SciOps 2022

Astrophysics with INODE

Dr. Srividya Subramanian
MPE, Garching

Limitations of Existing Data Exploration Tools

Input

- Limited Query Exploration Capabilities
- Knowledge of SQL (or SPARQL, etc)
- Knowledge of the database schema
- Well-formed information needs

Output

- No interpretation of results
- No explanation of system choices/answers
- No clue how to proceed next

Data

- Static, known Schema
- Hard to understand the attribute names and foreign key relations
- Hard to link and query new, but related databases

[1]SQL = Structured Query Language for relational databases
[2]SPARQL = SPARQL Protocol and RDF Query Language for graph databases

BIG DATA …

SDSS, HETDEX, 4MOST,
DESI, PFS, EUCLID

Open Questions …

- How did the Universe begin and how will it end ?
- Galaxy formation and evolution
- Structure of the Universe
- Dark matter
- and many many more ……..

BIG DATA …

SDSS, HETDEX, 4MOST,
DESI, PFS, EUCLID

Data is the new oil …
we need the right tools to leverage it !

Open Questions …

- How did the Universe begin and how will it end ?
- Galaxy formation and evolution
- Structure of the Universe
- Dark matter
- and many many more ……..

BIG DATA …

SDSS, HETDEX, 4MOST,
DESI, PFS, EUCLID

Data is the new oil …
we need the right tools to leverage it !

Open Questions …

- How did the Universe begin and how will it end ?
- Galaxy formation and evolution
- Structure of the Universe
- Dark matter
- and many many more ……..

INODE !

INODE - Intelligent Open Data Exploration

- a platform to access to open datasets through NL
- An end to end DE system
- for a wide range of users

Usecases :

Horizon 2020 EU project

Our Partners

zhaw · ATHENA Research & Innovation Information Technologies · Fraunhofer IGD · SIRIS Academy · cnrs · unibz · SIB Swiss Institute of Bioinformatics · Max-Planck-Gesellschaft

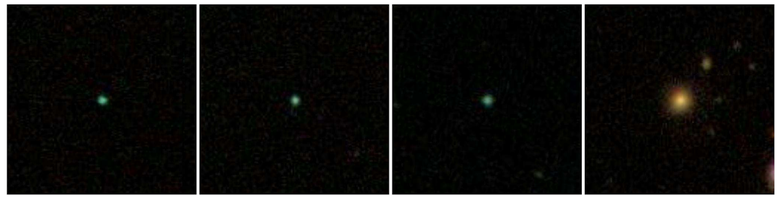
SDSS

- the most detailed three-dimensional maps of the Universe ever made
- Imaging and spectroscopic observations
- DR 16 (SDSS IV)

Catalog - database

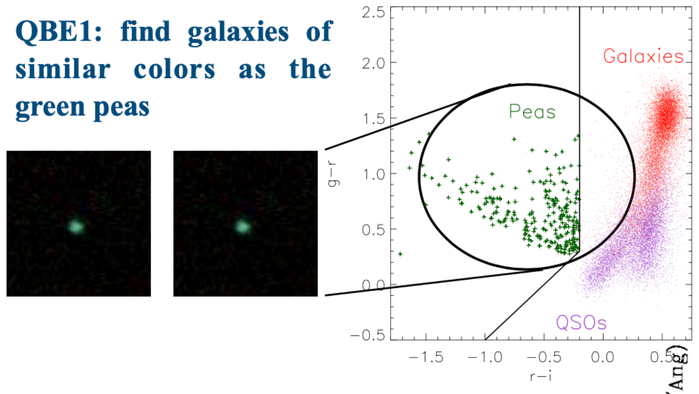
- 3 major categories - photo group of tables, spectro group tables and meta tables (with infos and docs)
- EBOSS, APOGEE-2, MaNGA (including MaStar), BOSS, APOGEE, SEGUE-2, LEGACY, Supernova, SEGUE
- relational database management system (RDBMS), organized in **128 Tables (59 Views)**

Greenpea galaxies



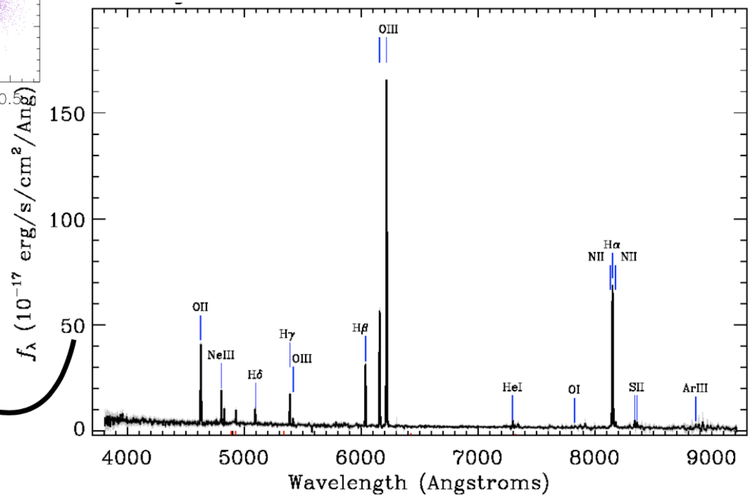
```
SELECT *
FROM PhotoObj AS p
  JOIN SpecObj AS s ON s.bestobjid = p.objid
  JOIN galSpecLine as L ON s.specObjId=L.specObjId
WHERE
  s.class = 'GALAXY'
  AND s.z between 0.11 AND 0.36
  AND p.r >= 18 and p.r <= 20.5
  AND p.petrorad_r < 2
  And p.u-p.r <= 2.5 and p.r-p.i <= 0.2 and p.r-p.z <= 0.5
  And p.g-p.r >= p.r-p.i+0.5 and p.u-p.r  >= 2.5*(p.r-p.z)
  AND oiii_5007_eqw < -100
  AND s.zwarning=0
```

QBE1: find galaxies of similar colors as the green peas

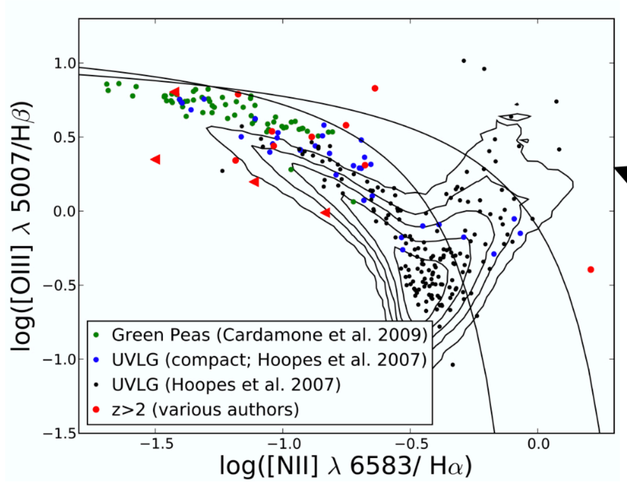


Characterize these galaxies and find similarities in size and redshift etc.

QBE2: Down select the group further according to magnitude, size, color and redshift.

