Finanziato dall'Unione europea
NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA
ICSC
Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

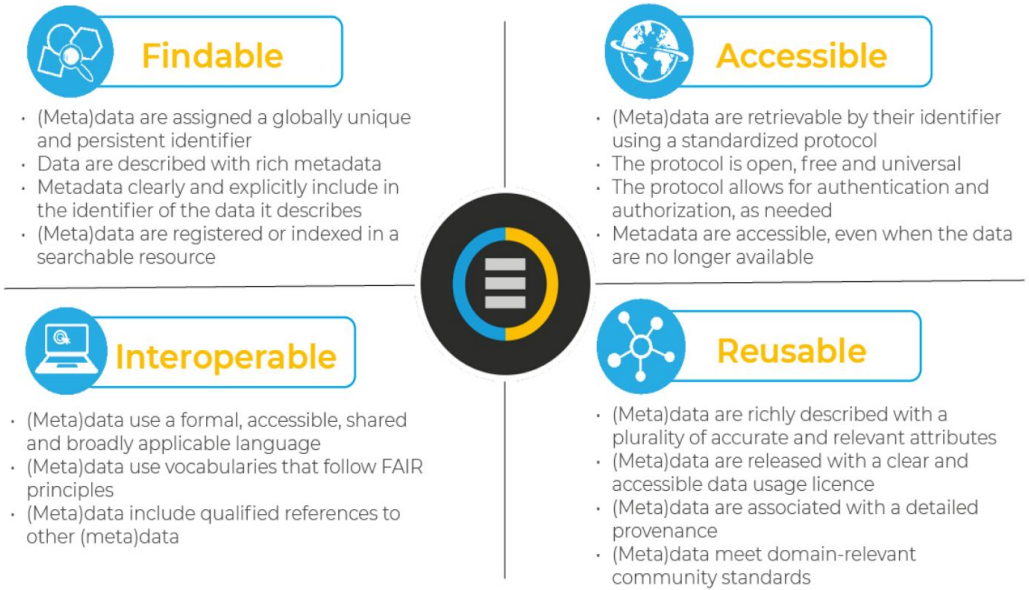
# Interoperable Data Lake
# WP2

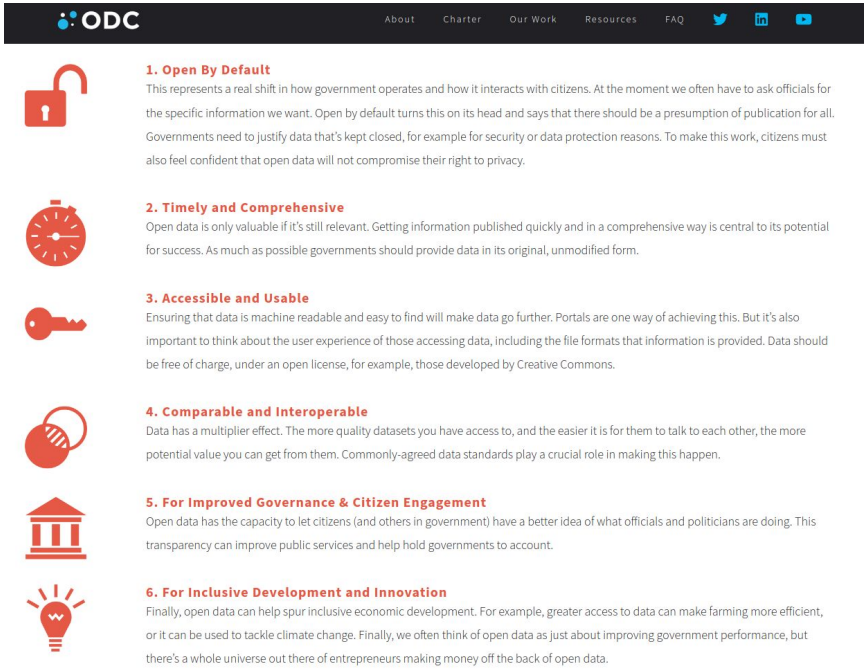
Cristina Knapic - INAF

**IDL Kick-off meeting, Bologna** October 6 2023

# Interoperable (FAIR&Open) Data Lake

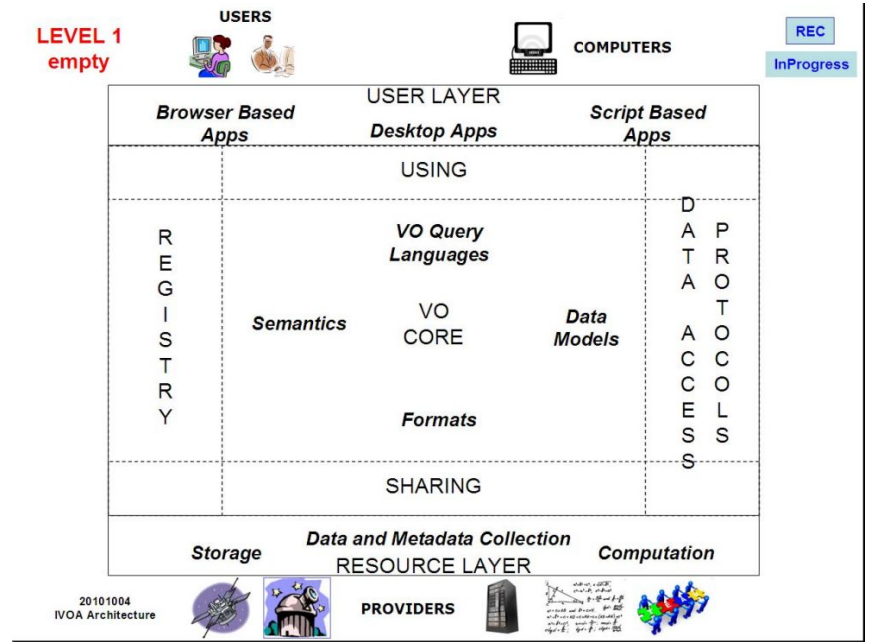
The IDL project will create a federated data infrastructure. In particular, WP2 will contribute in the analysis of a suitable data characterization and proper metadata description in order to follow the FAIR and Open Data principles.

