Lecture 11 – Metadata management

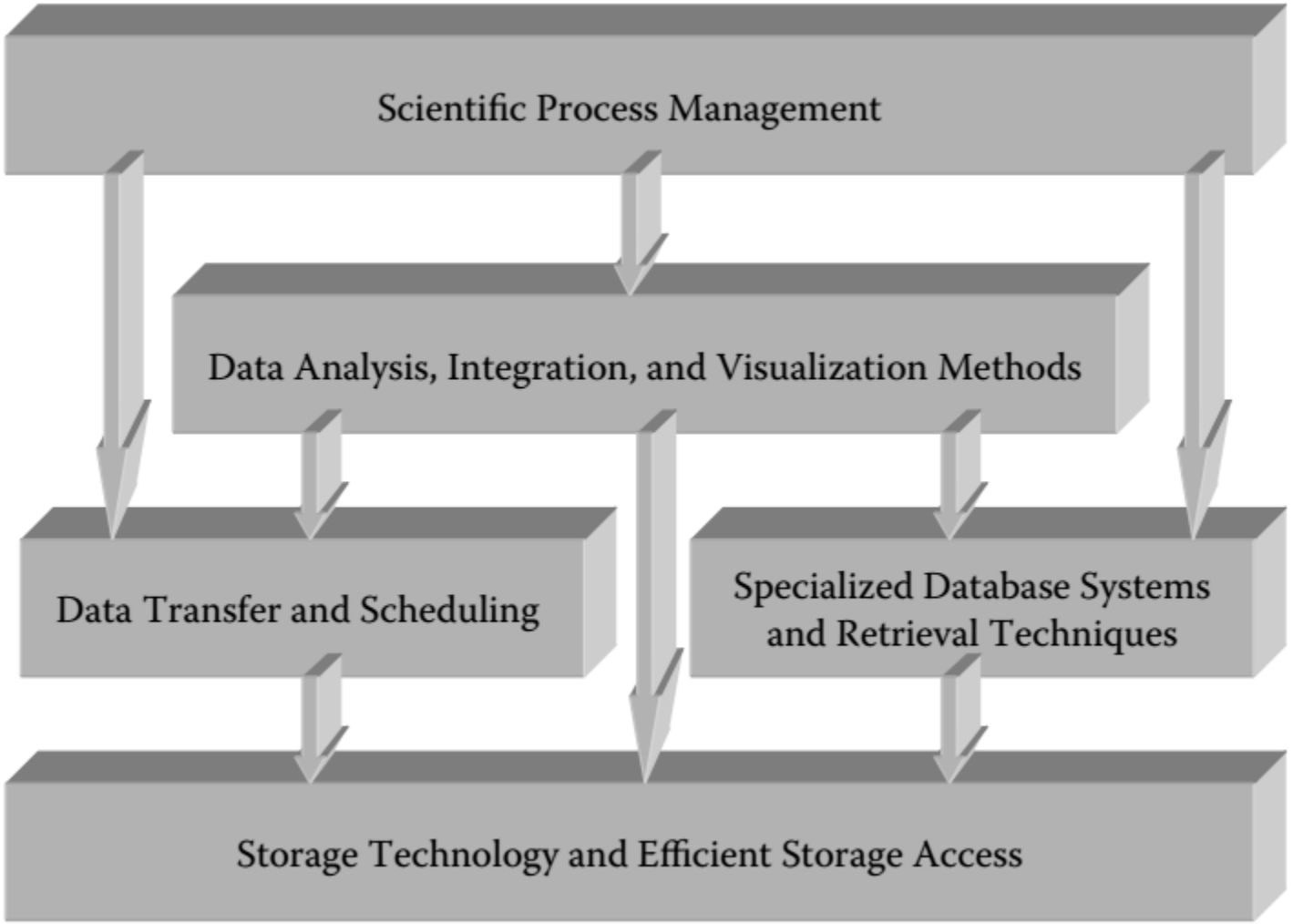
Open Data Management & the Cloud

(Data Science & Scientific Computing / UniTS – DMG)

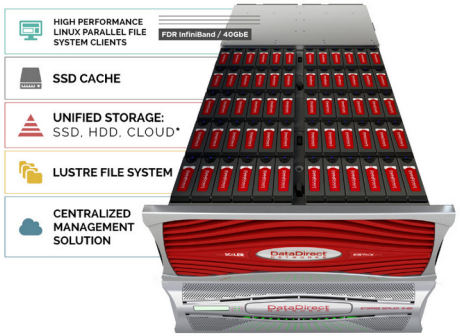
Scientific data management

- During the scientific exploration process, from the data generation phase to the data analysis phase, data management involves five main aspects
  - the efficient access to storage systems, in particular, parallel file systems, to write and read large volumes of data
  - A second aspect is the efficient data movement and management of storage spaces
  - techniques for automatically optimizing the physical organization of data, necessary for fast analysis
  - how to effectively perform complex data analysis and searches over large datasets
  - the automation of multistep scientific process workflows