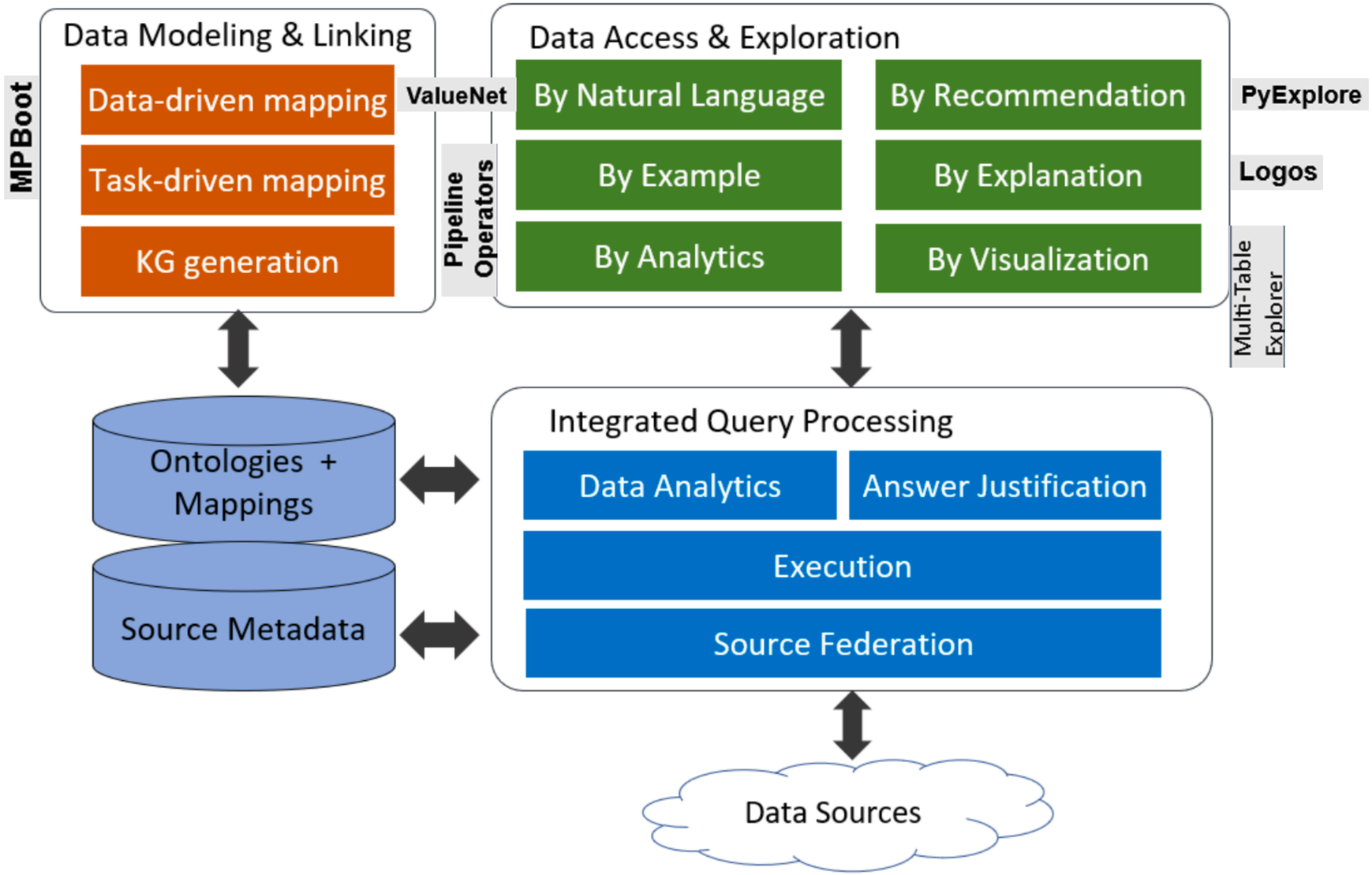


Discover similar galaxies at different redshifts.



Characterize spectral properties like emission line strength, equivalent widths, star formation rates etc.

QBE3: find galaxies of similar relative line ratios and star formation rates.



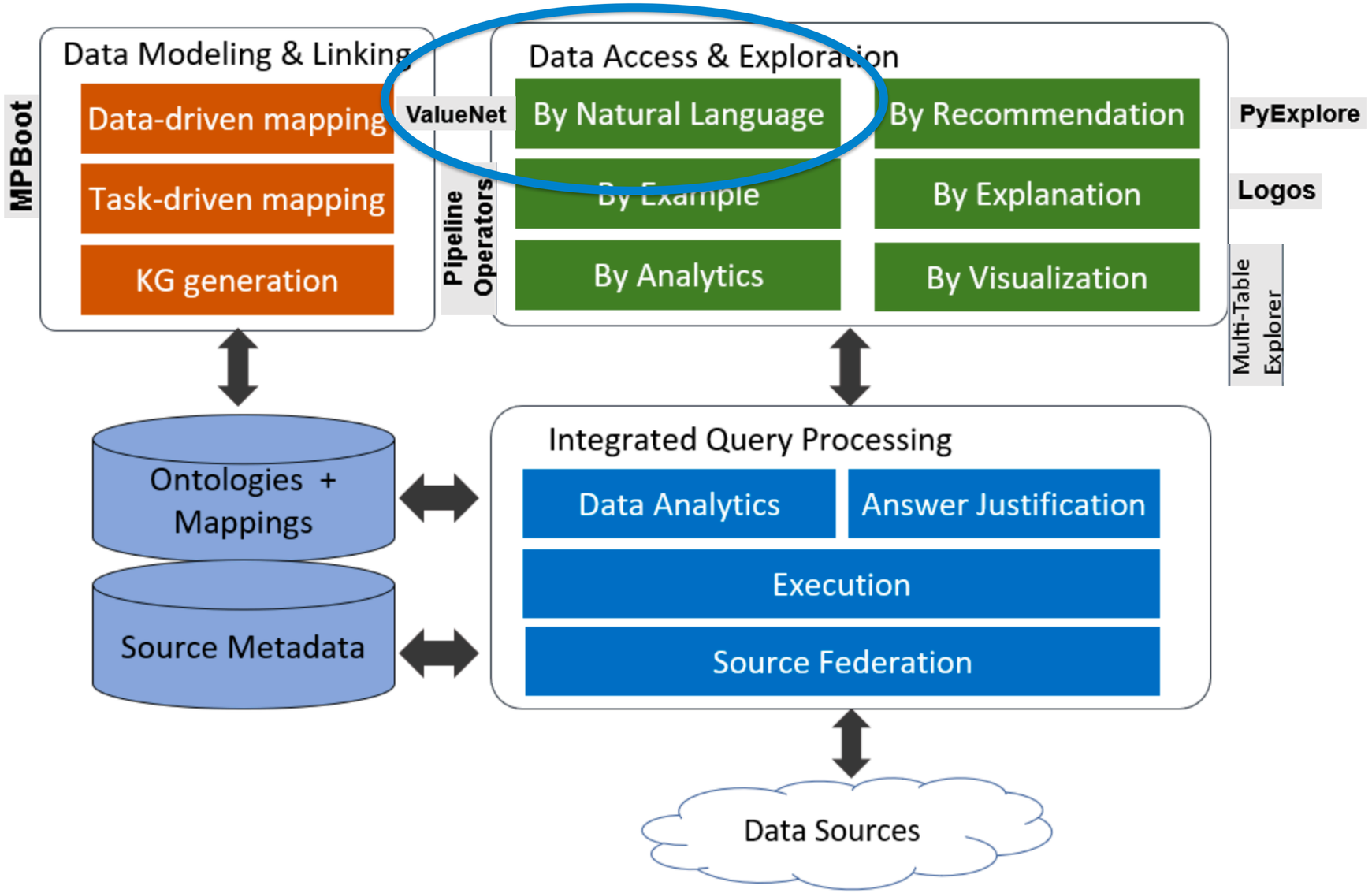
INODE Architecture

Open Data Dialog

Open Data Linking

Backend Services

NL query : Find carbon star



INODE Architecture

Open Data Dialog

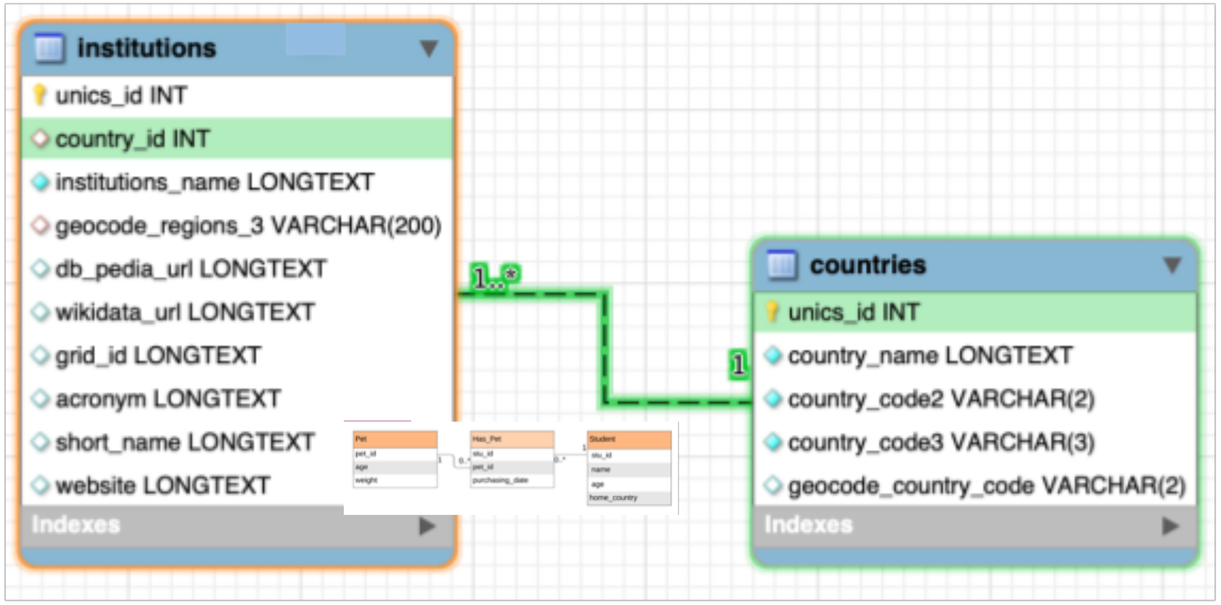
Open Data Linking

Backend Services

Querying a Relational Database in Natural Language

Question:
Find all of the institutions located in Italy.

