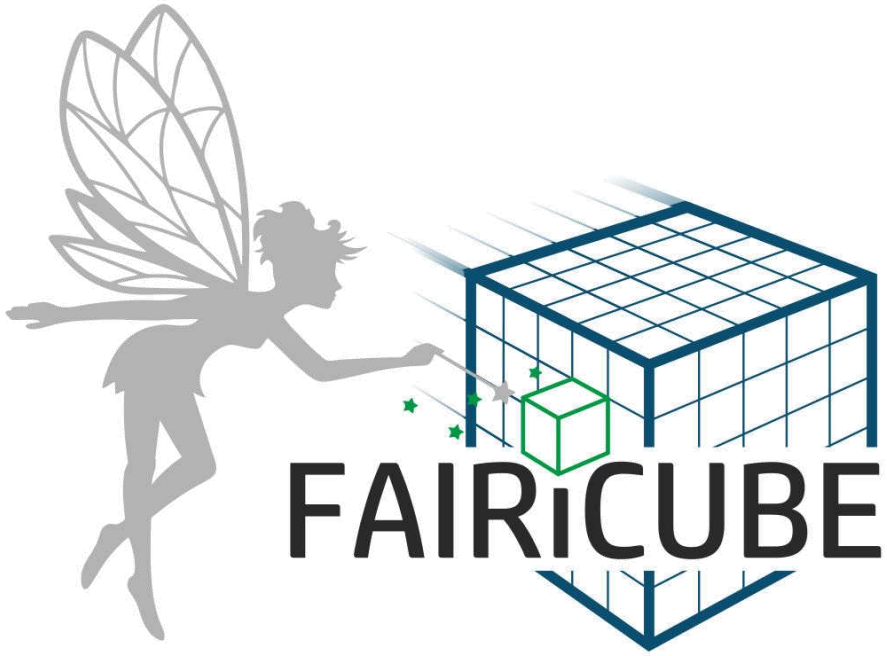


BRAINSTORMING ON A ROADMAP FOR FUTURE DEVELOPMENTS OF THE KB SERVICES & METADATA PIPELINE

EPSIT, 21/05/2024

BRAINSTORMING ON A ROADMAP FOR FUTURE DEVELOPMENTS OF THE KB SERVICES & METADATA PIPELINE

- Inclusion of computational resources information in a/p resources metadata
- Query Tool upgrade to query both a/p resrources and datasets metadata

INCLUSION OF COMPUTATIONAL RESOURCES INFORMATION IN A/P RESOURCES METADATA

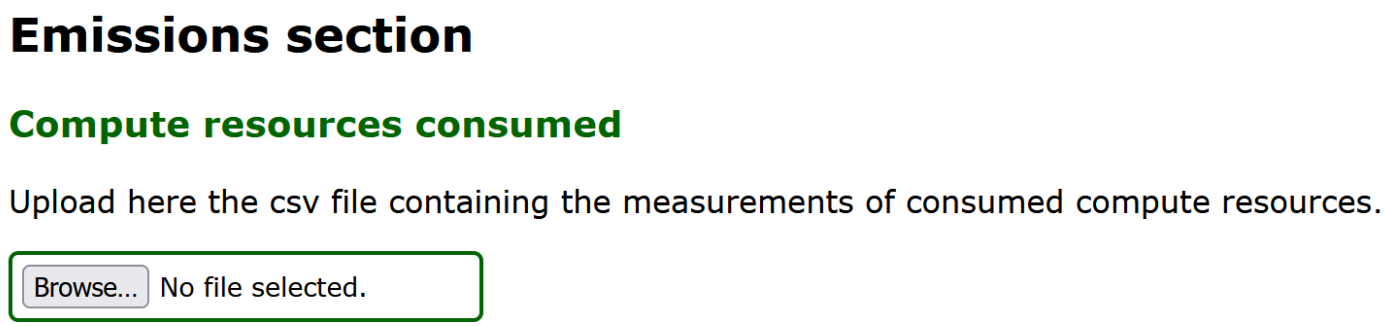
	Measure	Value
2	Data size in grid points	50000x32x32x3
3	Largest allocated array in grid points	[50000, 32, 32, 3]
4	Data size (MB)	0.07
5	Data read (MB)	0.12
6	Data written (MB)	2.24
7	Main memory available (MB)	31401.47
8	Main memory consumed (MB)	2581.38
9	Sum of allocated variable sizes (MB)	1406.26
10	Description of CPU/GPU	Machine type: x86_64 Processor type: x86_64 Number of physical cores: 4 Number of logical cores: 8 CPU frequency: 3.0432 GHz No GPU available
11	Wall time (s)	428.51
12	Energy consumed (W)	6.39
13	CO_2-equivalents [CO_2eq] (g)	2.46
14	Network traffic (MB)	22.8
15	Programming language	Python
16	Essential libraries	time os tensorflow numpy

