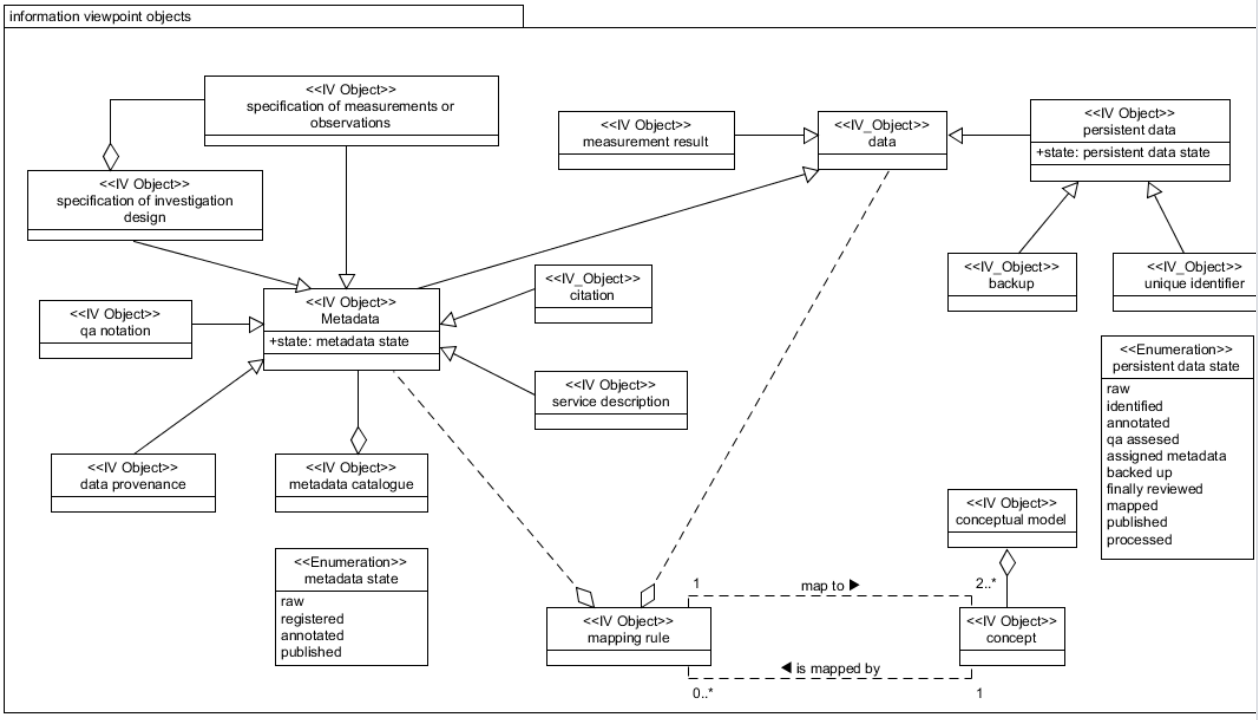
IV Information Objects

The IV of the ENVRI RM defines two main types of information objects: Data and Metadata.

Information objects are used to model the various types of data and metadata manipulated by the RI. The IV information objects can be grouped as follows.

- Data: research data processed by the RI:
  - Persistent data data
  - Unique identifiers for the data identification
  - Backup (of data)
- Metadata: data typically related to the design of observation and measurement models, complements data by providing more precise details.
  - Design specification of the observation and measurement
  - Description of the measurement procedure
  - Quality Assurance (QA) annotations
  - Concepts from a conceptual model, e.g. an ontology
  - Mapping rules which are used for the model-to-model transformations
  - Provenance records
  - Management metadata(The data used to identify the states of data and metadata objects)

Information Object Types



Notation

Information Object Definitions

backup

A copy of (persistent) data so it may be used to restore the original after a data loss event.

mapping rule

Configuration directives used for model-to-model transformation.

