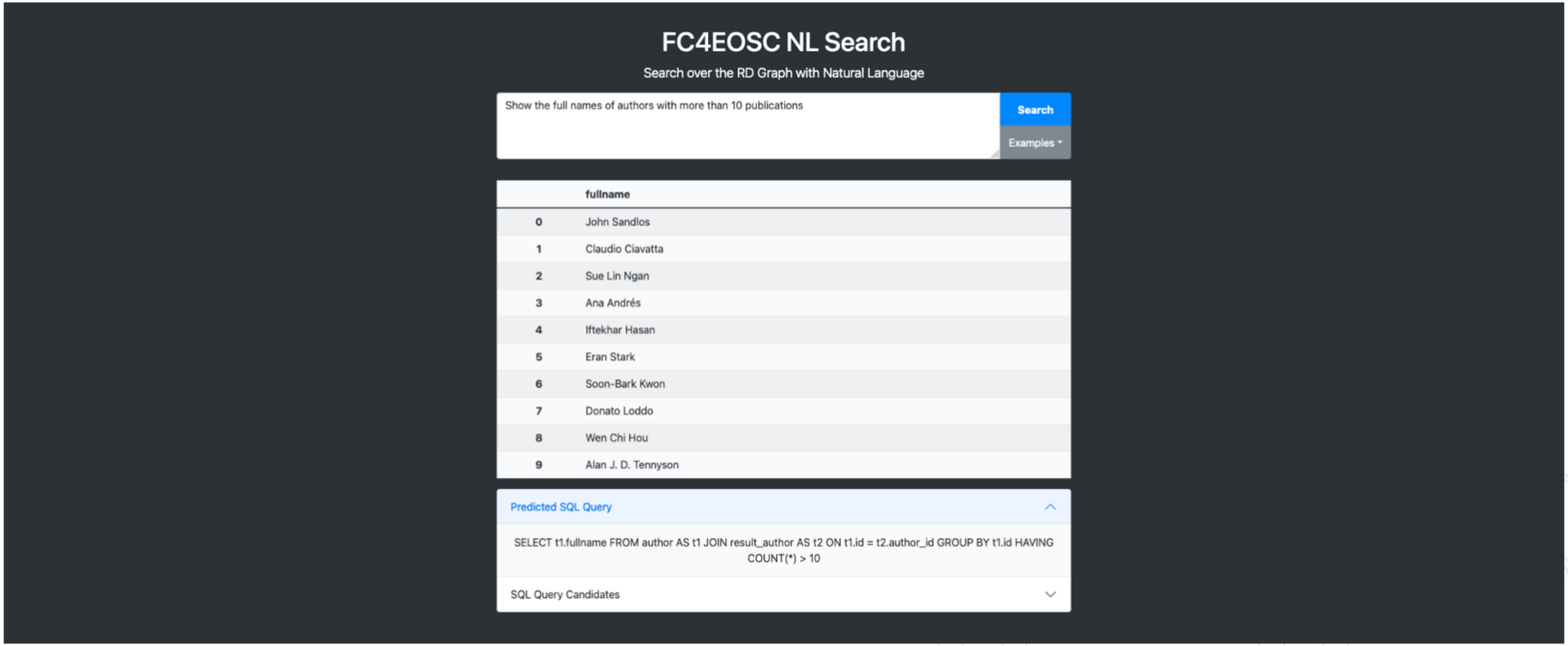
- Workflow for creating SKG-IF compliant RDGraph [1]
  - Dataset soon to be deposited on Zenodo
    - Currently available on gDrive [2]
- Community-specific sub-datasets supporting the activities in community recommendations

Next steps

- Import of RAiDs (exploit **Datacite)**
- API to become more compliant to SKG-IF

[1] https://skg-if.readthedocs.io/en/latest/products.html
[2] https://drive.google.com/drive/folders/1mMuJPTAI346_OWqCBRsdO5o_1krfOLAZ?usp=sharing