CSV example

- A first version of a tool capable to collect, at the end of the execution of an a/p resource, information about the computational resources spent and the characteristics of the executing machine, has been developed and tested. The information is stored in a csv file, that can be uploaded during the a/p resource metadata editing/ingestion.
- Currently, the csv file uploaded via the md-form, is stored in the KB-DB. The intention is to find a way to better exploit this information.
- Idea is to visualise the information stored in the csv files, adding a column in the query tool results to display the presence of csv files.

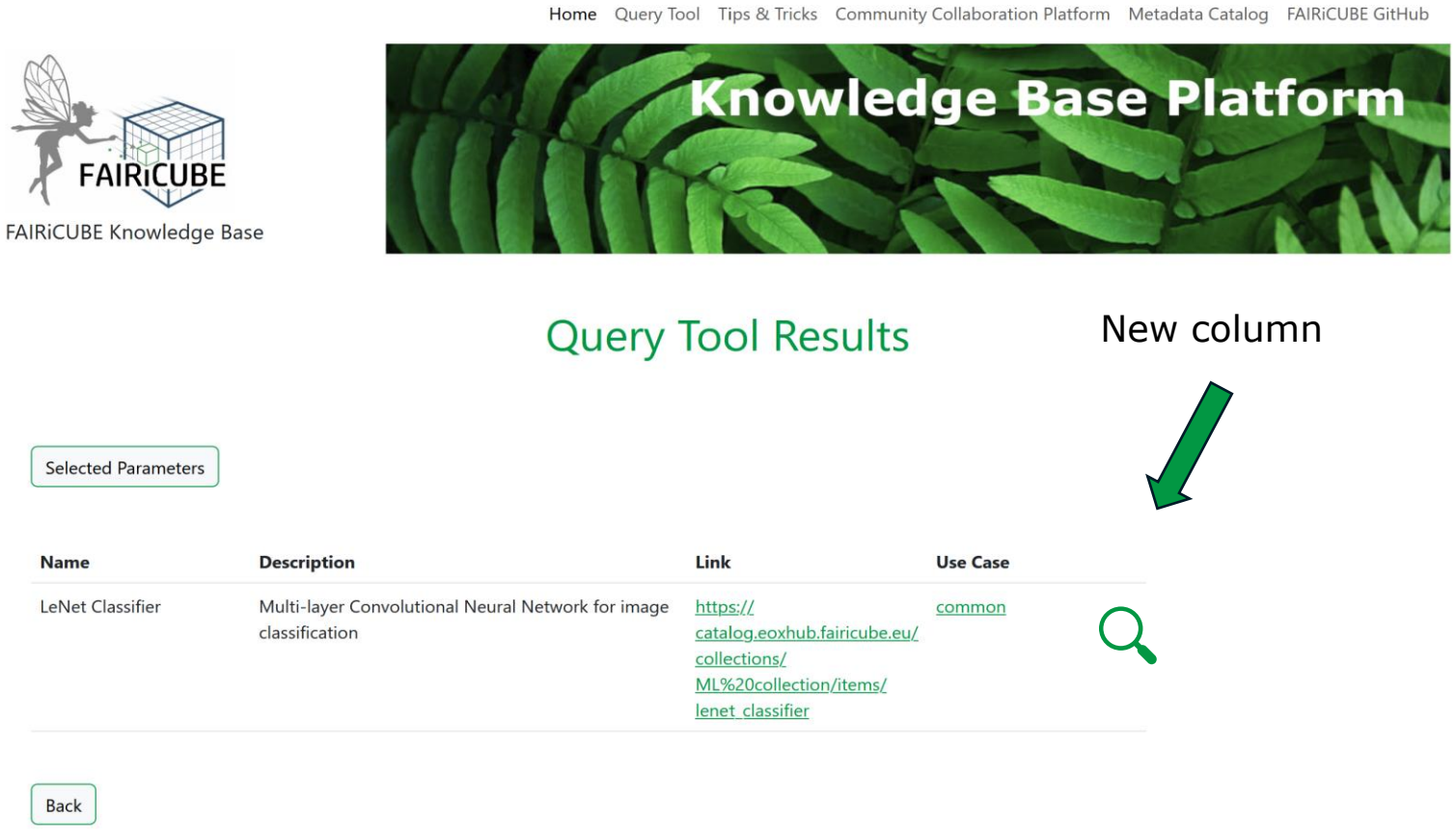
INCLUSION OF COMPUTATIONAL RESOURCES INFORMATION IN A/P RESOURCES METADATA

- Ingestion in the Knowledge Base DB, using the md-form, of the csv files containing information about the resources consumed