Data management technologies



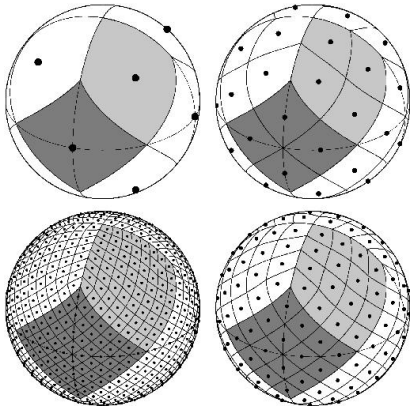
Data management technologies



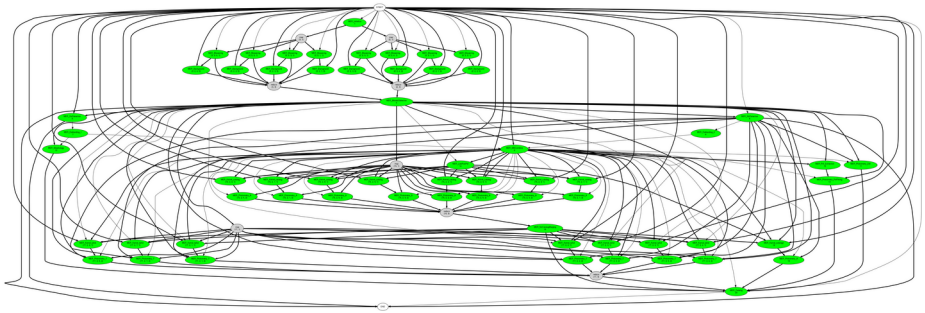
36 - 40

1.02PB-6.2PB HDD
926TB-4.76PB SSD

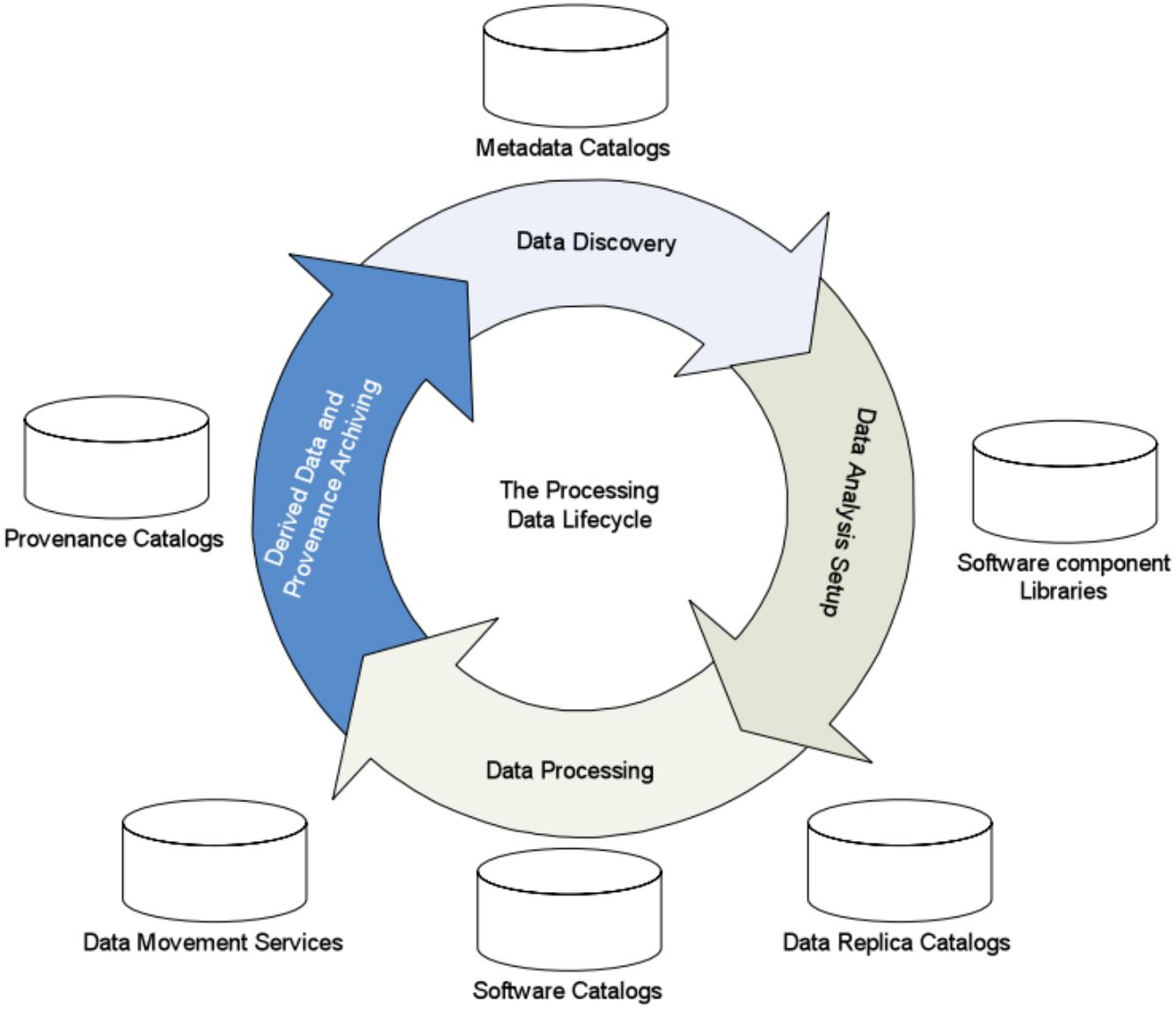
Up to 60GB/s / Rack



TensorFlow



The data lifecycle



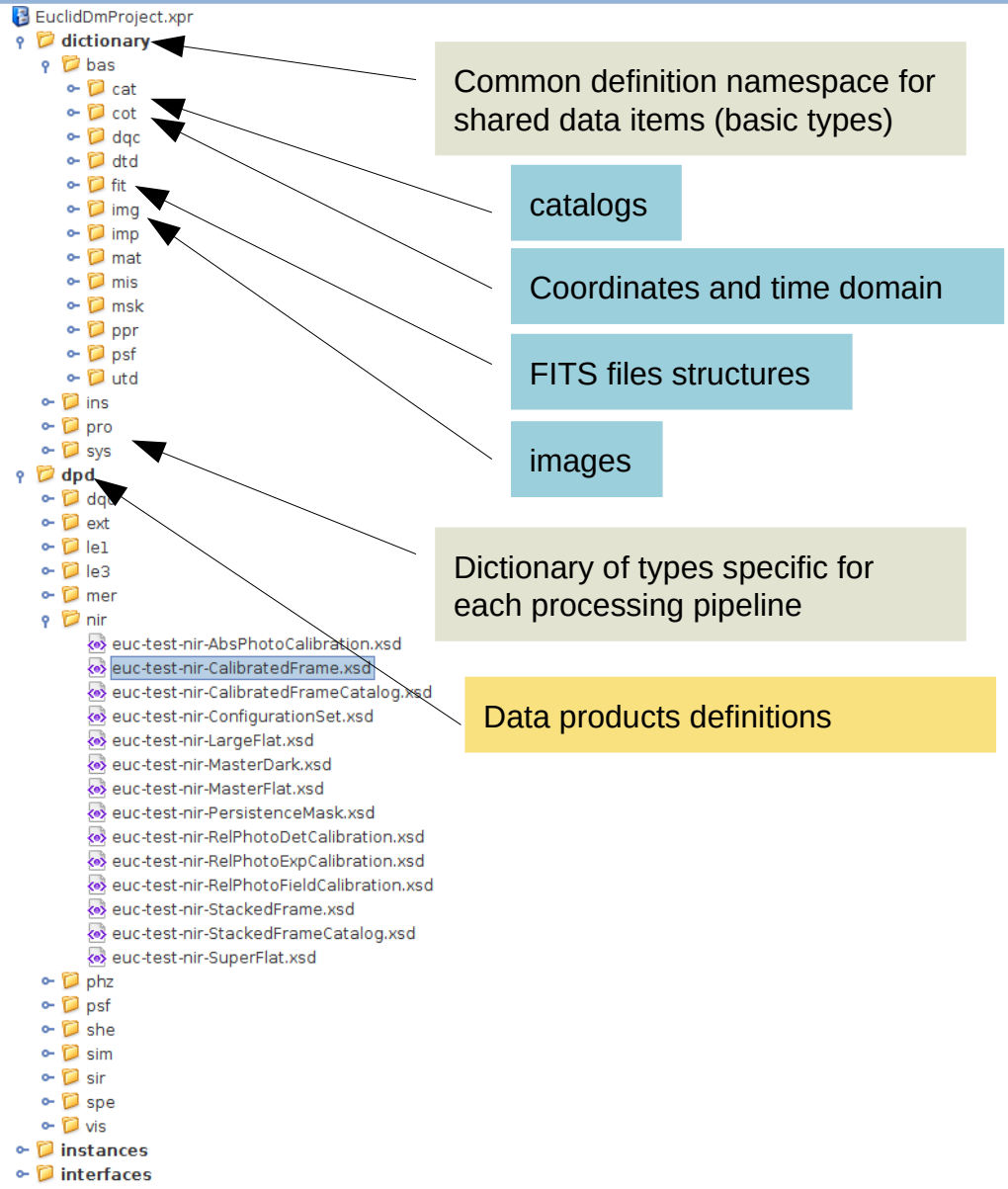
Metadata definition

- **Metadata**: “structured data about an object that supports functions associated with the designated object”
- *Structured data* implies a systematic ordering of data according to a metadata schema specification (see "**Metadata Vocabulary**")
- *Functions associated with the designated object*. The emphasis here is on the ability of metadata to support the activities and behaviors of an object. For example:
  - For example, ‘‘author,’’ ‘‘title,’’ and ‘‘subject’’ metadata facilitate the *discovery* of an information resource
  - An ‘‘invoice number’’, ‘‘product code’’, ‘‘credit card number (for payment)’’ and ‘‘date of financial transaction” metadata capture the *purchase* activity for a consumer good
- The above definition covers the dual function of the metadata:
  - describing the objects from a logical point of view
  - describing their physical and operational attributes.

