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

WP4 - Big Data Management -Storage and Archive

Cristina Knapic - INAF

Spoke 3 Technical Workshop, Trieste October 9 / 11, 2023

WP4 scientific rationale

- analysis of the key science product use cases and data model implementation;
- creation of a queryable data archive for the most relevant use cases, with performance focus on database and data retrieval in a distributed environment;
- data management plan implementation following the Spoke 3 WP0 directives;
- creation of a intuitive web application for data discovery and management, powered with appropriate tools for moving data to computing centres;

Technical Objectives, Methodologies and Solutions

- analysis of the data model of each considered use case, probably starting by the analysis of the data format and its content, trying to adhere to known standards;
- definition of the software architecture for the FAIR and Open data archive:
  - ingestion software depending on different data models, formats and storing paradigms;
  - implementation of a reliable and scalable process for data distribution along multiple sites;
  - implementation of a scalable and performant data base for metadata descriptors;
  - creation of a bridge between the data distribution process and the ancillary metadata database;
  - implementation of a web interface for data access suited with Authentication and Authorization mechanisms;
- for each collection, definition of a Data Management Plan and in particular customization of data flow, accessing rules and privacy policy;

Interoperable (FAIR&Open) Data

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

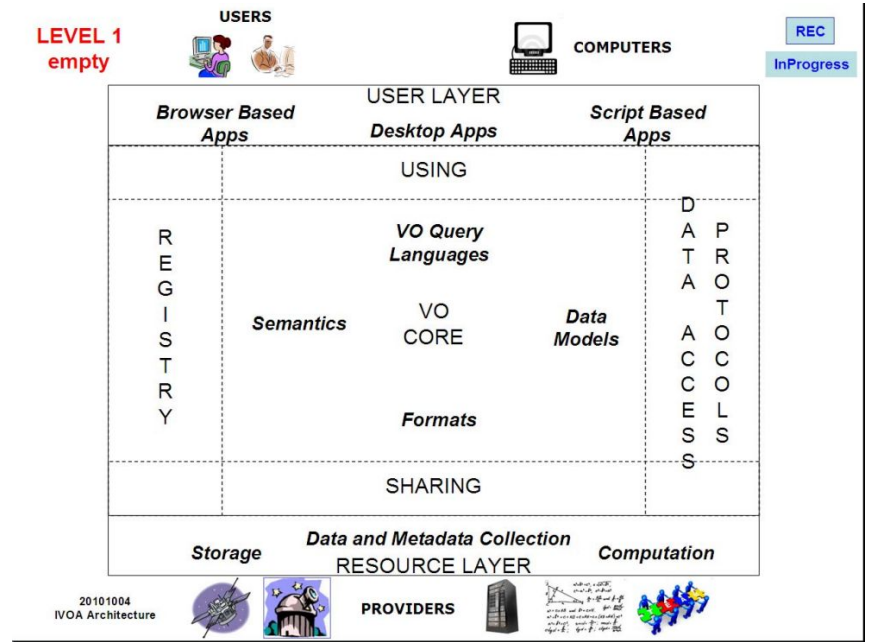
Open data has the capacity to let citizens (and others in government) have a better idea of what officials and politicians are doing. This transparency can improve public services and help hold governments to account.

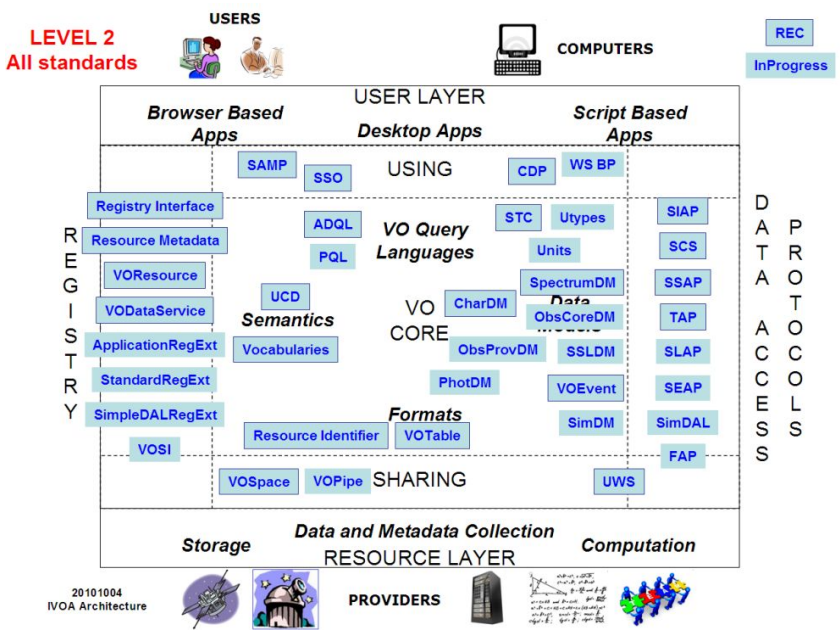
6. For Inclusive Development and Innovation

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.