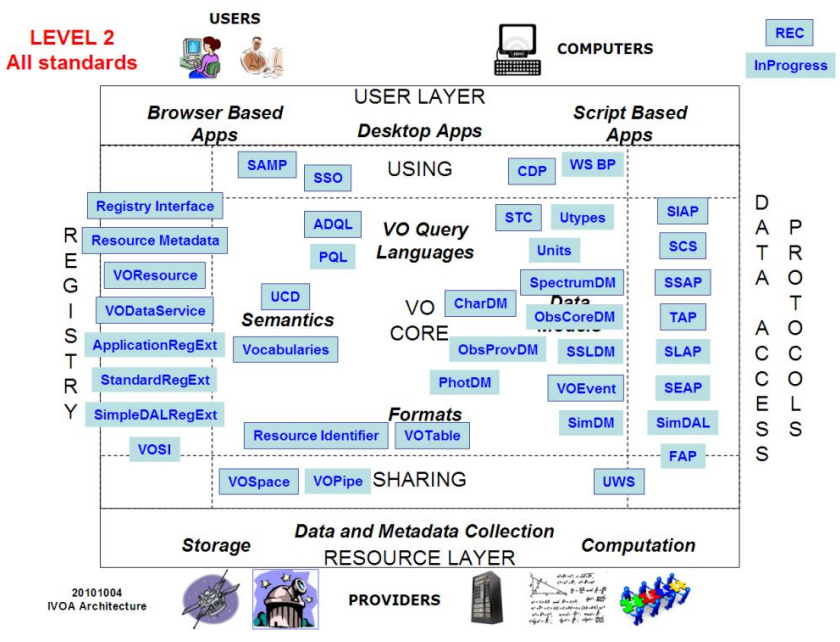


# Reference standards and starting points - 1

A key activity to obtain this objective is the definition and use of a metadata layer and standard, to be developed in close collaboration with data provider communities. IVOA Standards are already defined in the optic of interoperability between collections of different wavelengths. We will use and eventually extend the IVOA standards in order to include also the space economy products.