Natural language search



Natural language search

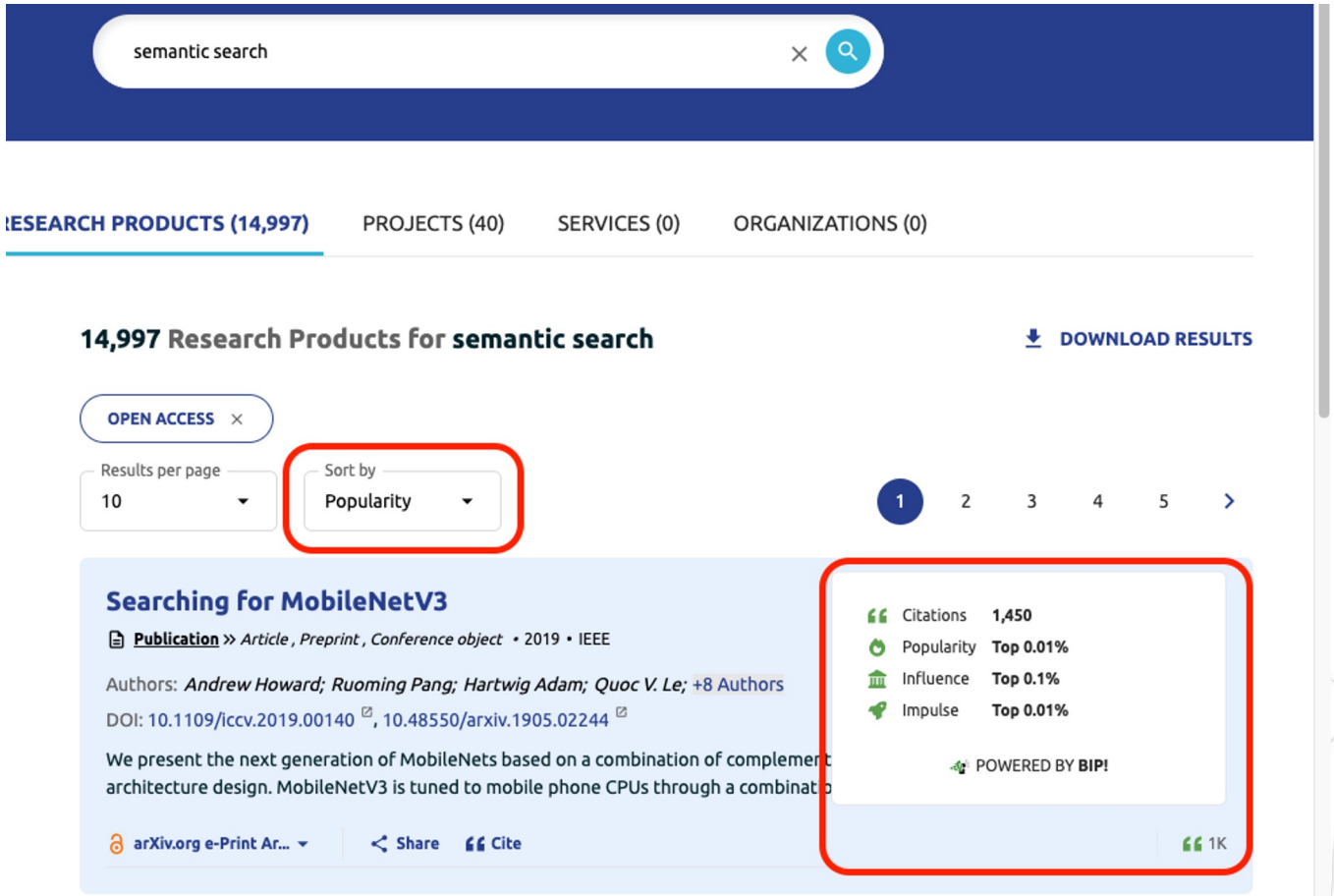
What has been achieved

- Creation of the NL Search component, using a Pre-trained Language Model, fine-tuned for Text-to-SQL
  - Fine-tuning done with the Spider multi-domain text-to-SQL dataset
- Creation of a complete pre/post-processing pipeline to optimize the use of the PLM
  - Handling and processing the necessary inputs of the model (question, DB schema, data)
  - Automatic correction and refinement techniques to improve the quality of generated SQL queries
- Connection of the component to a relational dump of the RDGraph, using sample communities
- Deployment of a demo and API to allow users and developers to access the service