Mapping rules can be transformation rules for:

- arithmetic values (mapping from one unit to another)
  from linear functions like k.x + d to multivariate functions

- ordinal and nominal values
  e.g. transforming classifications according to a classification system A to classification system B
- data descriptions (metadata or **Semantic Annotation** or QA annotation)
- parameter names and descriptions (can be n:m)
- method names and descriptions
- sampling descriptions

citation

A published, resolvable, token linking to a persistent data object via an identifier.

In information technology terms, a citation is a reference to published data which may include the information related to:

- the data source(s)
- the owner(s) of the data source(s)
- a description of the evaluation process, if available
- a timestamp marking the access time to the data sources, thus reflecting a certain version
- the equipment used for collecting the data (individual sensor or sensor network)

It is important that the citation is resolvable, which means that the identifiers point to live data sets and that the meaning of the items above are made clear.

concept

Identifier, name and definition of the meaning of a thing (abstract or real thing). Human readable definition by sentences, machine readable definition by relations to other concepts (machine readable sentences). It can also be meant for the smallest entity of a conceptual model. It can be part of a flat list of concepts, a hierarchical list of concepts, a hierarchical thesaurus or an ontology.

conceptual model

A collection of concepts, their attributes and their relations. It can be unstructured or structured (e.g. glossary, thesaurus, ontology). Usually the description of a concept and/or a relation defines the concept in a human readable form. Conceptual models can also be represented in machine readable formats, for instance RDFS or OWL. Those sentences can be used to construct a self description. It is common practice to provide both the human readable description and the machine readable description within the same system. In this sense, a conceptual model can also be seen as a collection of human and machine readable sentences. They can be local, developed within a project, or global, accepted and used by a wider community (such as GEMET or OBOE). Conceptual models can be used to annotate data (e.g. within a network of triple stores).

data

Research data processed by the RI. This is the base information object class from which all other information objects are derived

data provenance

Metadata that traces the origins of data and records all state changes of data during their lifecycle and their movements between storages.

A creation of an entry into the data provenance records triggered by any actions typically contains:

- date/time of action;
- actor;
- type of action;
- data identification.

Data provenance system is an annotation system for managing data provenances. Usually unique identifiers are used to refer the data in their different states and for the description of the different states.

measurement result

Quantitative, qualitative, or cataloguing determinations of magnitude, dimension, and uncertainty to the outputs of observation instruments, sensors, sensor networks, human observers and observer networks.

metadata

Data about data, in scientific applications is used to describe, explain, locate, or make it easier to retrieve, use, or manage a data resource.

There have been numerous attempts to classify the various types of metadata. As one example, NISO (National Information Standards Organisation) distinguishes between three types of metadata based on their functionality: Descriptive metadata, which describes a resource for purposes, such as discovery and identification; Structural metadata, which indicates how compound objects are put together; and Administrative metadata, which provides information to help manage a resource. But this is not restrictive. Different applications may have different ways to classify their own metadata.

Metadata is generally encoded in a metadata schema which defines a set of metadata elements and the rules governing the use of metadata elements to describe a resource. The characteristics of metadata schema normally include: the number of elements, the name of each element, and the meaning of each element. The definition or meaning of the elements is the semantics of the schema, typically the descriptions of the location, physical attributes, type (i.e., text or image, map or model), and form (i.e., print copy, electronic file). The value of each metadata element is the content. Sometimes there are content rules and syntax rules. The content rules specify how content should be formulated, representation constraints for content, allowable content values and so on. And the syntax rules specify how the elements and their content should be encoded. Some popular syntaxes used in scientific applications include:

- HTML (Hyper-Text Markup Language): **www.w3.org/MarkUp/**
- XML (eXtensible Markup Language): **www.w3.org/XML/**
- RDF (Resource Description Framework): **www.w3.org/RDF/**
- OWL (Web Ontology Language): **www.w3.org/2001/sw/**
- SGML (Standard Generalised Markup Language): **www.w3.org/MarkUp/SGML/**
- MARC (Machine Readable Cataloging): **www.loc.gov/marc/**
- MIME (Multipurpose Internet Mail Extensions): **www.ukoln.ac.uk/metadata/resources/mime/**

- DIME (Direct Internet Message Encapsulation): **xml.coverpages.org/draft-nielsen-dime-01.txt**

Such syntax encoding allows the metadata to be processed by a computer program.