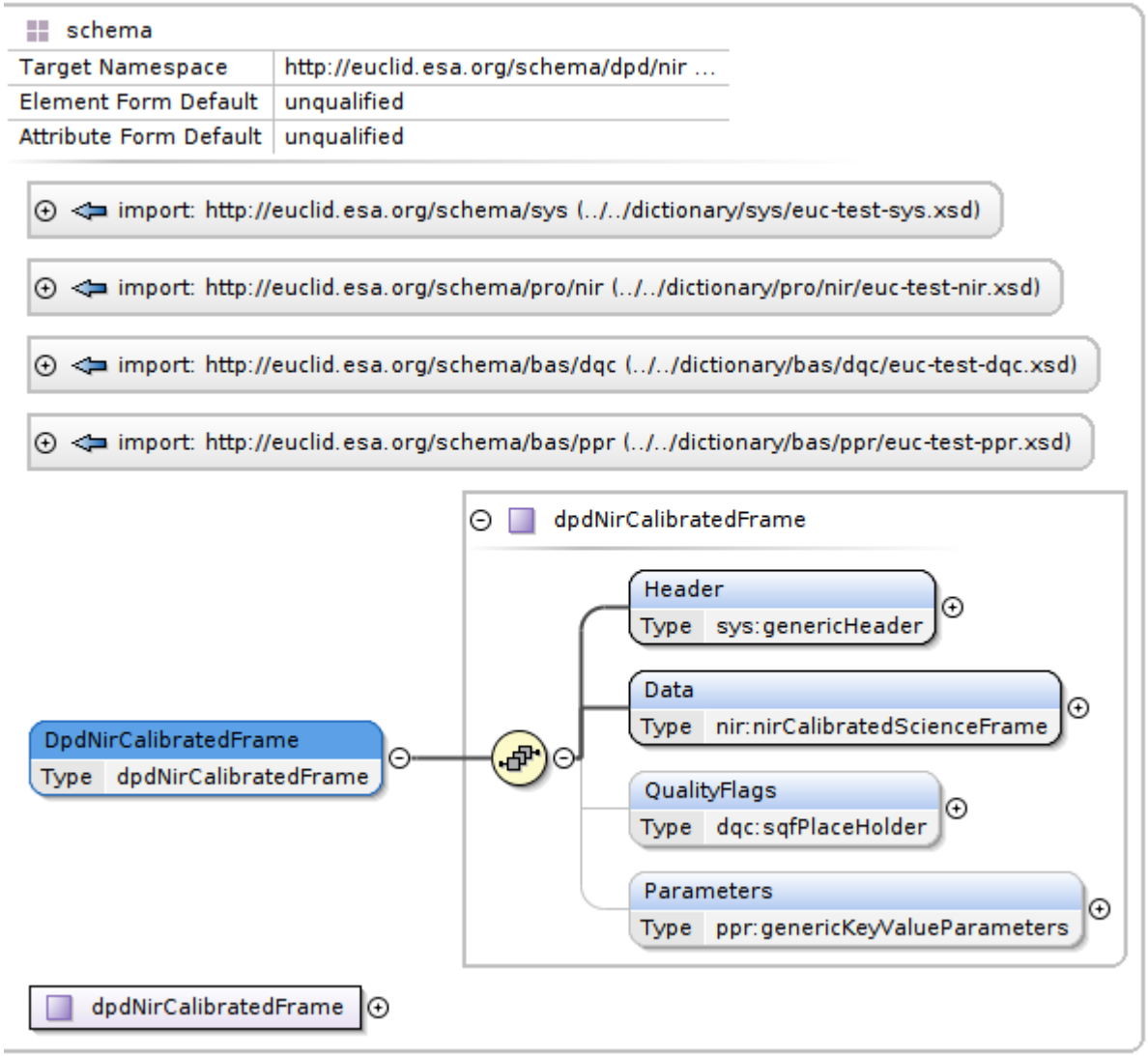
Metadata layers

- Metadata can be organized in layers. It can refer to:
  - raw data, e.g. coming from an instrument
  - information about the process of obtaining the raw data
  - derived data products
- This allows distinguishing different layers (or chains) of metadata: primary, secondary, tertiary, and so forth.
- Example with the satellite imaging domain: raw images taken by instruments in the satellite are sent to the ground stations.
  - Primary metadata includes:
    - the times when images were obtained and transferred
    - the instrument used to acquire them
    - The position to which each image refers
  - Secondary metadata:
    - Checking for gaps in the acquired data
    - Grouping of primary metadata information for a given instrument
    - Quality of the data in a given time period
    - Statistical summaries