Schema:

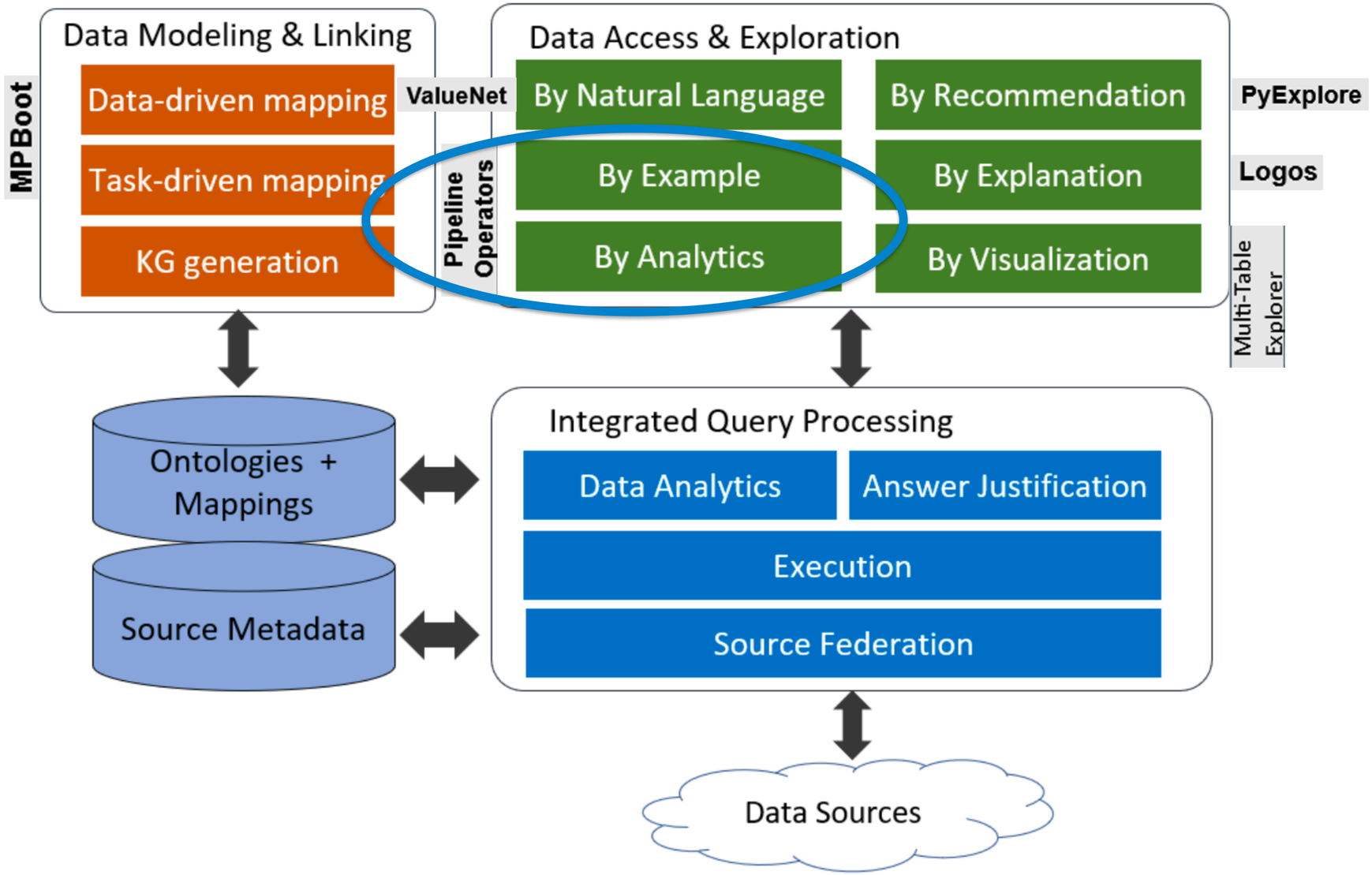


Query:

```
SELECT T1.institutions_name
FROM institutions AS T1
        JOIN countries AS T2 ON T1.country_id = T2.unics_id
WHERE T2.country_name = 'Italy'
```

ValueNet: NL - SQL transformer based Neural Network Architecture

- **Generate SQL** given a natural language question – end to end
- At its core a **neural network** – consisting of an encoder / decoder architecture
- Generates an **intermediate language – SemQL** – which abstracts technical details
- SemQL is **deterministically transformed** to SQL, or any other query language (e.g. SPARQL)
- Uses state of the art **pre-trained transformers** to understand the natural language question.



