



Finanziato  
dall'Unione europea  
NextGenerationEU



Ministero  
dell'Università  
e della Ricerca



Italiadomani  
PIANO NAZIONALE  
DI RIPRESA E RESILIENZA



# 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.



## Findable

- (Meta)data are assigned a globally unique and persistent identifier
- Data are described with rich metadata
- Metadata clearly and explicitly include in the identifier of the data it describes
- (Meta)data are registered or indexed in a searchable resource



## Accessible

- (Meta)data are retrievable by their identifier using a standardized protocol
- The protocol is open, free and universal
- The protocol allows for authentication and authorization, as needed
- Metadata are accessible, even when the data are no longer available



## Interoperable

- (Meta)data use a formal, accessible, shared and broadly applicable language
- (Meta)data use vocabularies that follow FAIR principles
- (Meta)data include qualified references to other (meta)data



## Reusable

- (Meta)data are richly described with a plurality of accurate and relevant attributes
- (Meta)data are released with a clear and accessible data usage licence
- (Meta)data are associated with a detailed provenance
- (Meta)data meet domain-relevant community standards



## ODC

About Charter Our Work Resources FAQ



### 1. Open By Default

This represents a real shift in how government operates and how it interacts with citizens. At the moment we often have to ask officials for the specific information we want. Open by default turns this on its head and says that there should be a presumption of publication for all. Governments need to justify data that's kept closed, for example for security or data protection reasons. To make this work, citizens must also feel confident that open data will not compromise their right to privacy.



### 2. Timely and Comprehensive

Open data is only valuable if it's still relevant. Getting information published quickly and in a comprehensive way is central to its potential for success. As much as possible governments should provide data in its original, unmodified form.



### 3. Accessible and Usable

Ensuring that data is machine readable and easy to find will make data go further. Portals are one way of achieving this. But it's also important to think about the user experience of those accessing data, including the file formats that information is provided. Data should be free of charge, under an open license, for example, those developed by Creative Commons.



### 4. Comparable and Interoperable

Data has a multiplier effect. The more quality datasets you have access to, and the easier it is for them to talk to each other, the more potential value you can get from them. Commonly-agreed data standards play a crucial role in making this happen.



### 5. For Improved Governance & Citizen Engagement

Open data has the capacity to let citizens (and others in government) have a better idea of what officials and politicians are doing. This transparency can improve public services and help hold governments to account.

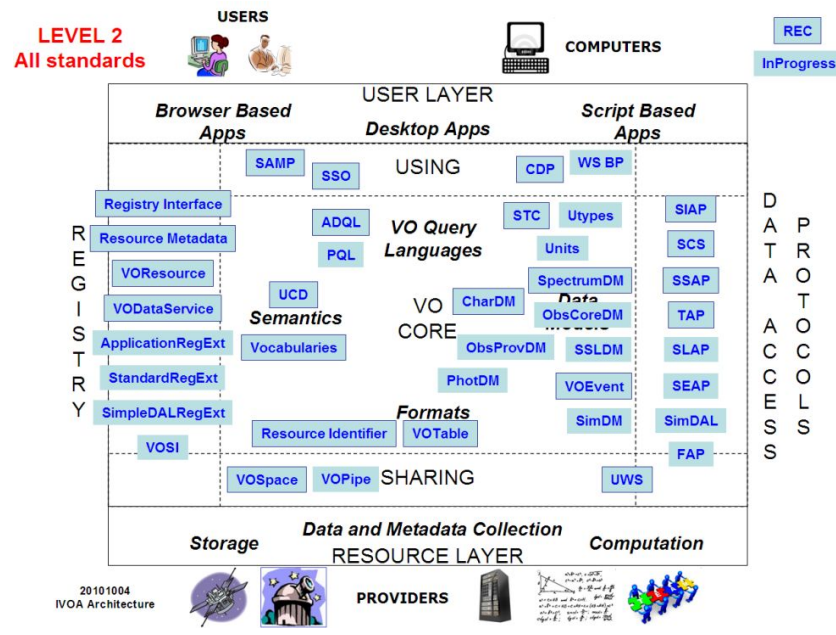
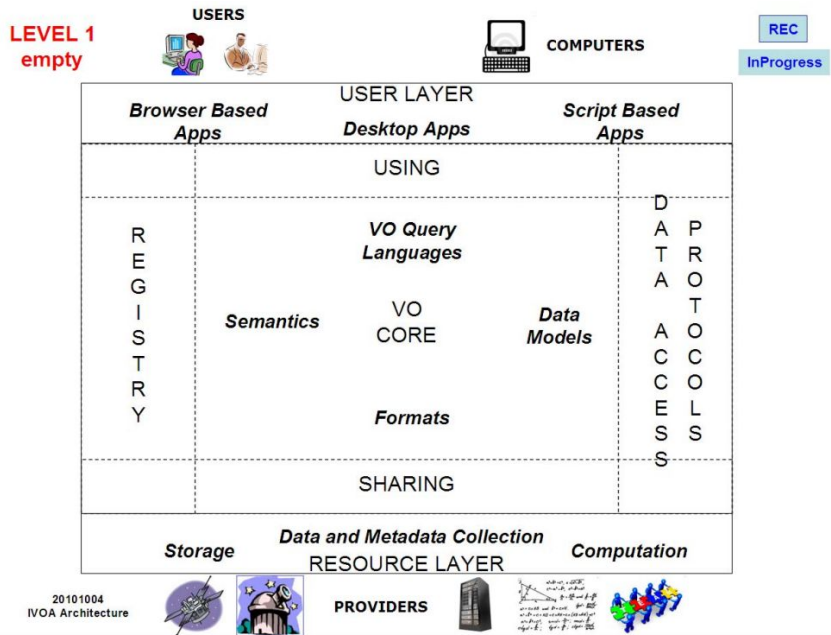