Next steps

- Training specifically for the RD Graph: Improve performance for our use cases
- Update the RD Graph instance with data and communities more relevant to the project
- Integration to the RDGraph Portals

Impact-based search



Impact-based search

What has been achieved

- Impact indicators (influence, popularity, impulse, citation count) for all research products are computed within the content provision pipeline of the Graph
- Impact indicators can be used to rank the results of the keyword-based search in the RDGraph portal
    - E.g., Retrieve publications for the keywords “Semantic search” and sort the results based on their number of citations by descending order
- Impact indicators can be used to filter the results of the keyword-based search in the RDGraph API
    - E.g., Retrieve only publications in Top 1% based on their calculated popularity score

Next steps

- Combine keyword-based relevance with impact indicators to rank keyword search results

Community recommendation profiles

What has been achieved

Developed three RDGraph services

S1: Author-centric research product recommendations (training using RDGraph's citation edges)

S2: Research field-specific product recommendations (online learning via user feedback)

S3: Similarity-based research product recommendations (training using textual data & metadata of research products)

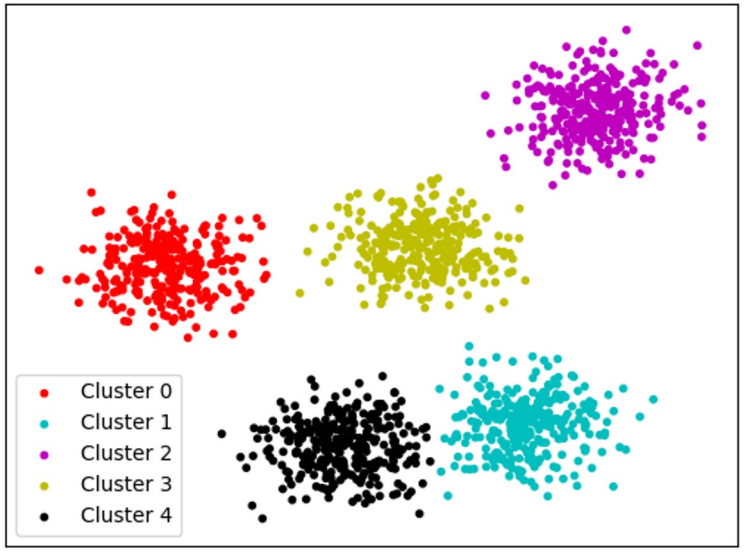
Next steps

- Integrate services into RDGraph's UI
- Enhance S3's performance and memory efficiency for scaling to 10M+ resources
- S4: Develop a community-based research product recommendation service