INODE Architecture

Open Data Dialog

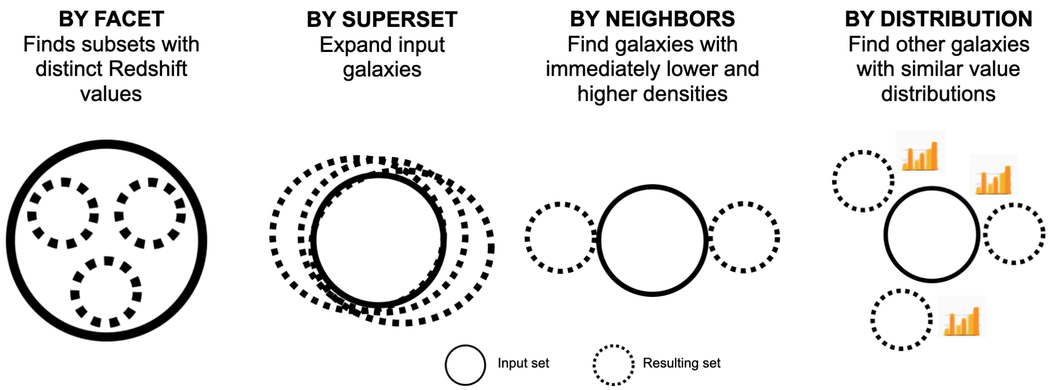
Open Data Linking

Backend Services

How to explore large datasets? Exploration Pipelines

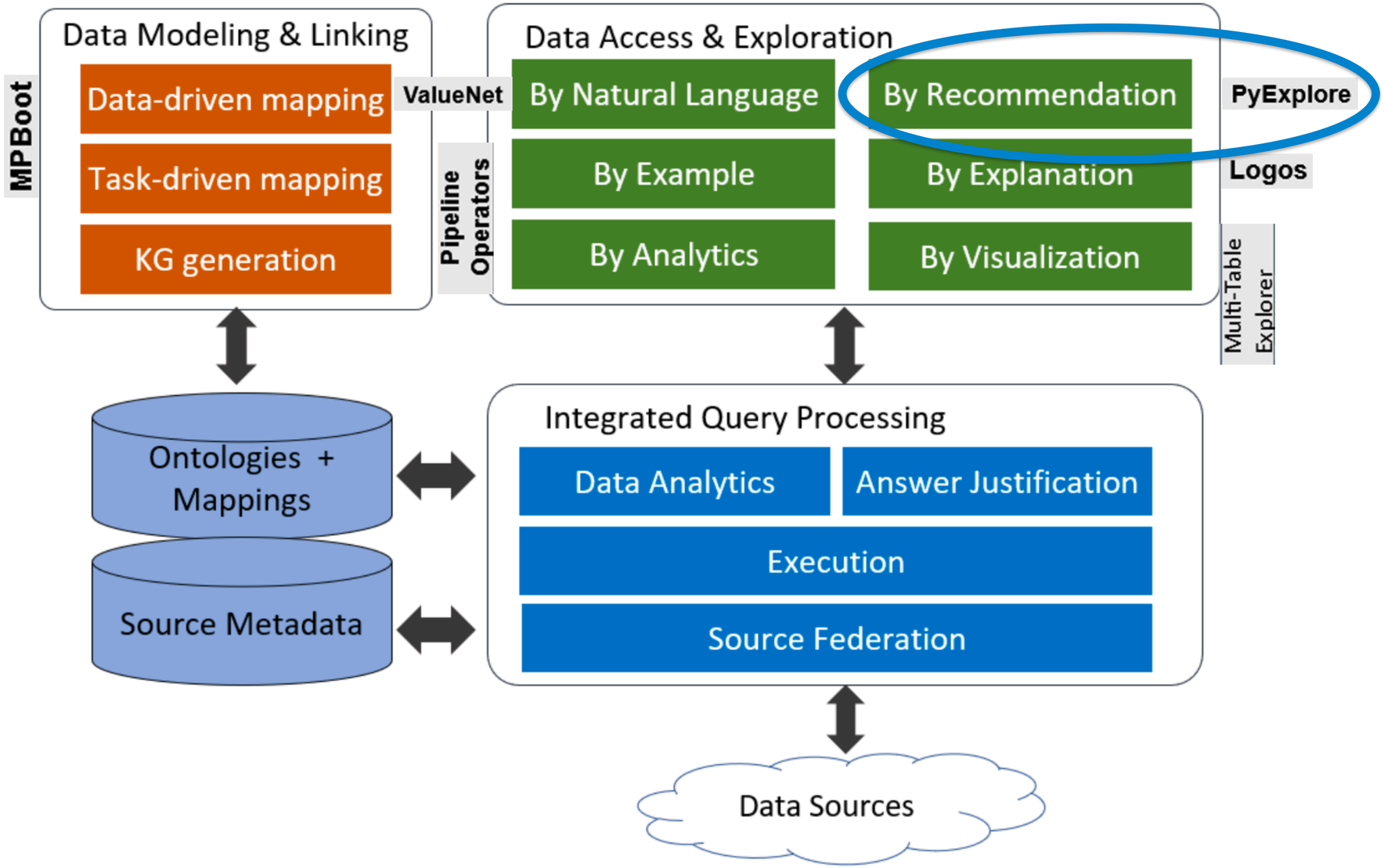
Exploration operators

- Four instance of by-example (example is a set of objects in solid lines)



- A sequence of exploration operators, closed under a set-based semantics
- An item set is defined with a conjunction of predicates

- https://bit.ly/dora_application
- http://www.inode-project.eu:18081/test/galaxies.html