Many standards for representing scientific metadata have been developed within disciplines, sub-disciplines or individual project or experiments. Some widely used scientific metadata standards include:

- Dublin Core: **purl.oclc.org/metadata/dublin_core/**
- CERIF (Common European Research Information Format): **www.eurocris.org**
- ISO 11179: **metadata-stds.org/11179/**

  ISO 19115 (by iso-tc 211): **www.isotc211.org**
- FGDC (The Federal Geographic Data Committee): **www.fgdc.gov/standards**
- INSPIRE: **inspire.jrc.ec.europa.eu**
- ISO 19115, Geographic information - metadata standard (metadata model closely related to INSPIRE) **www.iso.org**
- DDI (Data Documentation Initiative): **www.ddialliance.org**
- TEI (The Text Encoding Initiative): **www.tei-c.org**
- METS (Metadata Encoding and Transmission Standard): **www.loc.gov/standards/mets**
- MODS (Metadata Object Description Schema): **www.loc.gov/standards/mods/**
- OAIS (Reference Model for an Open Archival Information System)

Two aspects of metadata give rise to the complexity in management:

- Metadata are data, and data become metadata when they are used to describe other data. The transition happens under particular circumstances, for particular purposes, and with certain perspectives, as no data are always metadata. The set of circumstances, purposes, or perspectives for which some data are used as metadata is called the ‘context’. So metadata are data about data in some ‘context’.
- Metadata can be layered. This happens because data objects may move to different stages during their life in a digital environment requiring their association to different layers of metadata at each stage.

Metadata can be fused with the data. However, in many applications, such as a provenance system or a distributed satellite image annotation system, the metadata and data are often created and stored separately, as they may be generated by different users, in different computing processes, stored at different locations and in different types of storage. Often, there is more than one set of metadata related to a single data resource, e.g. when the existing metadata becomes insufficient, users may design new templates to make another metadata collection. Efficient software and tools are required to facilitate the management of the linkage between metadata and data. Such linkage relationship between metadata and data are vulnerable to failures in the processes that create and maintain them, and to failures in the systems that store their representations. It is important to devise methods that reduce these failures.

metadata catalogue

A collection of metadata, usually established to make the metadata available to a community. A metadata catalogue can be exposed through an access service.

metadata state

metadata state is an object property that determines the set of all sequences of actions (or traces) in which the metadata object can participate, at a given instant in time (as defined in ODP, ISO/IEC 10746-2).

In their lifecycle, metadata may have the states described in the following table.

State	Description
raw	metadata which are not yet registered or organised in a catalogue. Raw metadata are not shareable in this status.
registered	metadata which have been stored in a metadata catalogue.
annotated	metadata that are associated to concepts, describing their meaning
published	metadata made available to the public, the outside world. Metadata registered within public catalogues.

persistent data

Data is the representations of information dealt with by information systems and users thereof (as defined in ODP, ISO/IEC 10746-2). Persistent Data denotes data that are persisted (stored for the long-term).

persistent data state

Persistent Data state is an object property that determines the set of all sequences of actions (or traces) in which the object can participate, at a given instant in time (as defined in ODP, ISO/IEC 10746-2). The persistent data states and their changes as effects of actions are illustrated as **I V States**.

In their lifecycle, persistent data may have the states described in the following table.