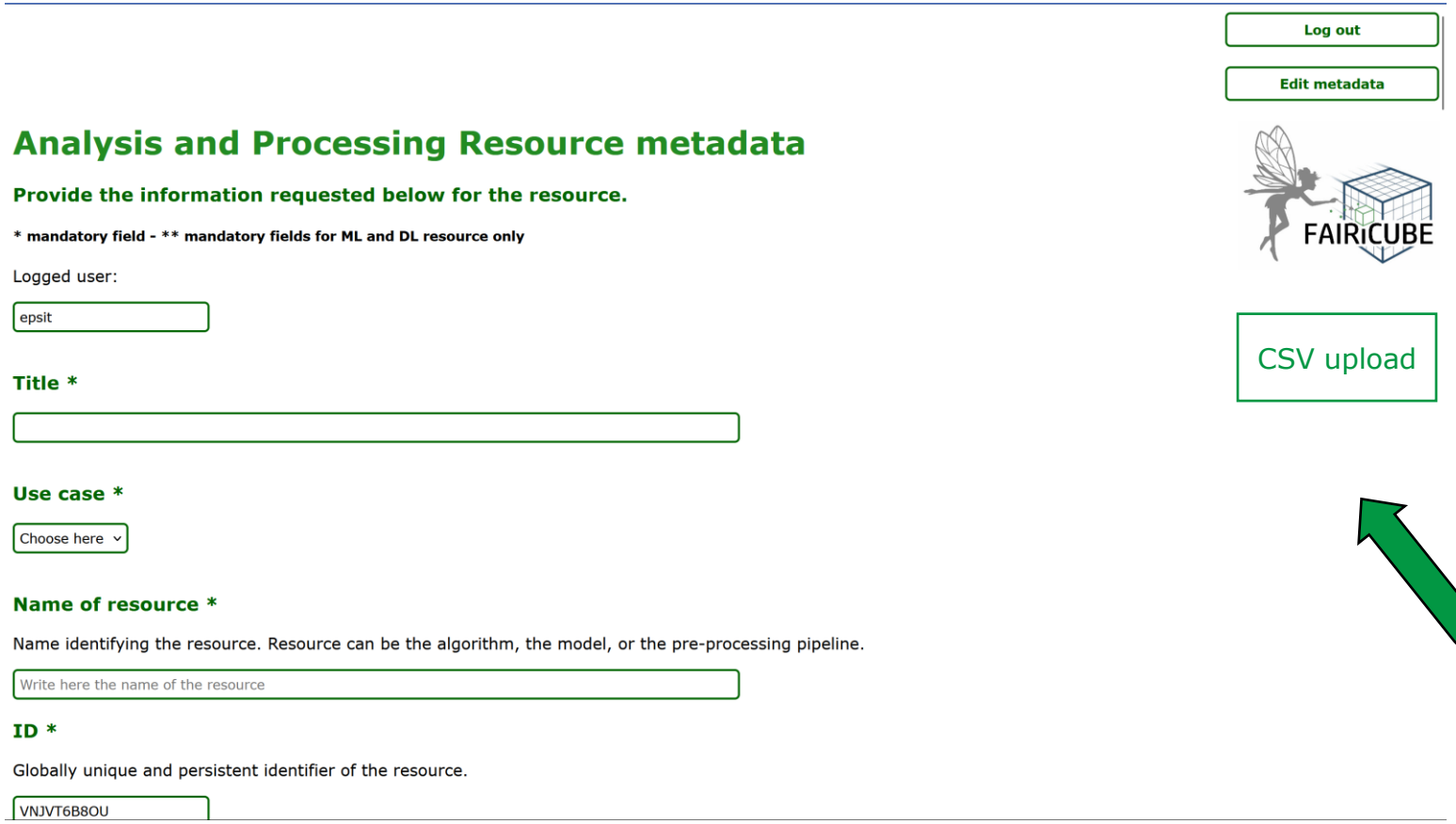


INCLUSION OF COMPUTATIONAL RESOURCES INFORMATION IN A/P RESOURCES METADATA

- Addition of a column in the query tool results to show the presence of csv files (and therefore the possibility to visualise the information on consumed resources)



INCLUSION OF COMPUTATIONAL RESOURCES INFORMATION IN A/P RESOURCES METADATA



- Because several csv files can be generated for each a/p resource (depending on the execution configuration and also by users different from a/p resource metadata editors), idea is to allow upload of multiple csv files, creating a new section in the md-form.