INODE Architecture

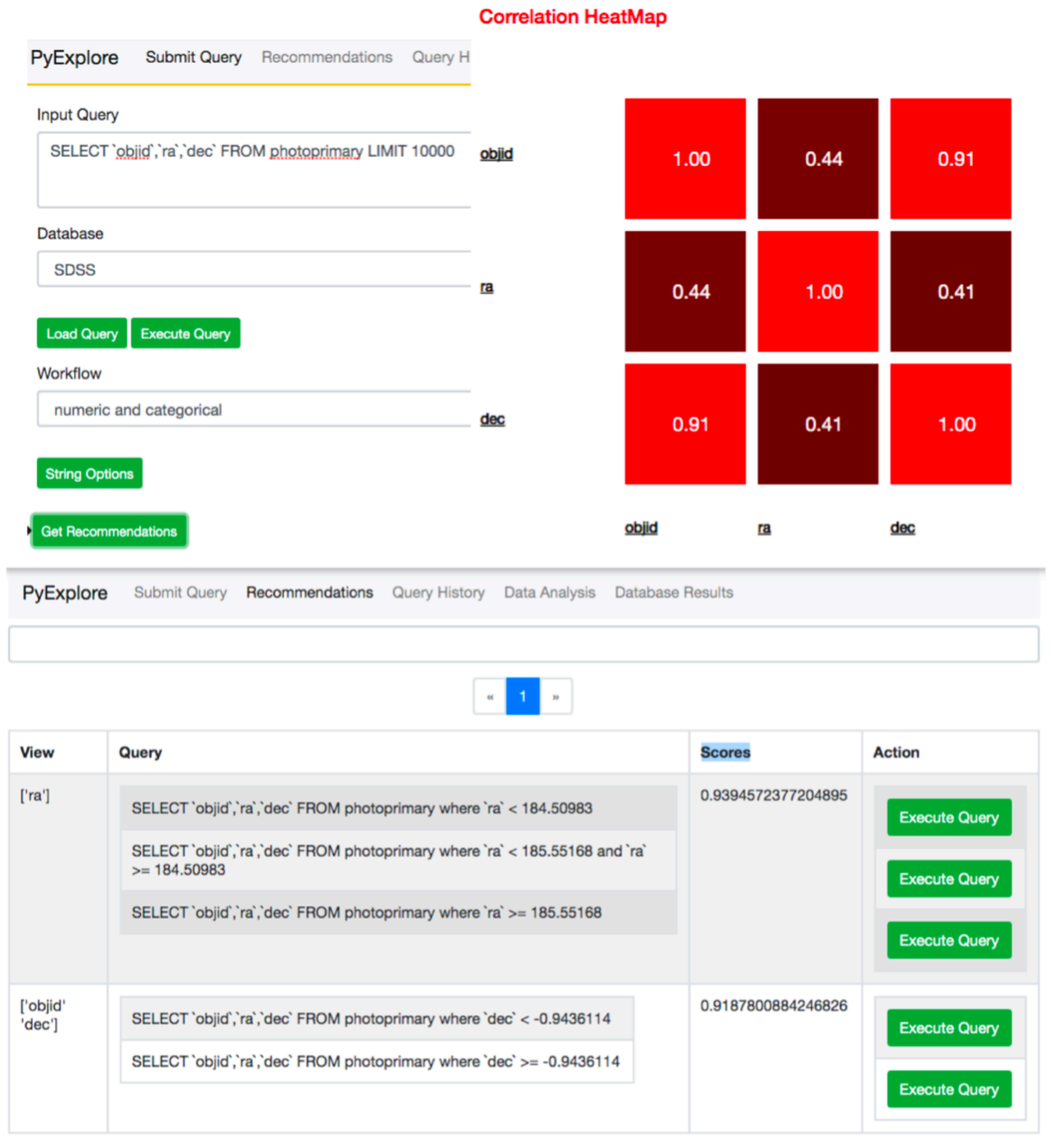
Open Data Dialog

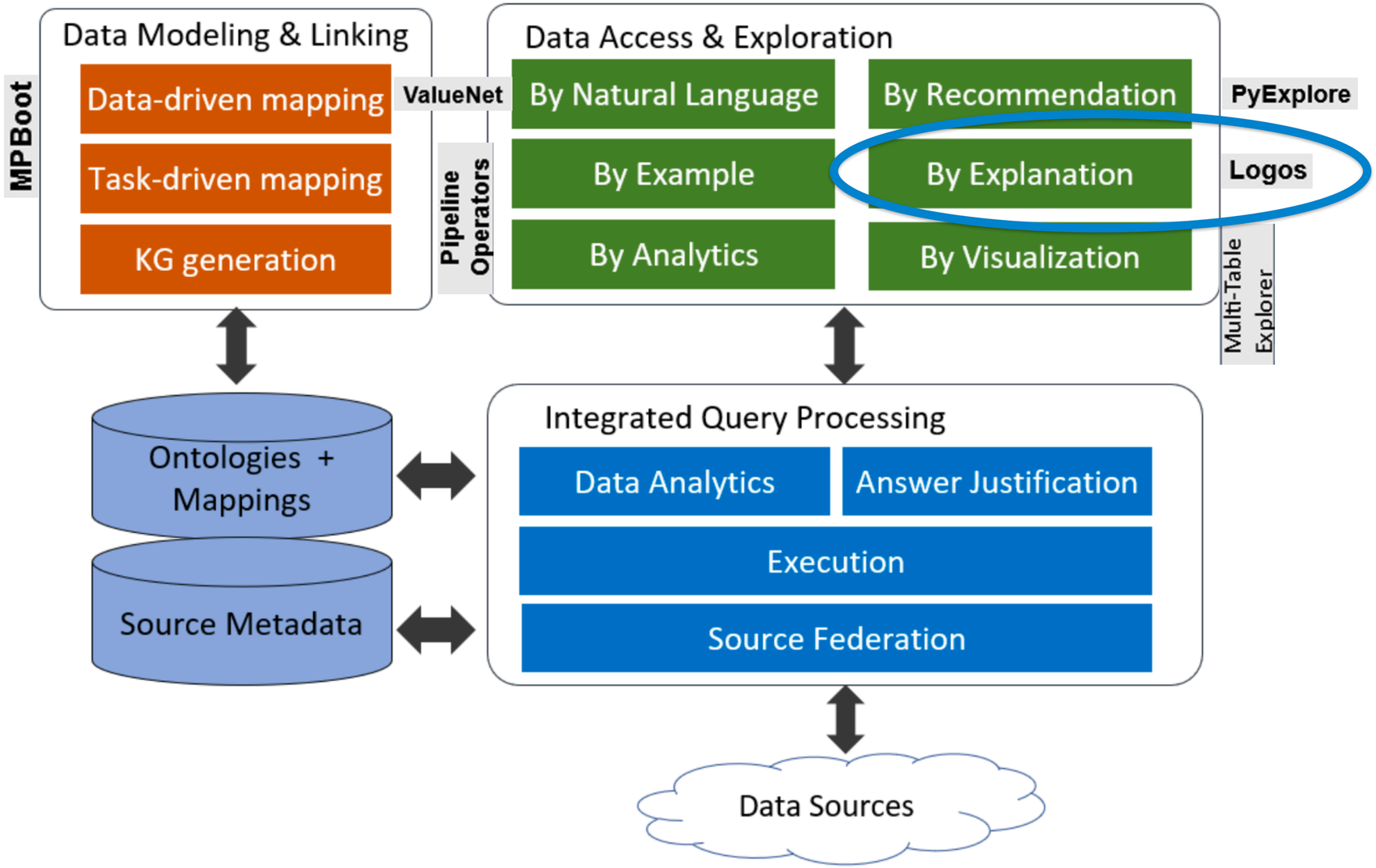
Open Data Linking

Backend Services

PyExplore - query Recommendations for Data Exploration without Query Logs

- ‘interesting’ subsets of query attributes — two notions: attribute correlation and diversity
- Clustering and query generation





INODE Architecture

Open Data Dialog

Open Data Linking

Backend Services

LOGOS : SQL - NL translator

database schema —> graph

nodes —> database relations and attributes
edges —> relationships between the nodes

- annotated with labels in NL

Example :

```
SELECT p.u, p.g, p.r, p.i, p.z FROM specobj s, photoobj p
WHERE s.bestobjid = p.objid AND s.class = 'QSO';
```

- **Logos v.1**: *"Find the u, g, r, i and z of photoobj associated with specobj whose class is QSO."*
- **Logos v.2**: *"Find the magnitude u, magnitude g, magnitude r, magnitude i and magnitude z of photometric objects corresponding to spectroscopic objects whose class is QSO."*

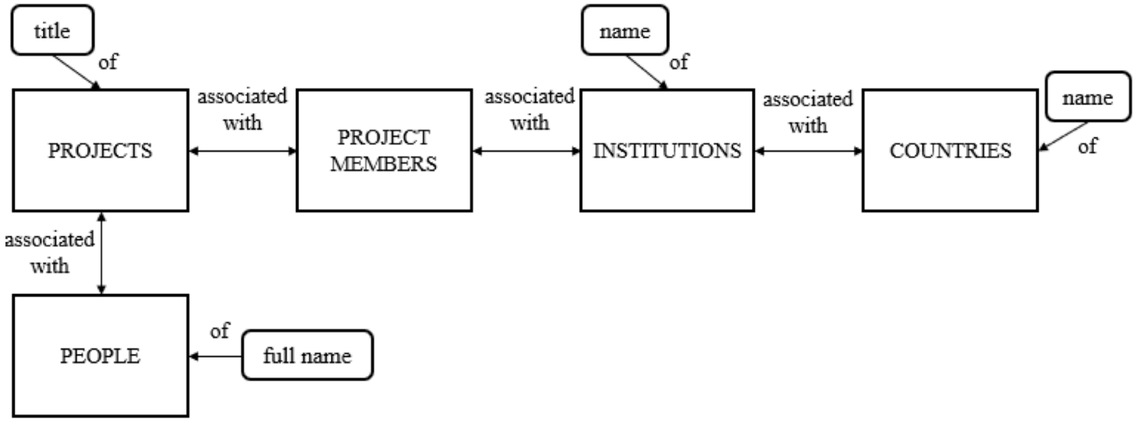


Figure 1: A subgraph of the CORDIS database graph.

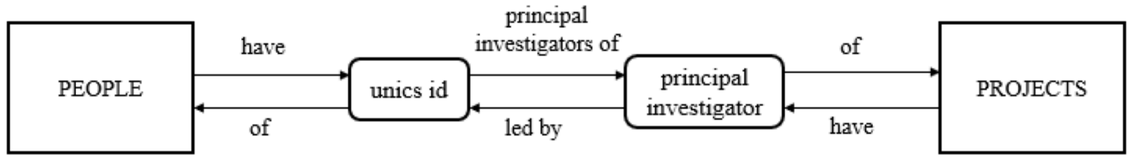
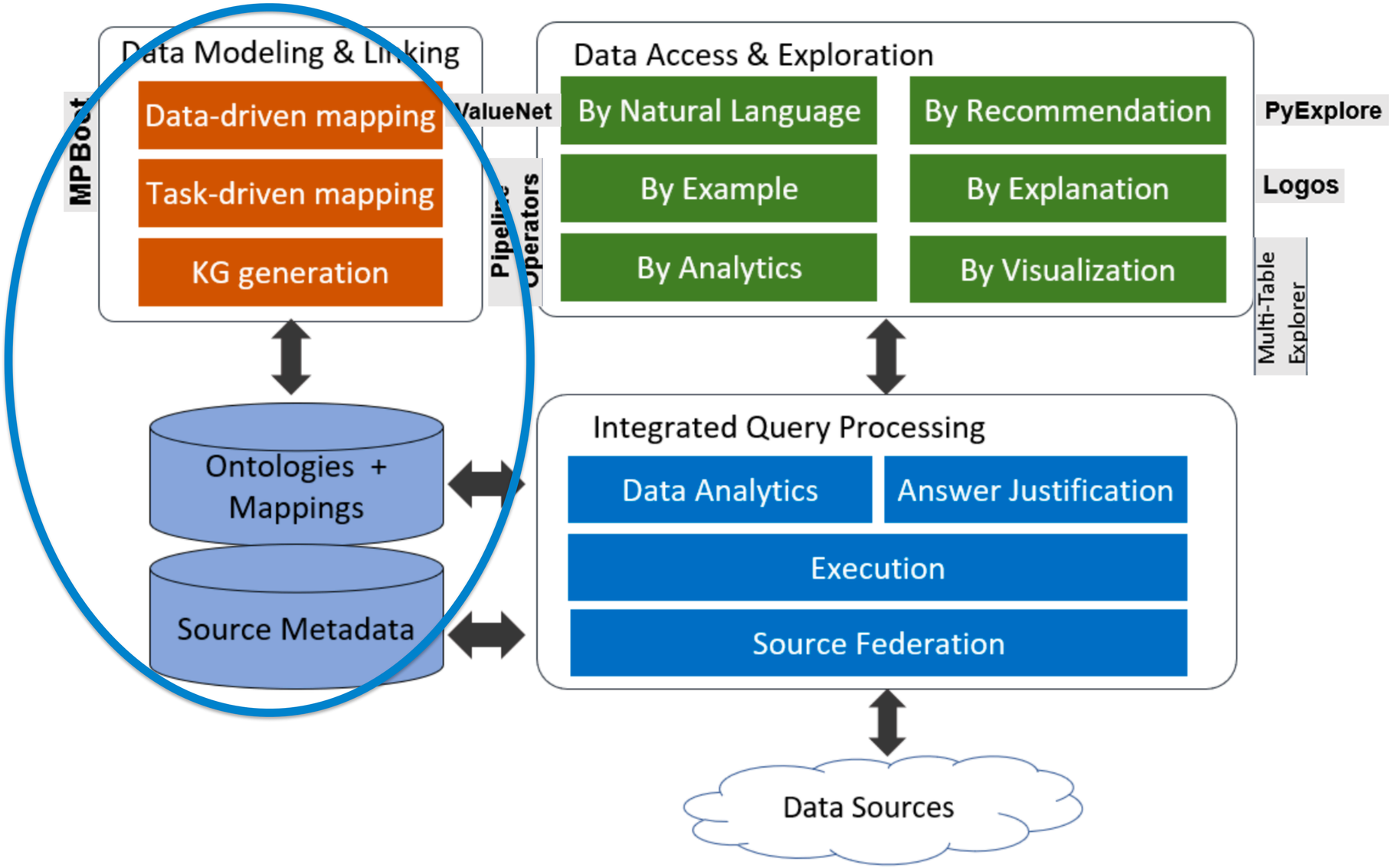


Figure 2: A join on the CORDIS database graph.