### 6. For Inclusive Development and Innovation

Finally, open data can help spur inclusive economic development. For example, greater access to data can make farming more efficient, or it can be used to tackle climate change. Finally, we often think of open data as just about improving government performance, but there's a whole universe out there of entrepreneurs making money off the back of open data.

# 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.

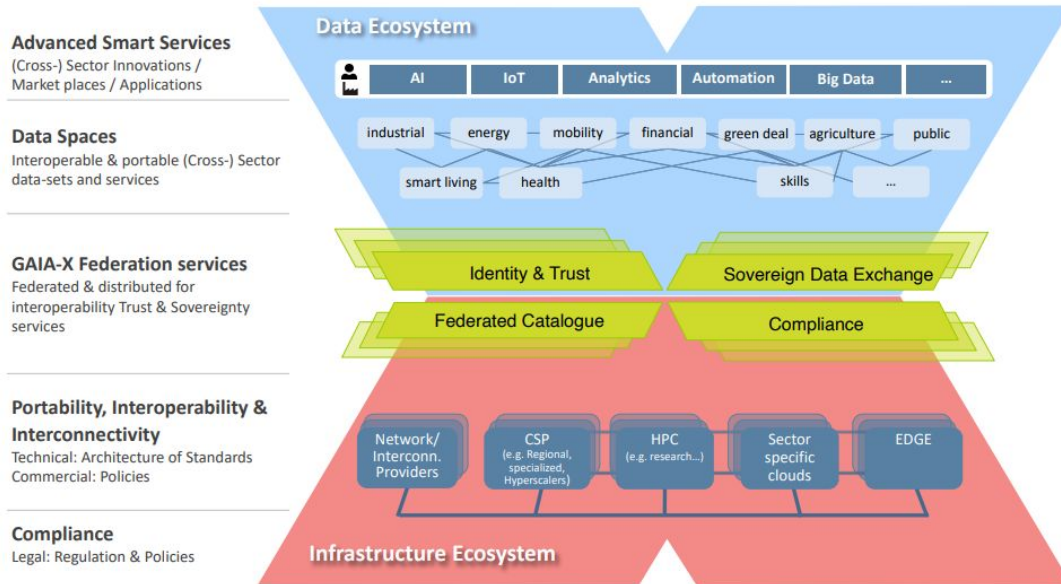




