

ENVRI

Services for the Environmental Community

D3.4 ENVRI Reference Model V1.1

Document identifier:	D3.4 ENVRI Reference Model V1.1
Date:	31/08/2013
Activity:	WP3
Lead Partner:	CU
Document Status:	FINAL
Dissemination Level:	PUBLIC
Document Link:	www.envir.eu/rm

ABSTRACT

It has been recognised that all ENVRI research infrastructures, although are very diverse, have some common characteristics, enabling them potentially to achieve a level of interoperability through the use of common standards for various functions. The objective of ENVRI Reference Model is to develop common ontological framework and standards for the description and characterisation of computational and storage infrastructures in order to achieve seamless interoperability between the heterogeneous resources of different infrastructures.

The ENVRI Reference Model is a work-in-progress, hosted by the ENVRI project, intended for interested parties to directly comment on and contribute to.



1. COPYRIGHT NOTICE

Copyright © Members of the ENVRI Collaboration, 2011. See www.ENVRI.eu for details of the ENVRI project and the collaboration. ENVRI (“**Common Operations of Environmental Research Infrastructures**”) is a project co-funded by the European Commission as a Coordination and Support Action within the 7th Framework Programme. ENVRI began in October 2011 and will run for 3 years. This work is licensed under the Creative Commons Attribution-Noncommercial 3.0 License. To view a copy of this license, visit <http://creativecommons.org/licenses/by/3.0/> or send a letter to Creative Commons, 171 Second Street, Suite 300, San Francisco, California, 94105, and USA. The work must be attributed by attaching the following reference to the copied elements: “Copyright © Members of the ENVRI Collaboration, 2011. See www.ENVRI.eu for details of the ENVRI project and the collaboration”. Using this document in a way and/or for purposes not foreseen in the license, requires the prior written permission of the copyright holders. The information contained in this document represents the views of the copyright holders as of the date such views are published.

2. DELIVERY SLIP

	Name	Partner/Activity	Date
From			
Reviewed by	Moderator: Reviewers:	Leonardo Candela (CNR) Ari Asmi (UNEL)	22/04/13
Approved by			

3. DOCUMENT LOG

Issue	Date	Comment	Author/Partner
1.0	31/03/13	The draft version of the ENVRI Reference Model which describes an abstract model with a set of concepts and terms to capture a minimal sets of requirements of ESFRI environmental research infrastructures.	Yin Chen (CU) Paul Martine (UEDIN) Herbert Schentz (EAA) Barbara Magagna(EAA) Zhiming Zhao(UoV) Alex Hardisty (CU) Alun Preece (CU) Malcolm Atkinson (UEDIN)
2.0	30/04/13	Internally reviewed version to be approved by project management and submitted to the Commission.	Yin Chen (CU) Paul Martine (UEDIN) Herbert Schentz (EAA) Barbara Magagna(EAA) Zhiming Zhao(UoV)



3.0	31/08/13	Modifications applied according to review opinions	Yin Chen (CU) Paul Martine (UEDIN) Herbert Schentz (EAA) Barbara Magagna(EAA) Zhiming Zhao(UoV) Alex Hardisty (CU) Alun Preece (CU) Malcolm Atkinson (UEDIN)
-----	----------	--	---

4. APPLICATION AREA

This document is a formal deliverable for the European Commission, applicable to all members of the ENVRI project, beneficiaries and Joint Research Unit members, as well as its collaborating projects.

5. DOCUMENT AMENDMENT PROCEDURE

Amendments, comments and suggestions should be sent to the authors.

6. TERMINOLOGY

A complete project glossary is provided at the following page: <http://www.ENVRI.eu/glossary>. The terminology of concepts and terms defined in this document is provided in Appendix A.

7. PROJECT SUMMARY

Frontier environmental research increasingly depends on a wide range of data and advanced capabilities to process and analyse them. The ENVRI project, “Common Operations of Environmental Research infrastructures” is a collaboration in the ESFRI Environment Cluster, with support from ICT experts, to develop common e-science components and services for their facilities. The results will speed up the construction of these infrastructures and will allow scientists to use the data and software from each facility to enable multi-disciplinary science.