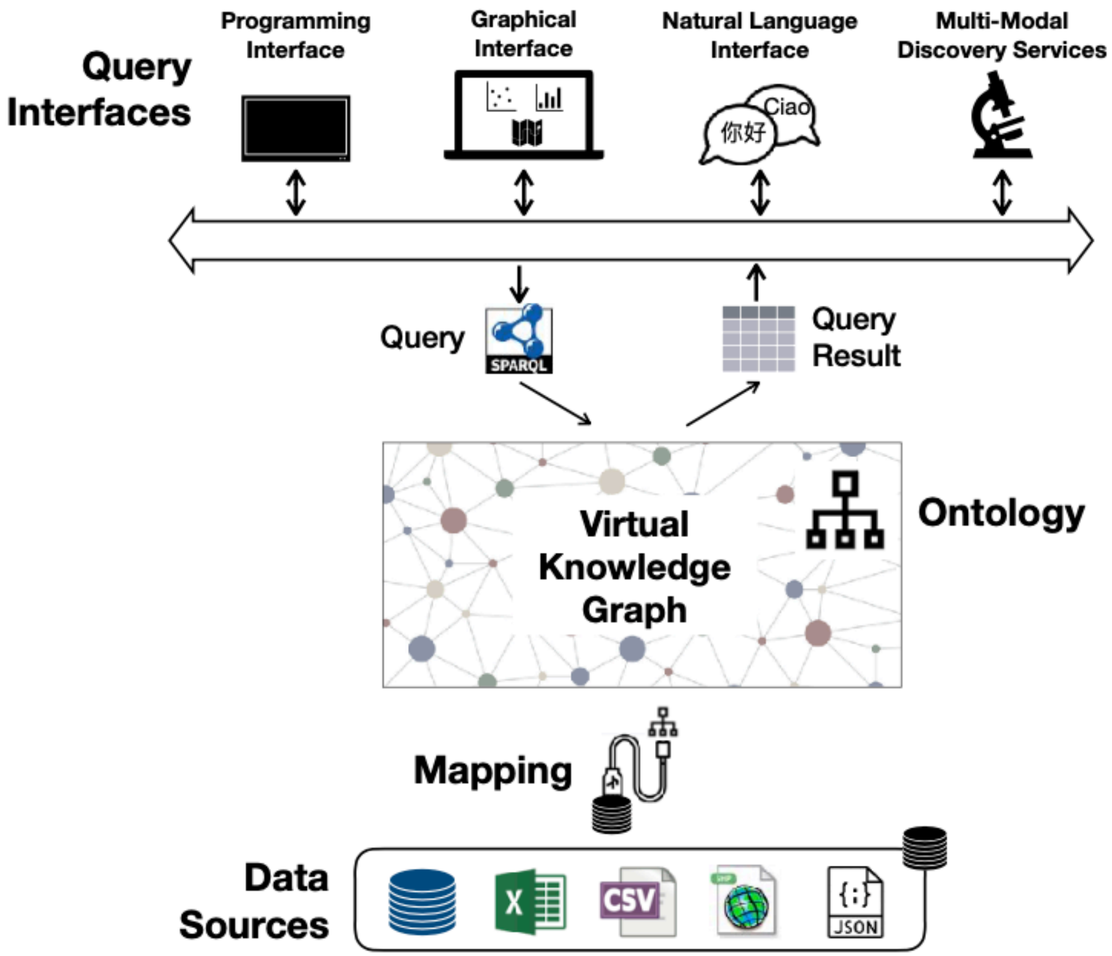
INODE Architecture

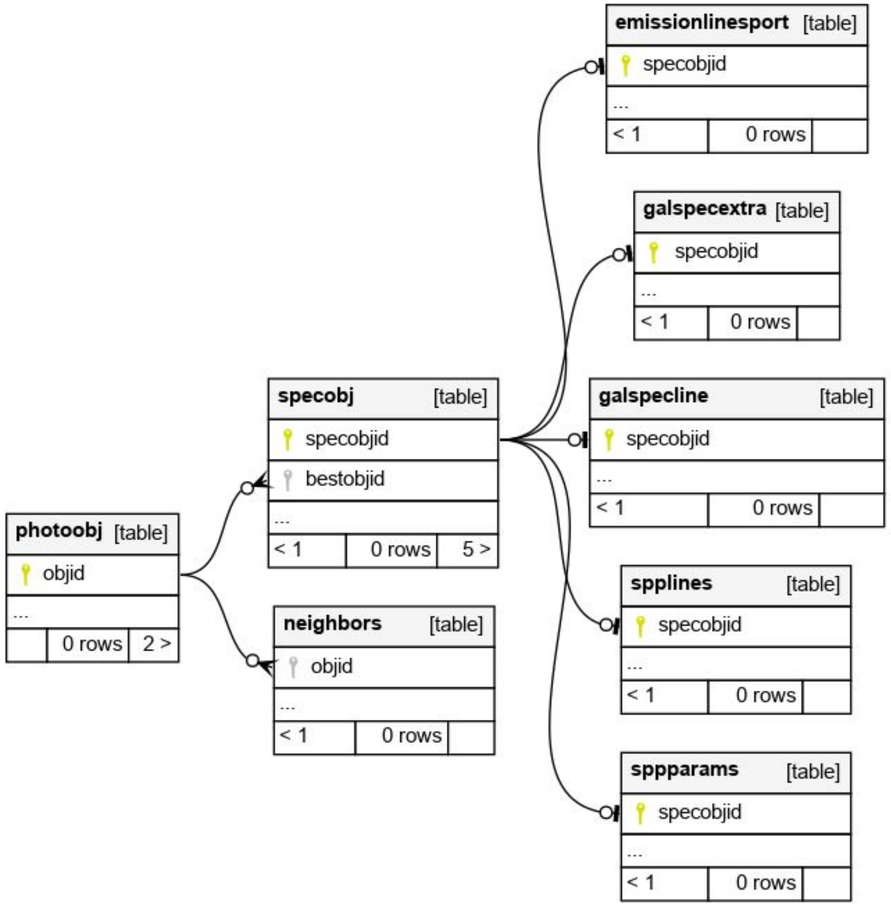
Open Data Dialog

Open Data Linking

Backend Services

Knowledge Graphs for Data Access (within INODE)