CONSIDERATIONS

- Do we agree that it is enough to offer only the visualisation of the computational resources and not even the possibility to query them (via predefined and/or customised queries)?

QUERY TOOL UPGRADE TO QUERY BOTH A/P RESOURCES AND DATASETS METADATA

- WHY?

- To allow retrieving information in reply to the following question:

  "Which datasets for biodiversity UCs and which algorithms on those datasets have been used?"

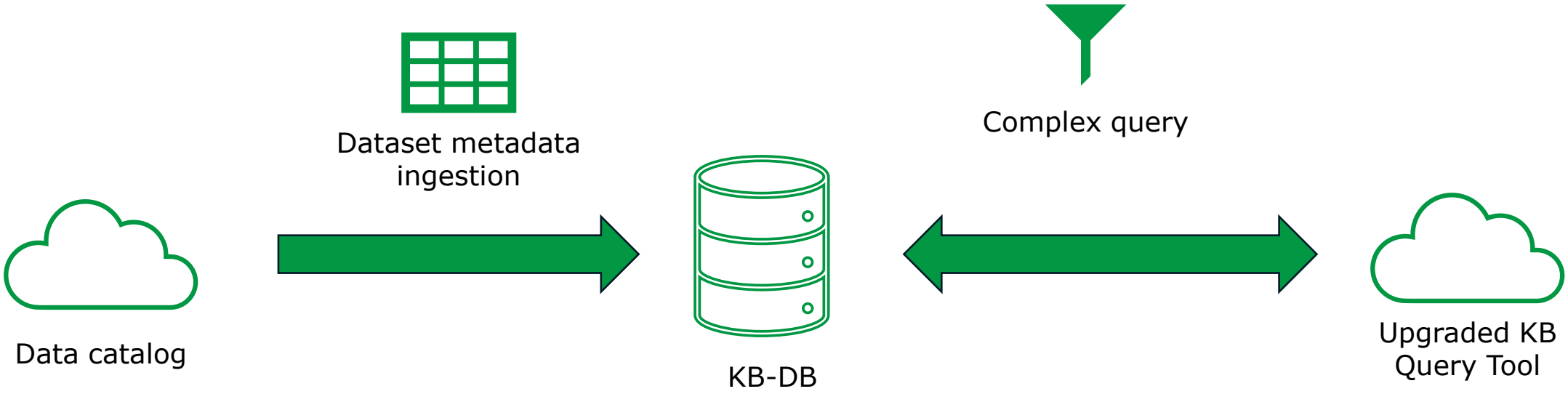
QUERY TOOL UPGRADE TO QUERY BOTH A/P RESOURCES AND DATASETS METADATA

We have envisaged 2 alternative options:

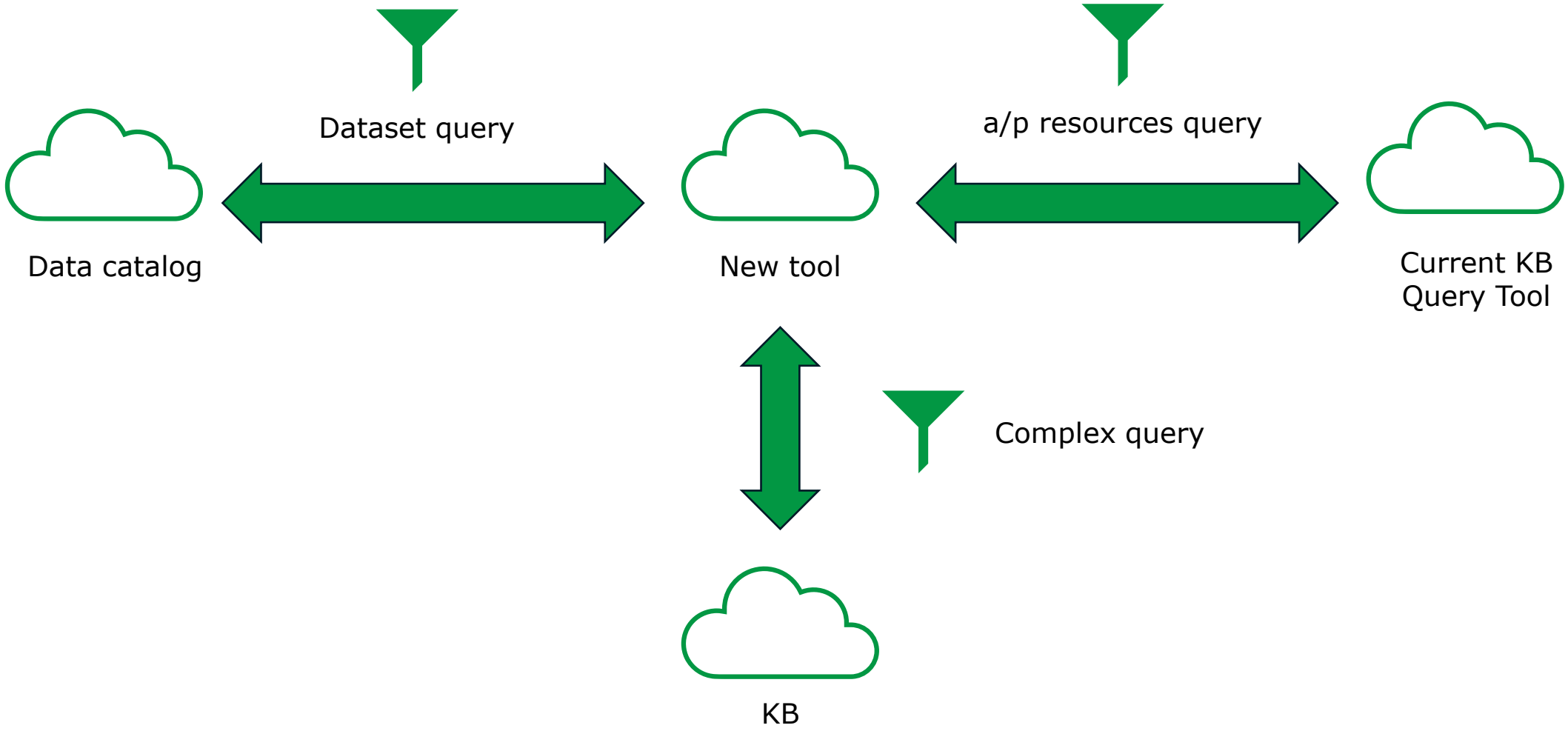
- **Option a**
  - Ingest in the KB database all datasets metadata present in the catalog (creating an ad-hoc routine to be launched periodically)
  - Up-grade the QT to enable the execution of “cross-table” queries, like the example above

- **Option b**
  - Create a new tool capable to search simultaneously the catalog and the current QT to perform a “combined query”

OPTION A


Dataset metadata ingestion
Complex query
Data catalog
KB-DB
Upgraded KB Query Tool

OPTION B


Dataset query
a/p resources query
Data catalog
New tool
Current KB
Query Tool
Complex query
KB

CONSIDERATIONS

- **Option a**
  - shorter development time
  - need to create a procedure to ingest the datasets metadata in the KB-DB
- **Option b**
  - longer and more complex to develop (and by whom?)