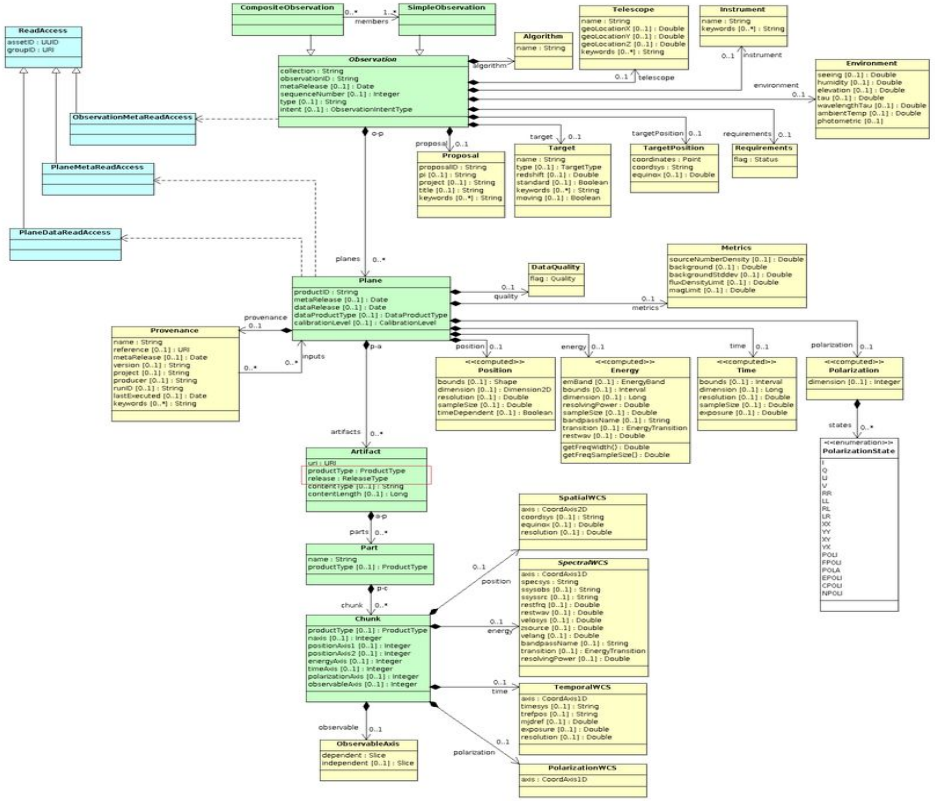
Reference standards and starting points - 1

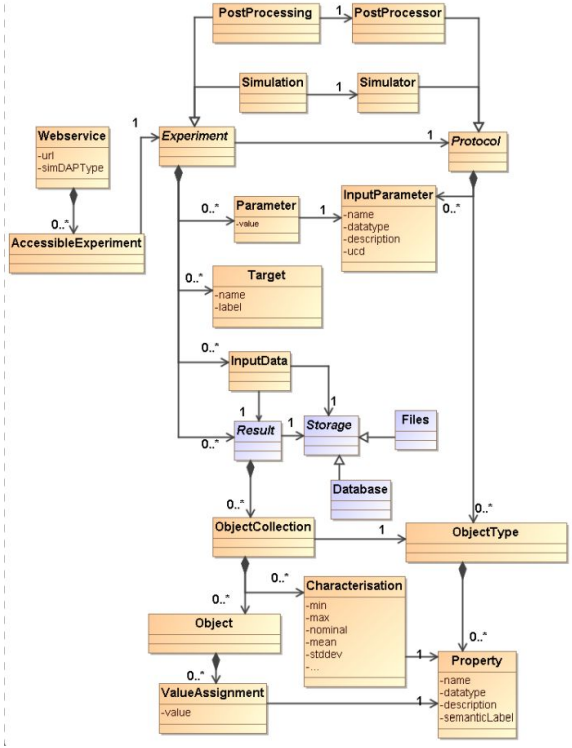
A key activity to obtain this objective is the definition and use of a metadata layer and standard, to be developed in close collaboration with data provider communities. IVOA Standards are already defined in the optic of interoperability between collections of different wavelengths. We will use and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