RAiD inference service

OpenAIRE Graph

Metapath2Vec
DeepWalk



Including meaningful relationships for a RAiD

- research products versions
- institutions
- projects
- authors

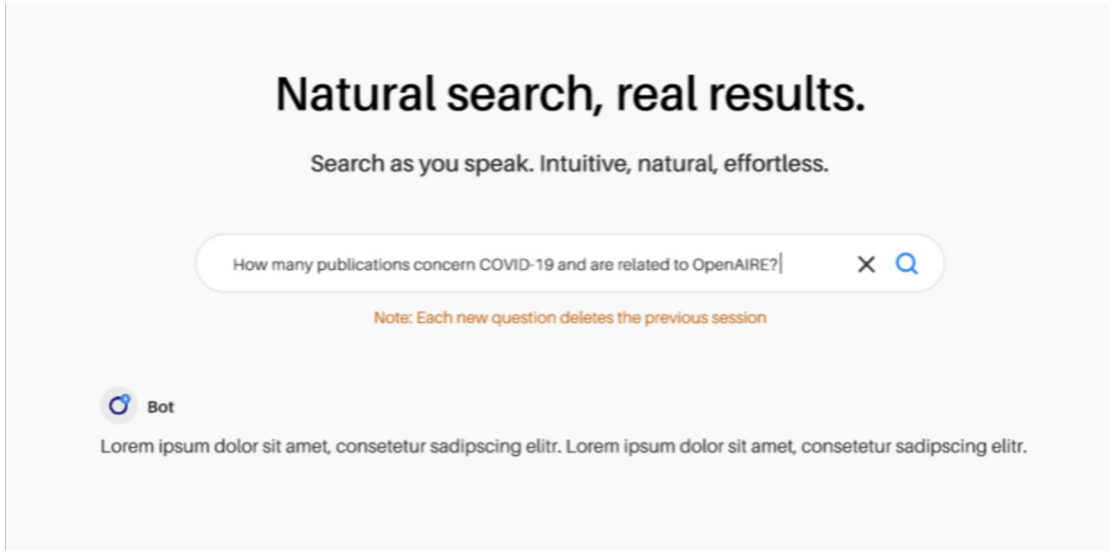
- Each cluster represents a RAiD

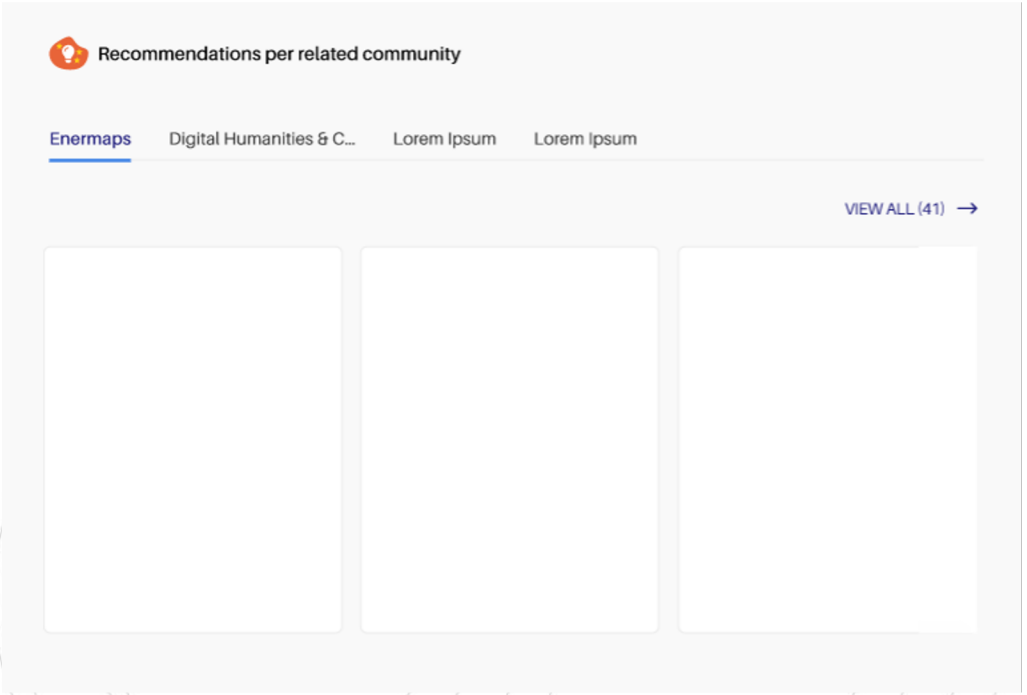
Based on a 2-stage process

- stage 1: creation of research product N-dimensional embeddings
- stage 2: clustering of embeddings to create RAiDs

first batch of recommendations: by the end of June

RDGraph portals





Thank you!

thanasis.vergoulis@openaire.eu

Also do not miss the CRIS2024 Tutorial titled “**Facilitating the discovery and identification of EOSC resources**” (Thursday 10:30am-12:00am CEST)