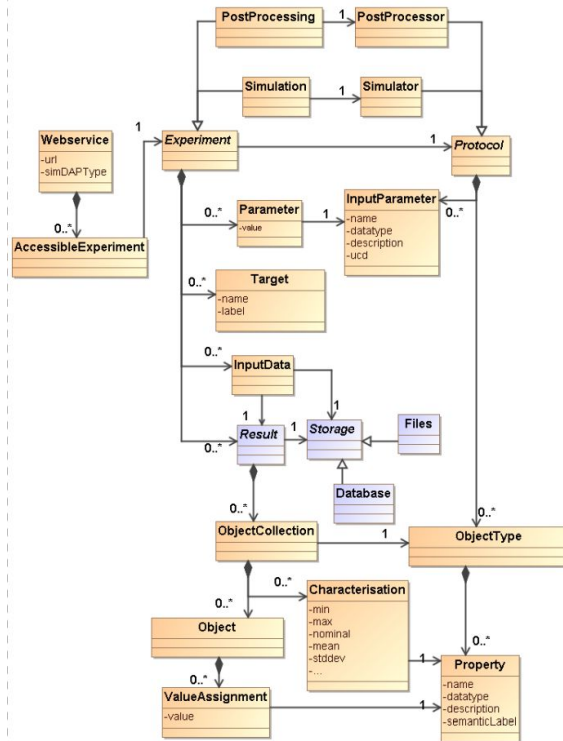
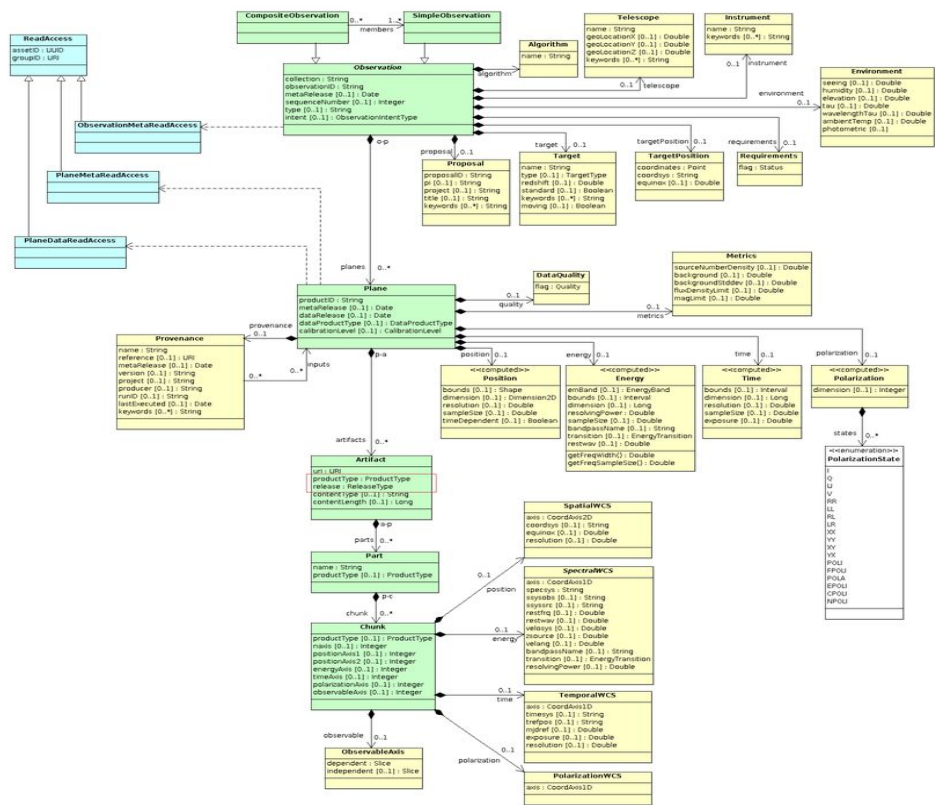
# Reference Standards and starting points - 2

## Industrial standards for data interoperability, identity and trust

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



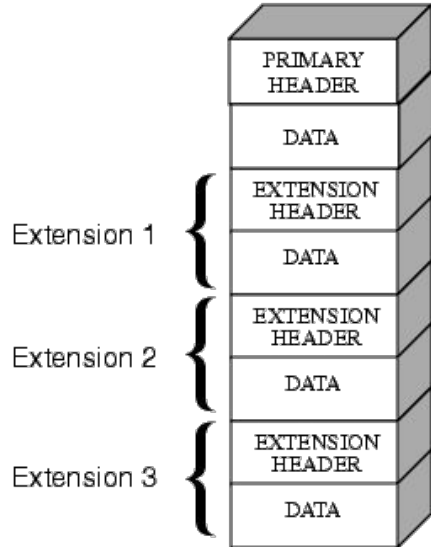
# Astrophysics data models



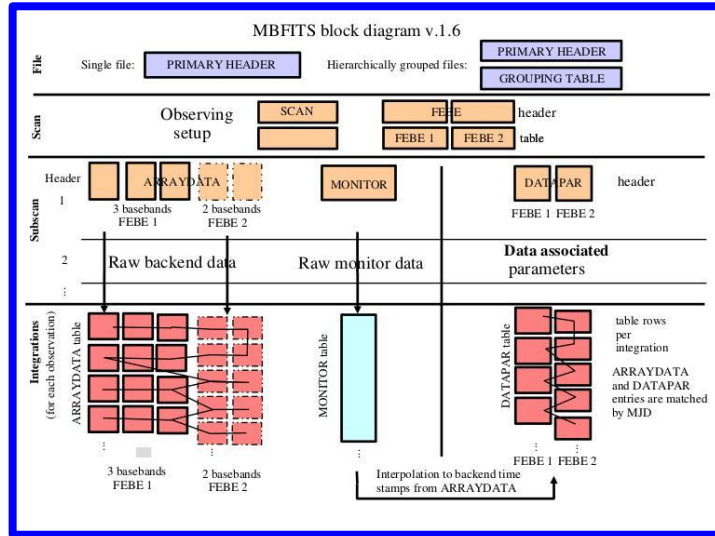
# Astrophysics data formats

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



Thanks for attention