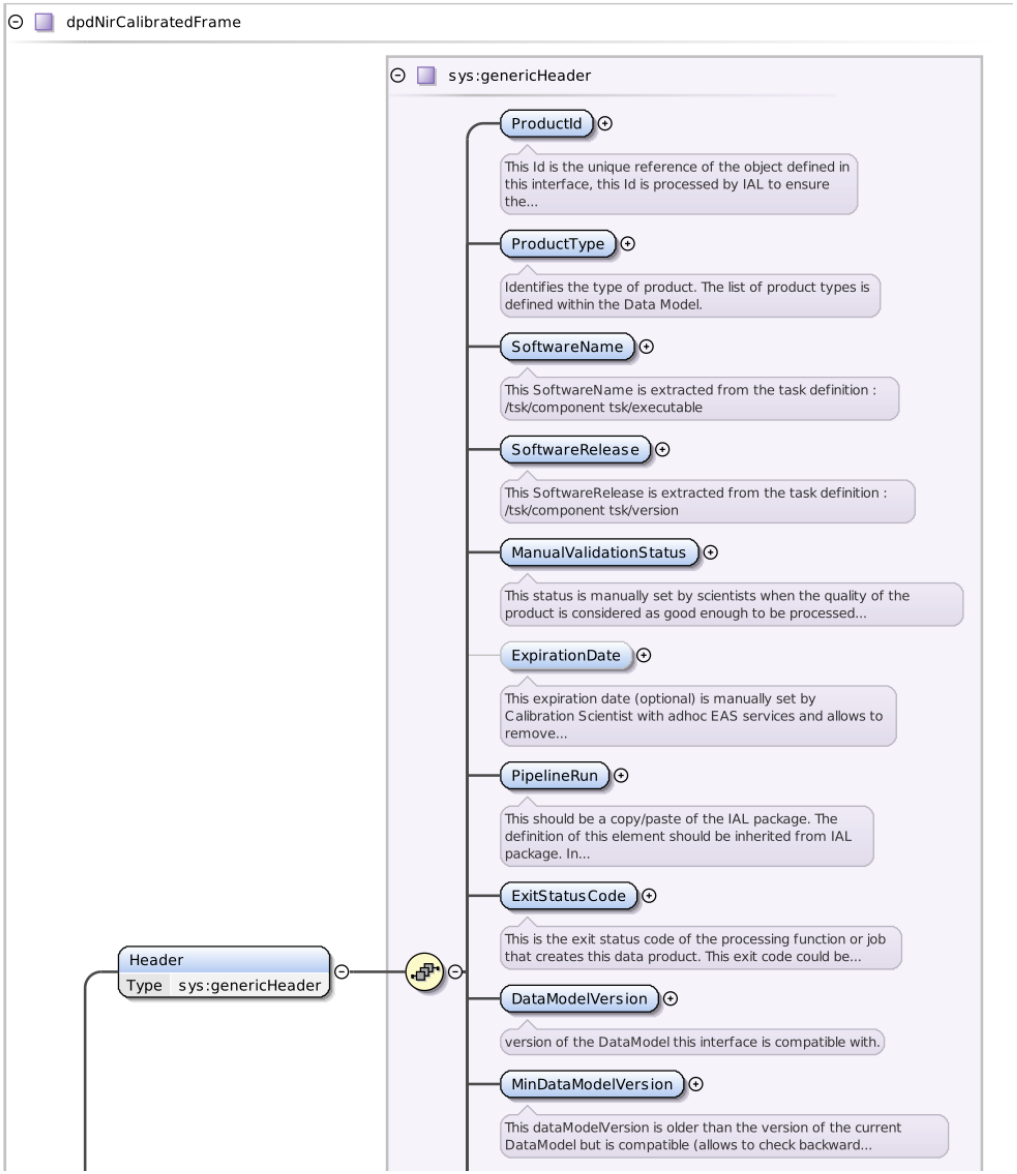
Example: Euclid project data model



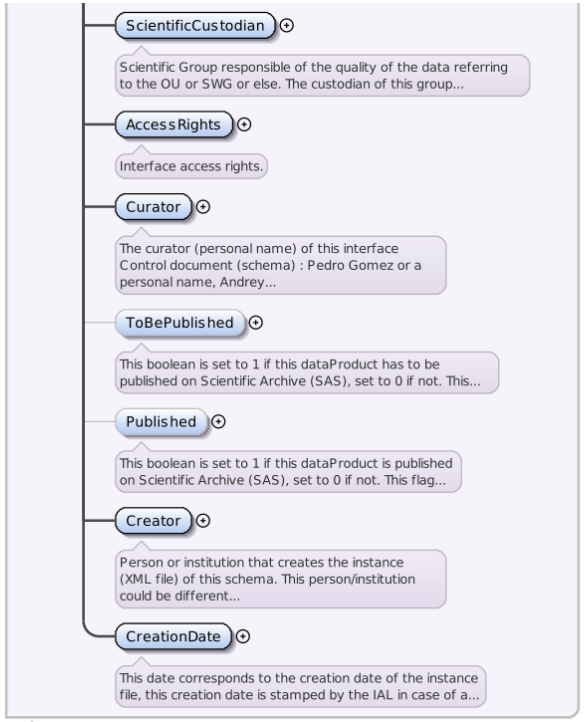
Example: Euclid project data model

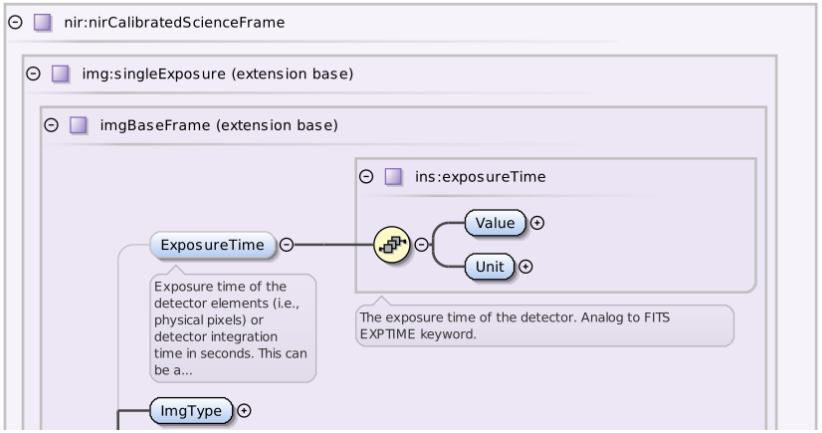


Example: Euclid project data model

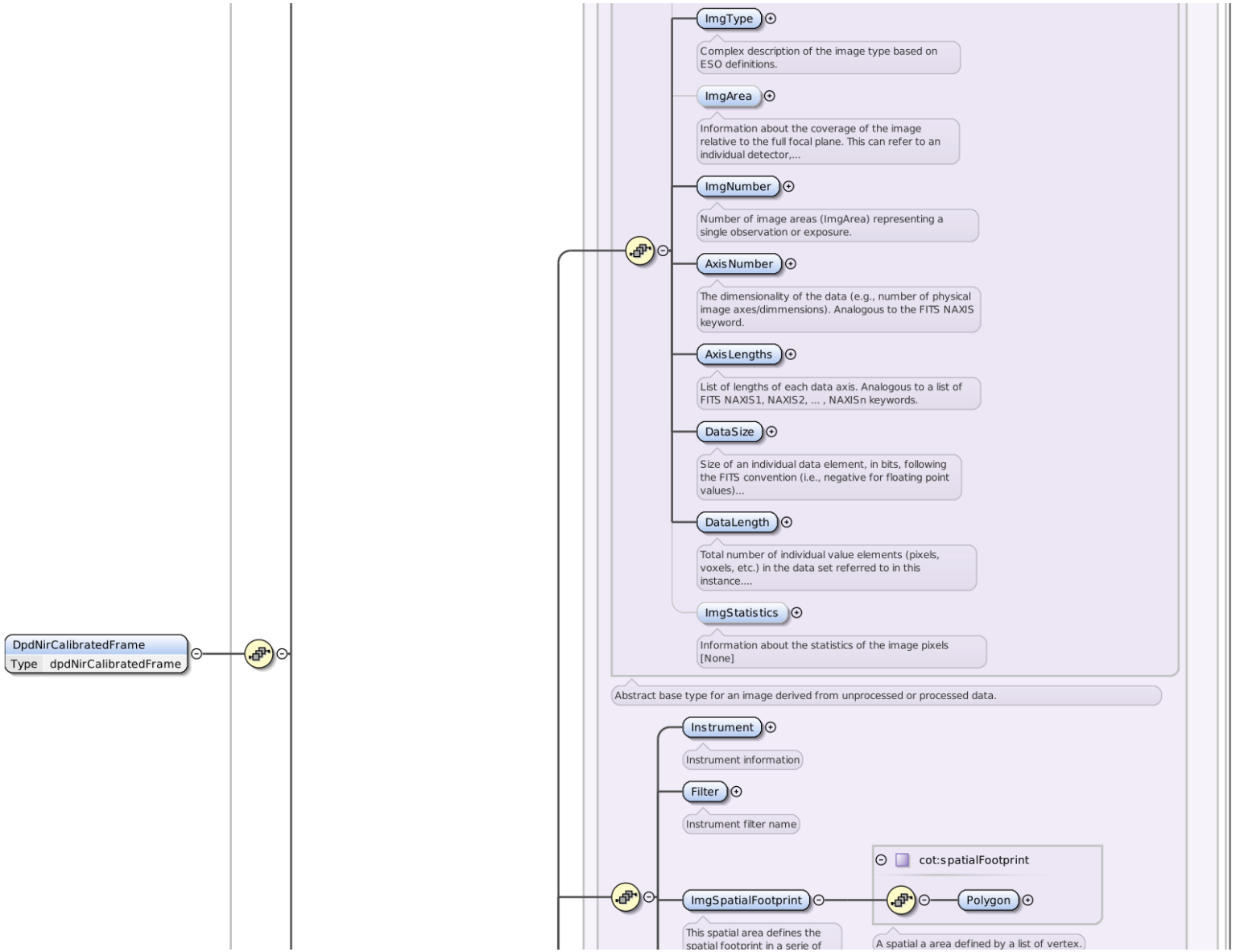


Example: Euclid project data model

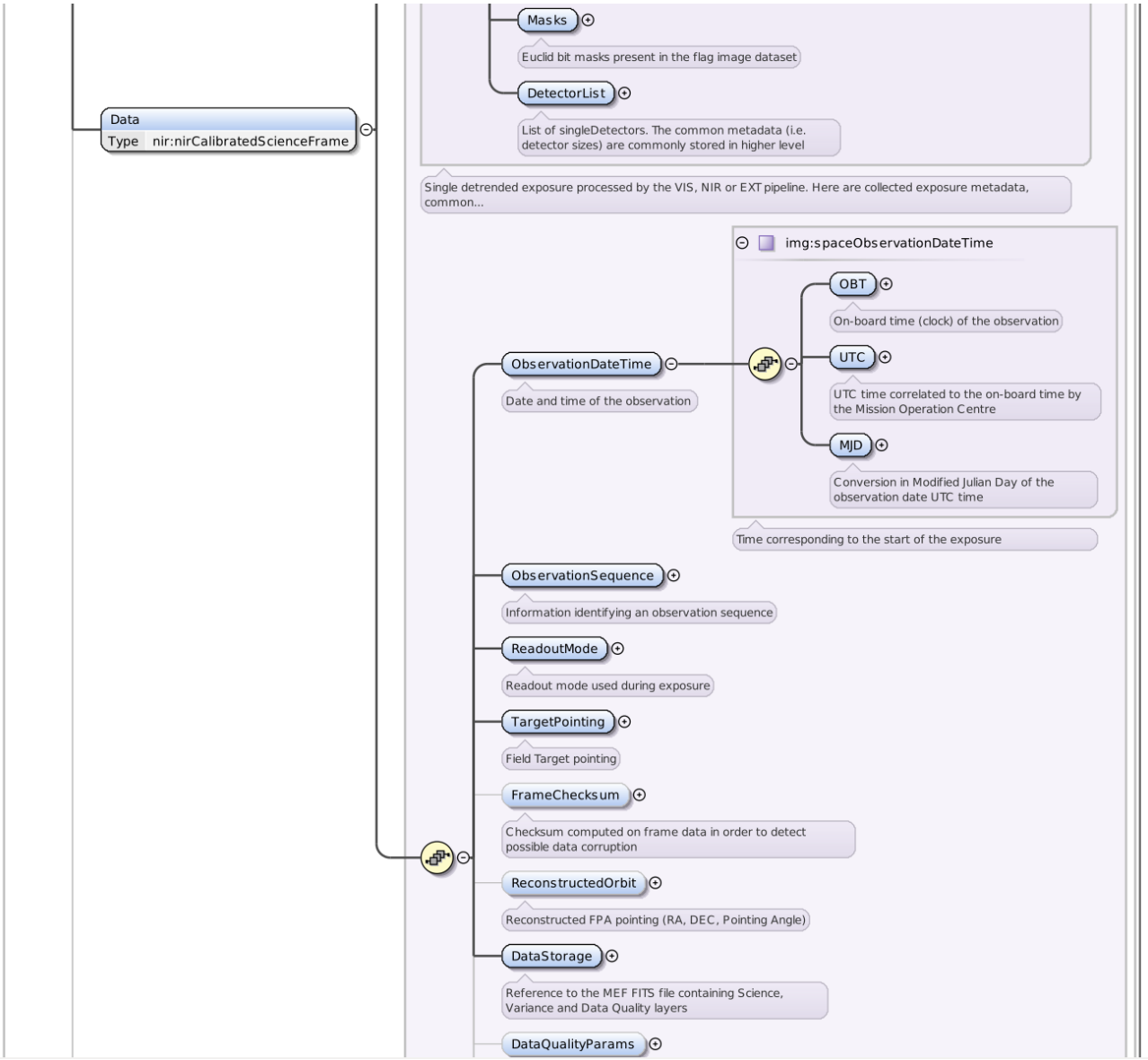




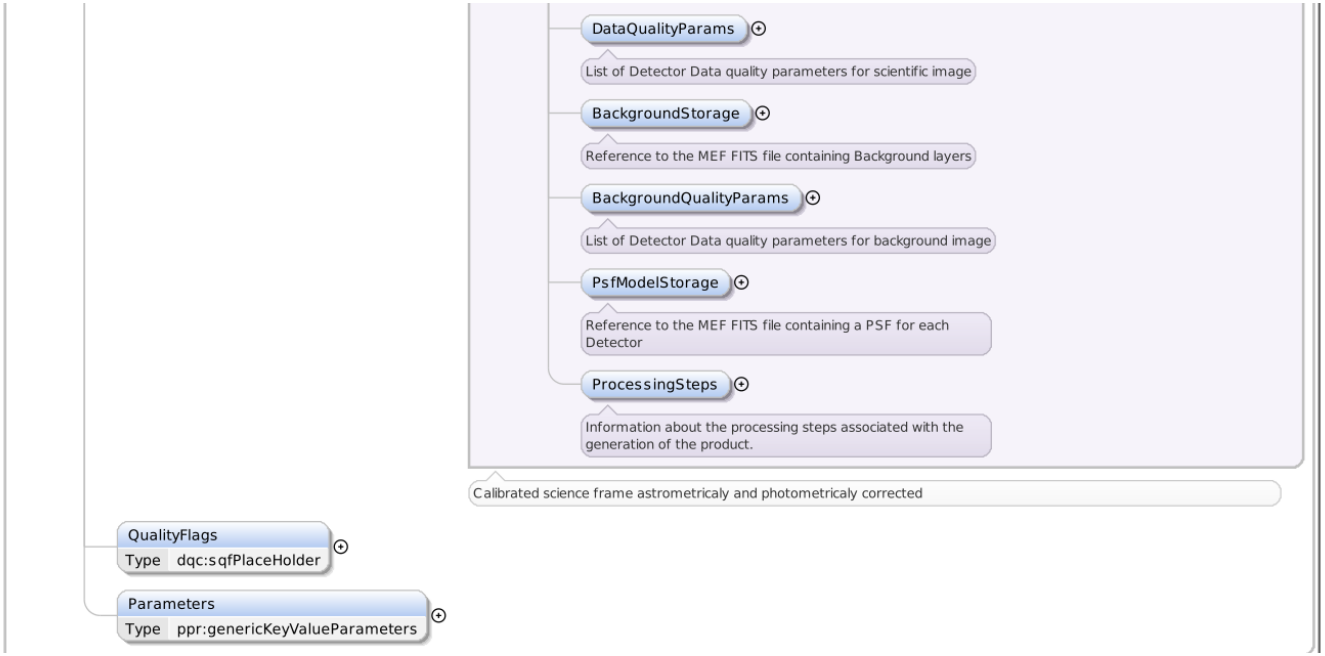
Example: Euclid project data model



Example: Euclid project data model



Example: Euclid project data model



Ontologies

- Together with controlled vocabularies and thesauri, **ontologies** have become one of the most common means to specify the structure of metadata in scientific applications
- Ontologies are normally defined as “formal, explicit specifications of shared conceptualizations”
- A *conceptualization* is an abstract model of some phenomenon in the world derived by having identified the relevant concepts of that phenomenon.
- *Explicit* means that the type of concepts used, and the constraints on their use, are explicitly defined
- *Formal* refers to the fact that the ontology should be machine readable
- *Shared* reflects the notion that an ontology captures consensual knowledge; that is, it is not private view for some individual, but accepted by a group