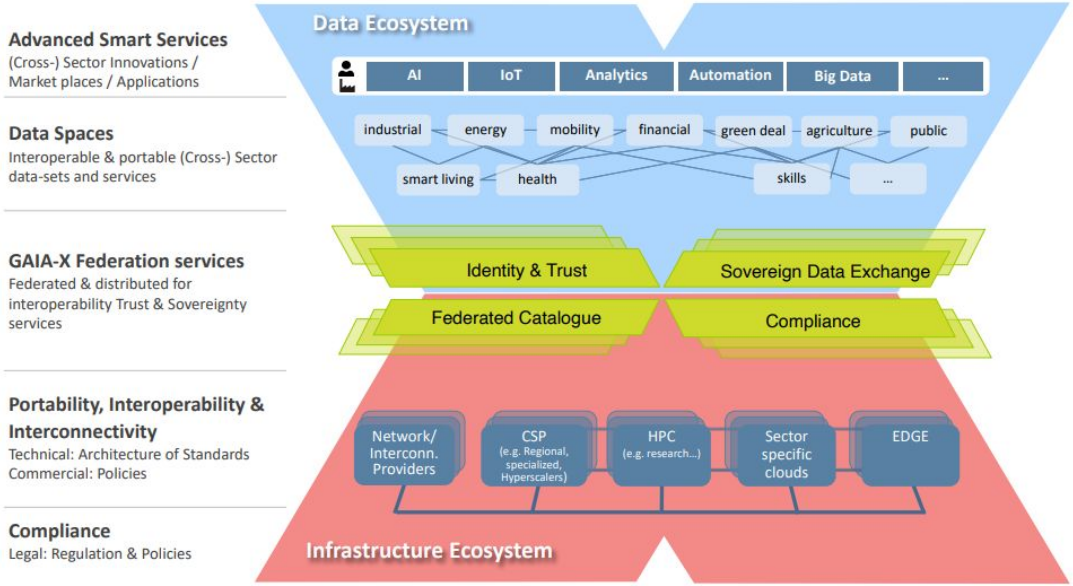



# Reference Standards and starting points - 2

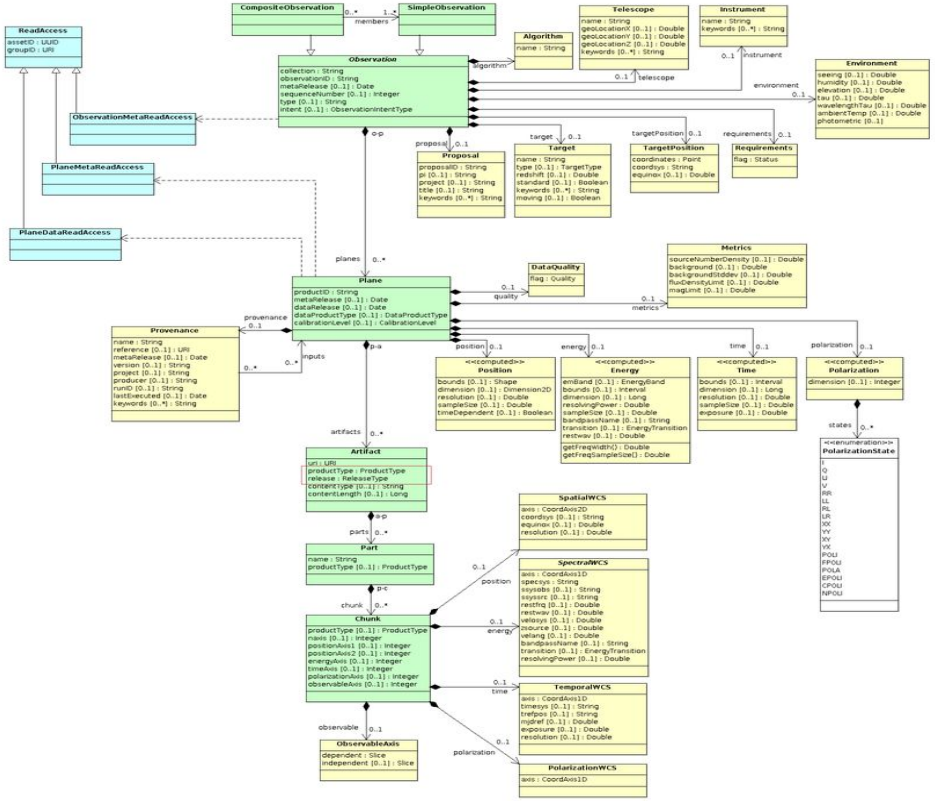
## Industrial standards for data interoperability, identity and trust

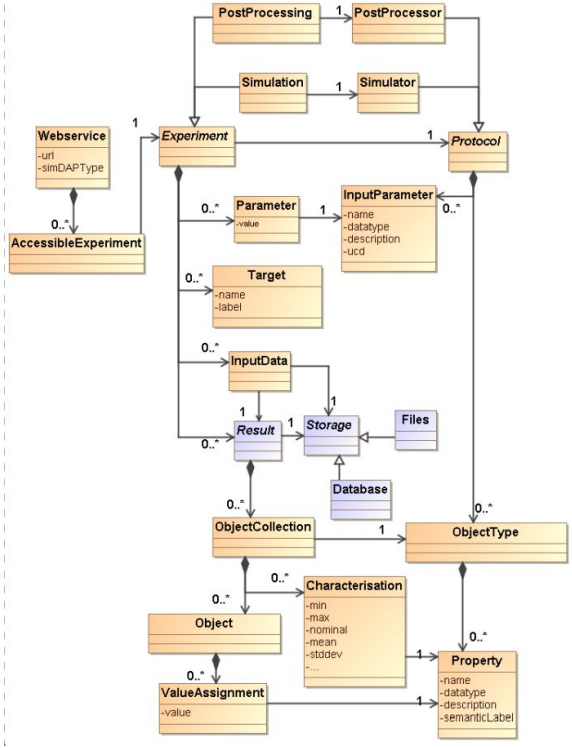
### The Gaia-X ecosystem of services and data