State	Description
raw	data derived from the primary results of observations or measurements
identified	data which has been assigned a unique identifier
annotated	data that are associated to concepts, describing their meaning
qa assessed	data that have undergone checks and are associated with descriptions of the results of those checks.

assigned metadata	data that are associated to metadata which describe those data
backed up	data that of which an identical copy has been stored securely
finally reviewed	data that have undergone a final review and therefore will not be changed any more
mapped	data that are mapped to a certain conceptual model
published	data that are presented to the outside world
processed	data that have undergone a processing (evaluation, transformation)

Note

The state 'raw' refers to data as received into the ICT elements of the research infrastructure. Some pre-processing may or may not have been carried out closer to where measurements and observations were made

These states are referential states. The instantiated chain of data lifecyle can be expressed in data provenance.

qa notation

Notation of the result of a Quality Assessment. This notation can be a nominal value out of a classification system up to a comprehensive (machine readable) description of the whole QA process.

In practice, this can be:

- simple flags like "valid" / "invalid" up to comprehensive descriptions like
- "data set to invalid by xxxxxx on ddmmyy because of yyyyyyy"

QA notation can be seen as a special annotation. To allow sharing with other users, the QA notation should be unambiguously described so as to be understood by others or interpretable by software tools.

service description

Description of services and processes available for reuse. The description is needed to facilitate usage. The service description usually includes a reference to a service or process making it available for reuse within a research infrastructure or within an open network like the Internet. Usually such descriptions include the accessibility of the service, the description of the interfaces, the description of behavior and/or implemented algorithms. Such descriptions are usually done along service description standards (e.g. WSDL, web service description language). Within some service description languages, semantic descriptions of the services and/or interfaces are possible (e.g. SAWSDL, Semantic Annotations for WSDL).

specification of investigation design

This is the background data needed to understand the overall goal of the measurement or observation. It could be the sampling design of observation stations, the network design, the description of the setup parameters (interval of measurements) and so on. It usually contains important data for the allowed evaluations of research results (e.g. the question of whether a sampling design was done randomly or by stratification determines which statistical methods can be applied).

Investigations (and hence measurement and observation results) need not be quantitative. They can also be qualitative results (like "healthy", "ill") or classifications (like assignments to biological taxa). It is important for data processing to know whether they are quantitative or qualitative.

The specification of investigation design can be seen as part of metadata or as part of the **Semantic Annotation**. It is important that this description follows certain standards and it is desirable that the description is machine readable.

specification of measurements or observations

The description of the measurement/observation which specifies:

- what is measured/observed;
- how it is measured/observed (including processes/metods and instruments to be used);
- by whom it is measured/observed (including project, organisation and experimenter/observer profile); and
- what the temporal design is (single / multiple measurements / interval of measurement etc.)

Note

This specification can be included as metadata or as **Semantic Annotation** of the scientific data to be collected. It is important that such a design specification is both explicit and correct, so as to be understood or interpreted by external users or software tools. Ideally, a machine readable specification is desired.

unique identifier (UID)

With reference to a given type of data, objects a unique identifier (UID) is any identifier which is guaranteed to be unique among all identifiers used for those type of objects and for a specific purpose.

There are 3 main generation strategies:

- serial numbers, assigned incrementally;
- random numbers, selected from a number space much larger than the maximum (or expected) number of objects to be identified. Although not really unique, some identifiers of this type may be appropriate for identifying objects in many practical applications and are, with abuse of language, still referred to as "unique";
- names or codes allocated by choice which are forced to be unique by keeping a central registry.

The above methods can be combined, hierarchically or singly, to create other generation schemes which guarantee uniqueness.

In many cases, a single object may have more than one unique identifier, each of which identifies it for a different purpose. For example, a single object can be assigned with the following identifiers:

- global: unique for a higher level community
- local: unique for the subcommunity

The critical issues of unique identifiers include but not limited to:

- long term persistence – without efficient management tools, UIDs can be lost;
- resolvability -- without efficient management tools, the linkage between a UID and its associated contents can be lost.