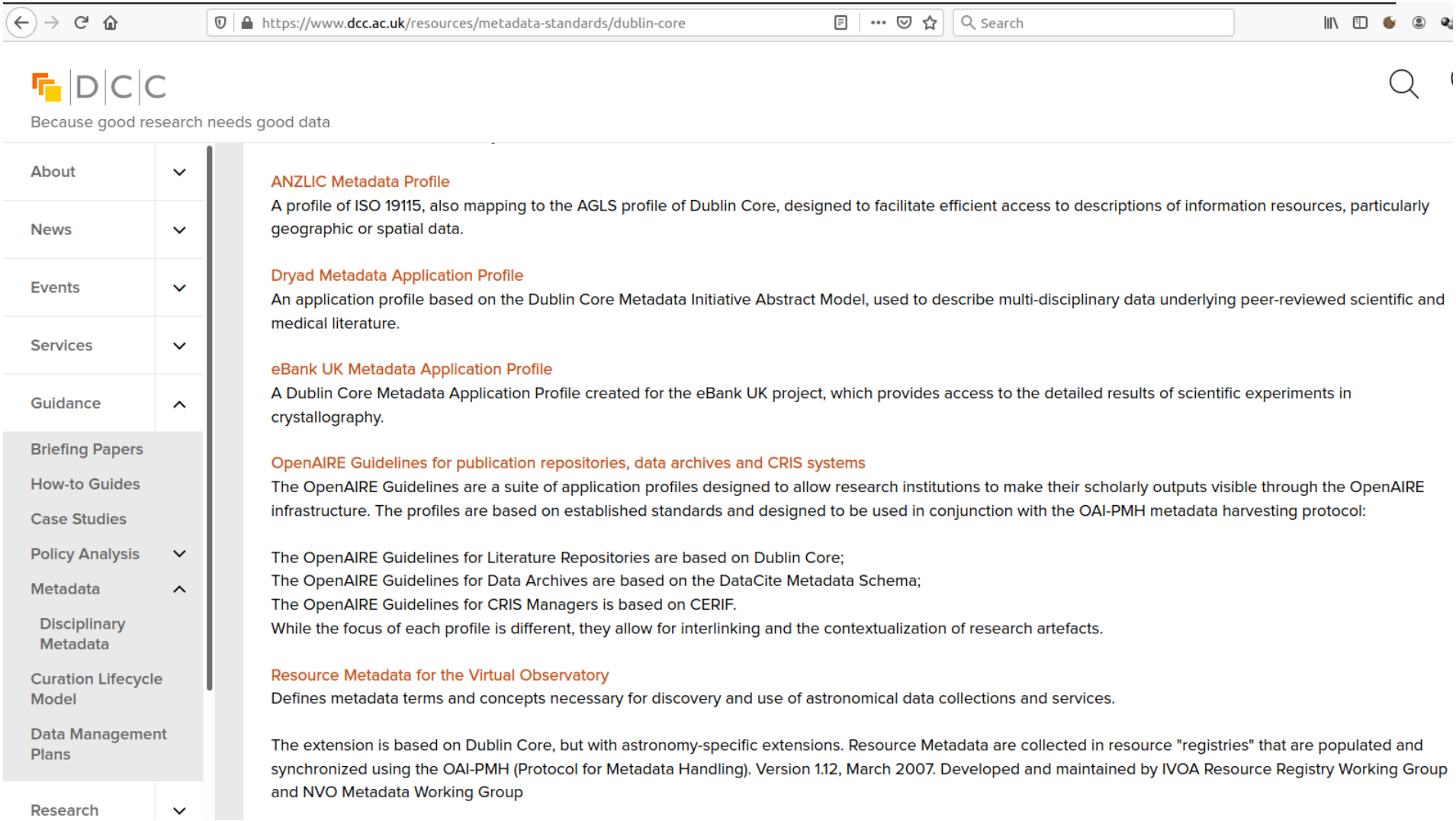
Ontologies formalization

- Not all ontologies have the same degree of formality;
- Given this fact, ontologies are usually classified either as lightweight or heavyweight.
- An example of the former would be *Dublin Core*, which is being widely used to specify simple characteristics of electronic resources
  - it specifies a predefined set of features such as creator, date, contributor, description, format, and the like
- An example of the latter is the Ontology of Astronomical Object Types (https://www.ivoa.net/documents/Notes/AstrObjectOntology/20100117/NOTE-AstrObjectOntology-1.3-20100117.html)
- Lightweight ontologies can be specified in simpler formal ontology languages like the **Resource Description Framework** (RDF) Schema
- Heavyweight ontologies require more complex languages like the **Web Ontology Language** (OWL)

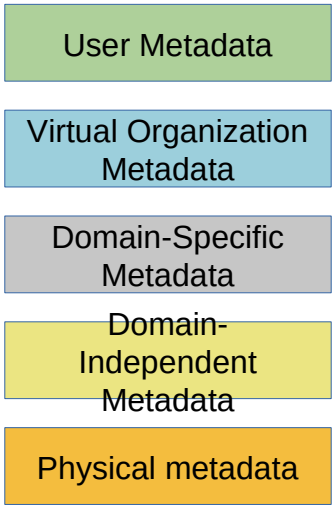
Metadata schema and attributes

- Metadata attributes that are elements of a metadata schema can encompass a variety of information
- Some metadata is **application independent**, such as the creation time, and author, as described in Dublin Core
- Other metadata is **application dependent** and may include attributes such as the duration of an experiment, temperature of the device, and others
- Many applications have expanded the Dublin Core schema to include application-dependent attributes (see next slide)
- Definitions provided in the Dublin Core or its extensions may be instantiated in a variety of standard forms, e.g. XML, and with a variety of mechanisms, e.g. OWL, or RDBMSs
  - Consequently, the **exact names** and rendering of the **values** may depend on the particular form in which they are represented

Dublin core extensions



Metadata types



- At the lowest level , **physical metadata** includes information about the physical storage systems as well as replica location metadata
- **Domain-Independent metadata** includes generic attributes, such as logical names, creator, modifier, data content, authorization, etc.
- **Domain-Specific metadata** attributes are often defined by metadata ontologies developed by the application communities
- A **virtual organization** that includes multiple scientific institutions may define attributes for characterizing data sets
- **Individual users** may want to associate metadata attributes such as annotations to data items or collections

Domain-Independent metadata

- We can identify a number of logical categories for domain-independent metadata
- In the following we provide categories related to data file handling:
  - Logical File metadata
  - Logical collection metadata
  - Logical view metadata
  - Authorization metadata
  - Audit metadata
  - Creation and transformation history metadata (provenance)

Approaches to data integration

- More and more domain-specific data management infrastructures are built to allow users **easy access to scientific data**, often through comprehensive web portals
- Traditional **data integration** basically follows a schema-matching approach in which related schema components (relations and attributes) from different sources are identified and homogenized
- Integration aims at providing a single conceptual view over the data managed at sources
- Using this view, the data can either be physically integrated at a single site (physical integration of data)
- Or the data can be queried in a uniform and transparent fashion. This approach results in a **federated** or multi-database system (logical integration)

Integration through Interoperability

- Interoperability among heterogeneous and **distributed data sources** is a fundamental requirement not only in the context of scientific data management, but in any type of distributed computing infrastructure.
- **Interoperability** is generally defined as “the ability of two or more systems or components to exchange information and to use the information that has been exchanged.”
- Interoperability means that systems can **exchange information** and data using **standard protocols** and **formats**
- Interoperability among data repositories and applications has become a main driver to facilitate scientific data management and exploration on a large scale
- Grid computing infrastructures have significantly contributed to this development
- A more recent trend in these science initiatives is to increase interoperability aspects through **service oriented science**

Service oriented architectures

- **Service-oriented architecture** (SOA) allows an effective cooperation among data sources and data processing components hosted at different organizational units
- SOA supports reusability and interoperability of software and service components on the Web, thus increasing the efficiency of developing and composing new services
- In a SOA-based system, all data and process components are modeled as **Web services**
- Web services can be implemented using different technologies:
  - Web Services Description Language (WSDL) and Simple Objects Access Protocol (SOAP)
  - As RESTFul web services
  - XML-RPC

Registry (and catalog) services

- As scientific data are accumulating at an ever-increasing speed, it is very difficult if not impossible for users to know exactly the details of all the data that might be relevant to their project
- A **registry service** enables users—human or software—to locate, access, and make use of resources in an open, distributed system
  - it facilitates the retrieval, storage, and management of many kinds of resource descriptions
- Based on SOA, a registry service must support some fundamental interactions:
  - **publishing** resource descriptions so that they are accessible to prospective users
  - **discovering** resources of interest according to some set of search criteria
  - and then **interacting** with the **resource provider** to access the desired resources
- The terms ‘catalogue’ and ‘registry’ are often used interchangeably
- A registry is a specialized catalogue that exemplifies a formal registration process
- A registry is typically maintained by a registration authority, who assumes responsibility for complying with a set of policies and procedures for accessing and managing registry content

Catalog service example