Astrophysics data models

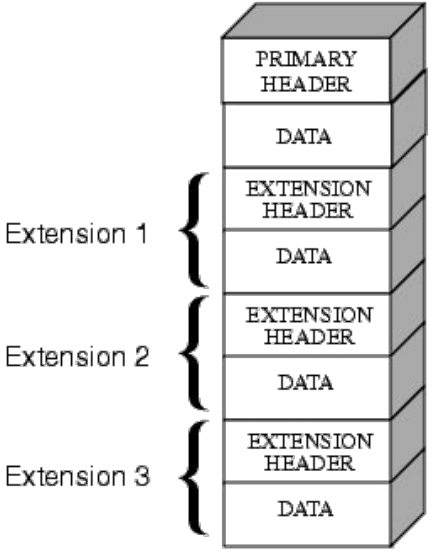




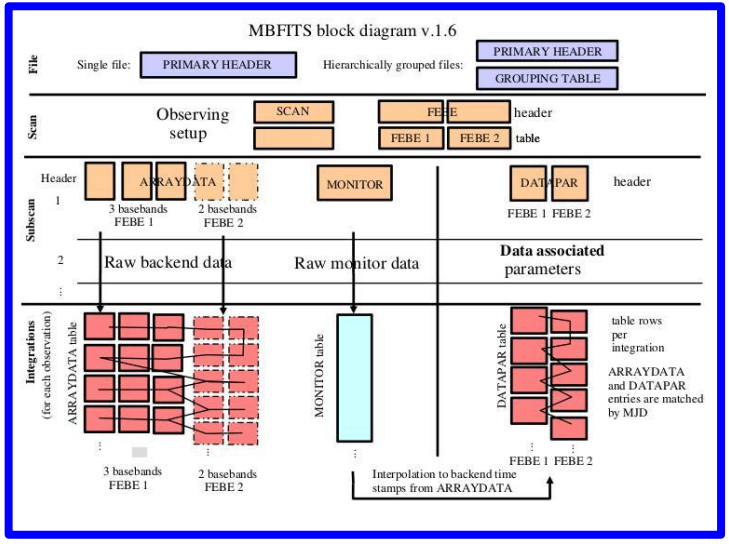
Astrophysics data formats

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



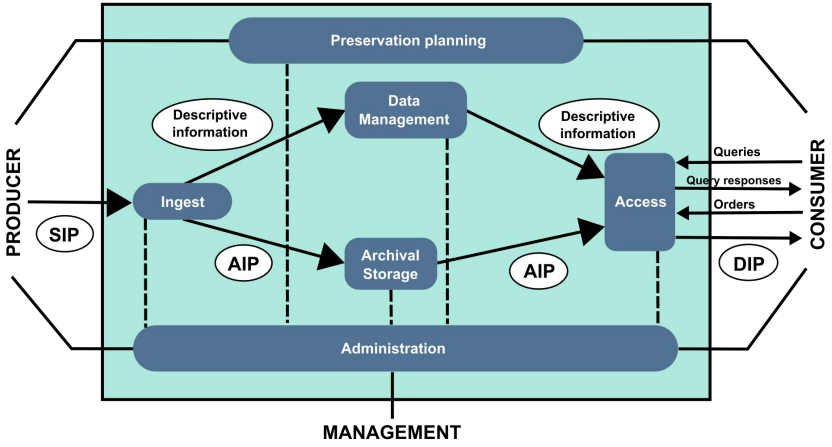
MBFits



Other formats:
* simulated and satellite DMs and formats
* custom or private DMs and formats

Data Management and storage

- definition of DMP for each collection;
- separation of function between data and metadata;
- configurable data ingestion software independent on formats, rel/no-rel database, storage type (local, remote, cloud) (G. Coran presentation);
- data distribution using rucio (D. Ciangottini / A. Adelfio presentations) ;
- smart web application for data retrieval;



Milestones

WP4 Big Data management, storage and archiving	**786 g**	**gio 01-09-22**	**gio 04-09-25**
T4.1 Big data managem. and Standardization	526 g	ven 02-09-22	ven 06-09-24
T4.2 Big Data Storage Optimiz.	636 g	gio 30-03-23	gio 04-09-25
T4.3 Archive design and implement	524 g	lun 04-09-23	gio 04-09-25

Criticalities

- short time to analyse the different Key Science projects since some of them are still not defined;
- resources are partially available;
- skills of persons involved (person to be involved on 1st of Dec.);
- lack of clear key indicators since no already defined use cases (except GAIA - gravitational waves).

Innovation Grant participation

- Interoperable Data Lake: involvement on the implementation of a data lake interoperable with IVOA and GAIA-X standards, optimization of the data base performances and identification of bottlenecks in peculiar data workflows;
- possible shared objective with HAMMON IG.

Interaction with industries and private companies will bring benefits in terms of database optimization techniques and will help the WP 4 in the development of ingestion code for fast data archiviation.

Thanks for attention

Reference Standards and starting points - 2

Industrial standards for data interoperability, identity and trust

The Gaia-X ecosystem of services and data

gaia-x