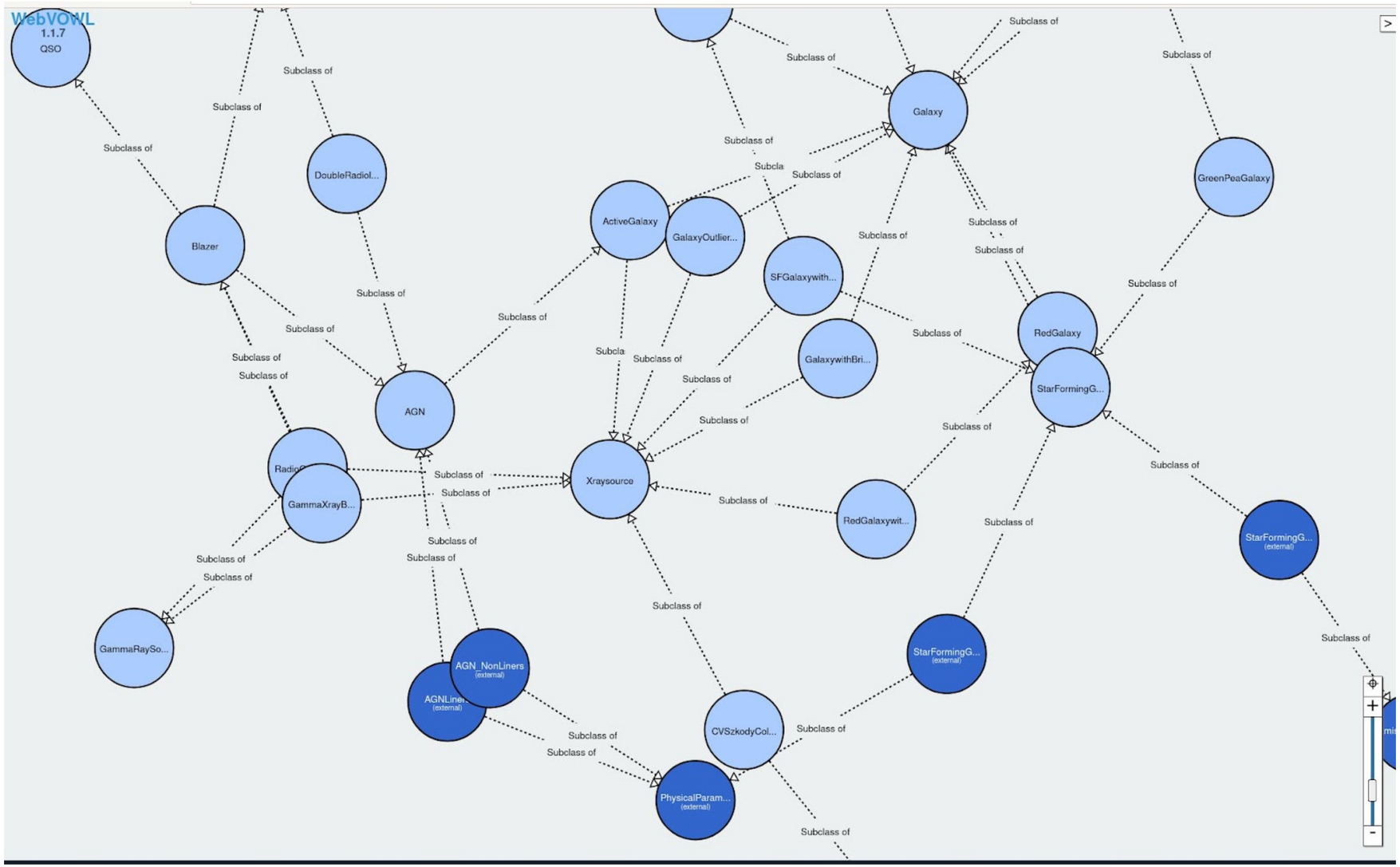


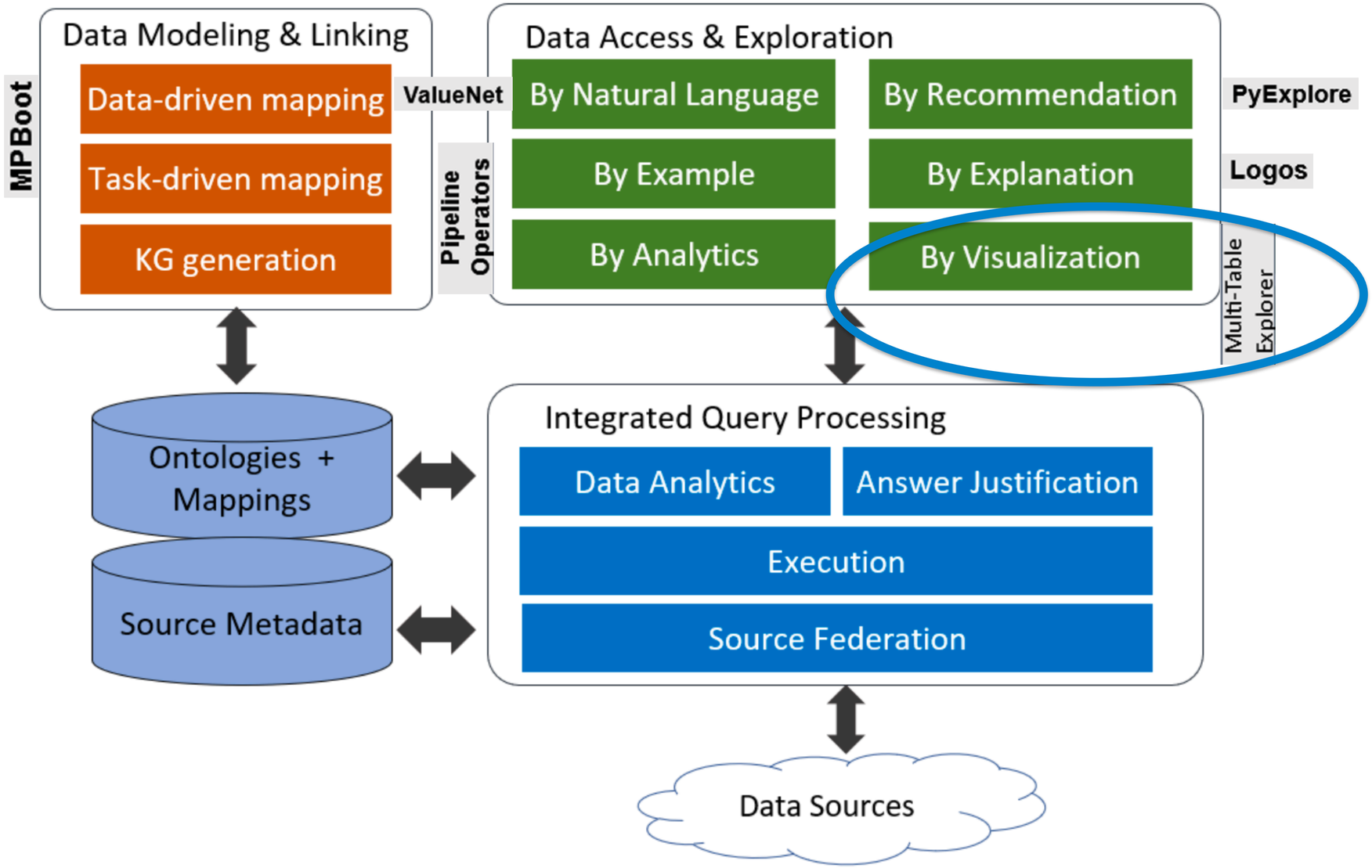
Subclasses

GALAXY
Starforming
Starburst
AGN

QSO

STAR
O, OB, B6, B9, A0, A0p, F2, F5, F9, G0, G2, G5, K1, K3, K5, K7, M0V, M2V, M1, M2, M3, M4, M5, M6, M7, M8, L0, L1, L2, L3, L4, L5, L5.5, L9, T2, Carbon, Carbon_lines, CarbonWD, CV





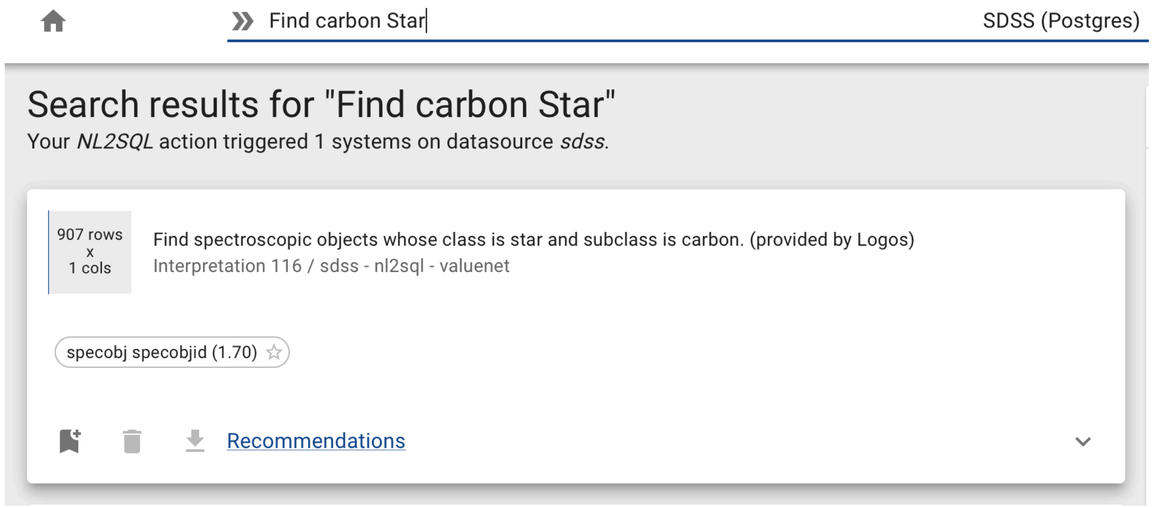
INODE Architecture

Open Data Dialog

Open Data Linking

Backend Services

NL query : Find carbon star



NL query : Find carbon star

Find carbon Star SDSS (Postgres)

Search results for "Find carbon Star"

Your *NL2SQL* action triggered 1 systems on datasource *sdss*.