gaia-x



# Astrophysics data models

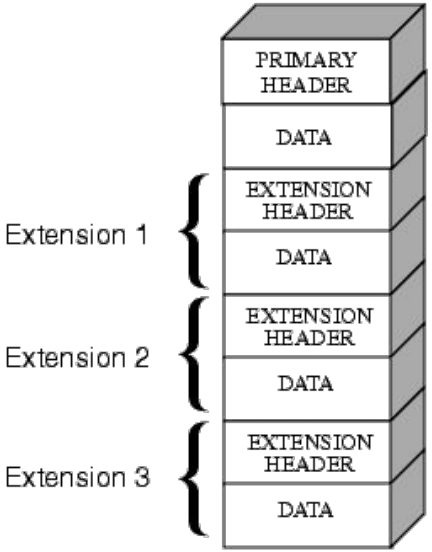


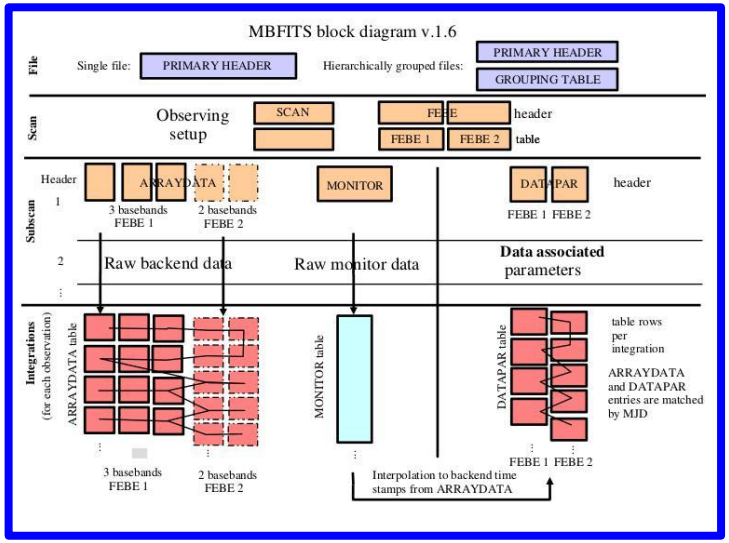
# Astrophysics data formats

The data model will include all relevant information about data, software for data reduction, analysis, data policy, and all the information for data filtering for retrieval.

FITS




MBFits




Other formats:
* satellite proprietary DMs and formats
* custom DMs and formats

# WP2 activities

**Task 2.1: Datasets identification**: selection of data coming from radioastronomy, space debris observations, numerical simulations;

**Task 2.2: Data Models and metadata definition**: Identification of a suitable data model and metadata definition, Implementation of an ingestion software;

**Task 2.3: Requirement analysis (for DB)**: choice, implementation, testing and verification of the database technology to be integrated on top of the data-lake. Investigation of no-sql dbs in connection with rucio;

**Task 2.4: Database deployment in ICSC infrastructure**: deployment of the ingestion software and the chosen DB into the ICSC infrastructure

**Task 2.5: Validation and testing**: performance test of the processing pipelines on a hybrid HPC/cloud

infrastructure.

# WP2 Outputs and Milestones

Definition of a performance oriented, scalable, VO and possibly GAIA-X compliant application for astronomical and satellite data ingestion in a HPC environment.

Application will be connected to the data lake technology carried on by the other WPs in a suitable manner to enhance the performance of all the system. Enhancement of db and storage performances for data discovery and access.

| Timing / Milestone | Activity |
|---|---|
| M12 / MS9 | Requirement analysis |
| M12 / MS9 | Data provisioning |
| M18 / MS10 | Database deployments |
| M24 / MS10 | Validation and testing |

# Documentation

| Document | Release |
|---|---|
| TN1 Dataset identified and data model vs1 | M12 |
| TN2 Technical report of the database | M18 |
| TN3 Final report on database and data model vs1.1 | M24 |

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
ICSC Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

Thanks for attention