907 rows x 1 cols

Find spectroscopic objects whose class is star and subclass is carbon. (provided by Logos)

Interpretation 116 / sdss - nl2sql - valuenet

specobj specobjid (1.70)

Recommendations

Interpretation 329

Produced by operator 323 of type `nl2sql/valuenet` on dataset sdss

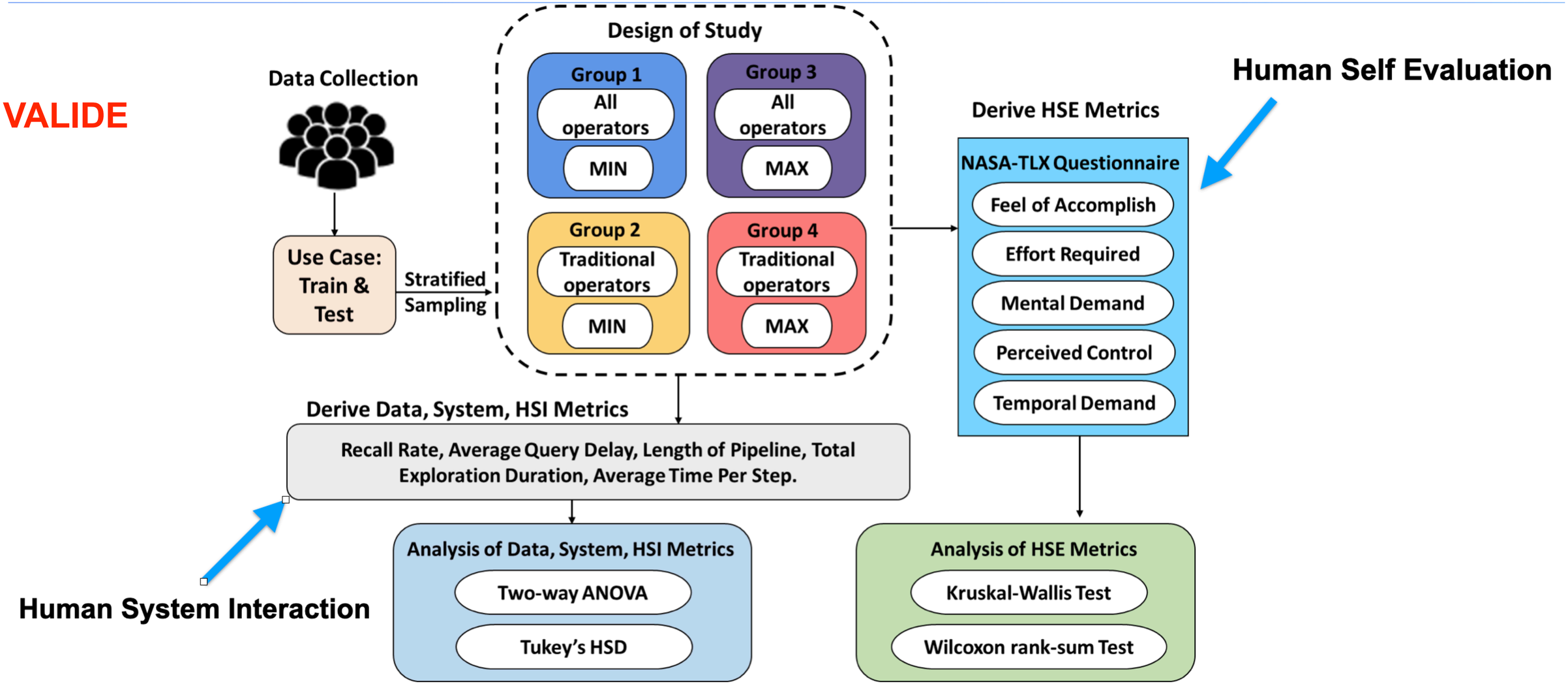

```
"root" : { 7 items
  "dataset" : string "sdss"
  "id" : int 329
  "operatorBaseType" : string "nl2sql"
  "operatorInvocation" : int 323
  "operatorSpecificType" : string "valuenet"
  "query" : string "SELECT T1.specobjid FROM specobj AS T1 WHERE T1.class = 'STAR' and T1.subclass = 'Carbon'"
  "table" : int 328
}
```


Table

```
SELECT T1.specobjid FROM specobj AS T1 WHERE T1.class = 'STAR' and T1.subclass = 'Carbon'
```

	Specobj specobjid
	907 rows 1 columns
	What are the objects of the class "STAR" and the class "Carbon"? Show the object ids. (provided by eQsplain) `SELECT T1.specobjid FROM specobj AS T1 WHERE T1.class = 'STAR' and T1.subclass = 'Carbon'`
	7988386943791812608 1 8350846451593793536 1 8849752605720858624 1
Row 0	1545951882948667392
Row 1	8354225800581763072
Row 2	8354215630099206144
Row 3	9256268462787153920
Row 4	7692185654755872768
Row 5	9250645010130685952
Row 6	9442027092600705024
Row 7	8350846451593793536

VALIDE

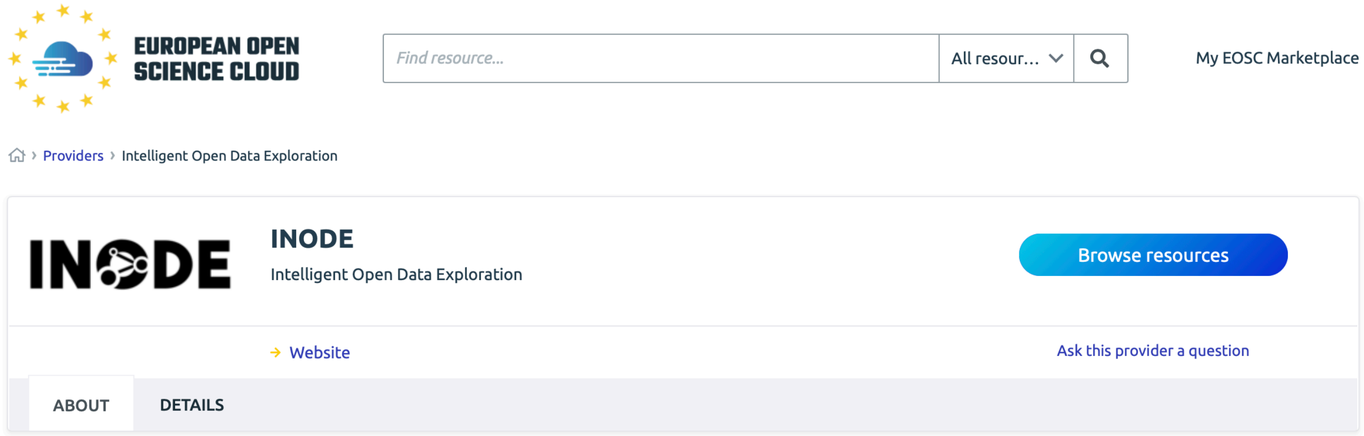


Evaluation framework -- system, data and human dimensions

Conclusions

- Building intelligent systems is not only fun but also enables access to data for a wide range of (non)-technical users
- We understand data faster and can also use it faster to generate scientific results or business value

- Further information:
  - http://www.inode-project.eu/
  - https://www.linkedin.com/in/project-inode/
  - Vision paper : “INODE: Building an End-to-End Data Exploration System in Practice”. Amer-Yahia, S., Koutrika, G., Bastian, F., Belmpas, T., Braschler, M., Brunner, U., ... & Stockinger, K. (2021). ACM SIGMOD Record 2021, https://arxiv.org/abs/2104.04194



INODE

INODE Use Cases News & Blog Publications Documents Events **Contact Us**

WORKSHOP ▪ Date: May 10, 2022

Upcoming! International INODE EOSC Workshop on the June 1, 2022 @ 10:00 (CET)

May 10, 2022@
Virtual zoom event

Please register via Eventbrite for participation.

10:00 - 10:05	Welcome to INODE (Kurt Stockinger, ZHAW, INODE Project Manager)
10:05 - 10:20	**INODE Use Cases**:
	• *Astrophysics* (Srividya Subramanian, Max Fabricius, MPI)
	• *Cancer Research* (Frederic Bastian, Tarcisio Mendes de Farias, SIB)
	• *Policy Making* (Guillem Rull, SIRIS)
10:20 - 10:40	Demos: **Data Exploration and Explanation in Natural Language** (NL):
	• *NL-to-SQL* (Kate Kosten, Yi Zhang, ZHAW)
	• *SQL-to-NL* (Stavroula Eleftheraki, George Katsogiannis, Athena)
10:40 - 11:20	Demos: **Interactive Data Exploration:**
	• *Query Builder* (Antonis Mandamadiotis, Athena)
	• *Multi Table Viewer* (Hendrik Lücke-Tieke, Fraunhofer)
	• *Query Recommendation* (Katerina Xagorari, Athena)
	• *Pipeline Operators* (Sihem Amer-Yahia, Aurélien Personnaz, Yogendra Patil, CNRS)
11:20 - 11:30	Q&A Session 1
11:30 - 12:00	Demos: **Data Integration and Knowledge Graphs:**
	• *Knowledge Graphs for Data Access* (Davide Lanti, Diego Calvanese, UNIBZ)
	• *Information Extraction, Database Enrichment and NL-to-Cypher* (Ellery Smith, ZHAW)
	• *Knowledge Graph Enrichment and Decision Support* (Dimitris Giagkos, Infili)
12:00 - 12:15	Q&A Session 2

https://www.inode-project.eu/events/international-inode-eosc-workshop-on-the-june-1-2022-10-00-cet

Thank you !