The target is on developing common capabilities including software and services of the environmental e-infrastructure communities. While the ENVRI infrastructures are very diverse, they face common challenges including data capture from distributed sensors, metadata standardisation, management of high volume data, workflow execution and data visualisation. The common standards, deployable services and tools developed will be adopted by each infrastructure as it progresses through its construction phase.

Two use cases, led by the most mature infrastructures, will focus the development work on separate requirements and solutions for data pre-processing of primary data and post-processing toward publishing.

The project will be based on a common reference model created by capturing the semantic resources of each ESFRI-ENV infrastructure. This model and the development driven by the test-bed deployments result in ready-to-use systems which can be integrated into the environmental research infrastructures.



ENVRI Common Operations of Environmental Research Infrastructures

The project puts emphasis on synergy between advanced developments, not only among the infrastructure facilities, but also with ICT providers and related e-science initiatives. These links will facilitate system deployment and the training of future researchers, and ensure that the inter-disciplinary capabilities established here remain sustainable beyond the lifetime of the project.

8. EXECUTIVE SUMMARY

This document describes the ENVRI Reference Model (ENVRI-RM) which is an abstract model with a set of concepts and terms capturing a set of requirements of environmental research infrastructures. Built on top of the Open Distributed Processing (ODP) framework, the Reference Model defines functional elements, data flow and dependencies that are common in ENVRI research infrastructures. The Reference Model can be used as the foundation for building reference architectures, and concrete implementations can be derived.

The ENVRI Reference Model is a work-in-progress, hosted by the ENVRI project, intended for interested parties to directly comment on and contribute to.

9. HOW TO READ

The document is organised as follows:

Section 1 introduces the motivation and background knowledge of the ENVRI-RM.

Section 2 presents an overview of the ENVRI-RM, which consists of 5 common subsystems identified in a pre-study work. The concepts of these entities and their relationship is discussed.

Section 3 detailed describes the ENVRI Reference Model from ODP three Viewpoints, the Science, Information and Computational.

Section 4 concludes this work.

Appendixes are not part of the model, and provided for the convenience of the reader.

Appendix A is the glossary of the document which consists of all concepts and terms defined throughout the ENVRI-RM.

Appendix B presents the full list of the required functionalities which is the result of the investigation of the common requirements of ENVRI Research Infrastructures.

Appendix C provides detailed dynamic schemata specified in Information Viewpoint (Section 4.2.3).

The intended audience of this document is the ENVRI community as well as other organisations or individuals that are interested in understanding the top level technical architecture which underpins the construction of such an architecture. In particular, the intended primary audience of this documents includes:

- Research Infrastructure implementation teams:
 - Architects, designers, and integrators;
 - Engineers – to enable them to be able to drill down directly to find required knowledge.
- Research Infrastructure Operations Teams, and
- Third party solution or component providers

The documents is also intended for Research Infrastructure leaders and Service Centre staffs.

The document can be read by others who want to better understand the ENVRI ongoing work, to gain understanding necessary to make contributions to the standardisation processes of environmental research infrastructures.

For the primary audience of the ENVRI-RM shall read the whole document.

For the leaders of research infrastructures, service centre staffs may want to read the introduction and background knowledge in **section 1** and model overview in **section 2**.

For readers who have general interests of the ENVRI reference model may want to read the **section 1** introduction.

10. REVISION HISTORY

This version (1.1) is the second major version, incorporating the following changes from the previous version (1.0):

- Revision of 3.2.1.1 Information Object
- Revision of 3.2.1.2 Information Action Types
- Revision of 3.2.21 Dynamic Schemata Overview
- Revision of 3.3 Computational Viewpoint
- Updated the related changes in figures and terminologies



TABLE OF CONTENTS

1	INTRODUCTION	9
1.1	Purpose and Scope	9
1.2	Rationale	9
1.3	Basis	11
1.4	Approaches	12
1.5	Conformance	12
1.6	Related Work	12
1.6.1	Related concepts	12
1.6.2	Related reference models	14
1.6.3	Other Related standards	16
2	MODEL OVERVIEW	17
2.1	Subsystems of Environmental Infrastructures	17
2.1.1	Data Acquisition	17
2.1.2	Data Curation	18
2.1.3	Data Access	18
2.1.4	Data Processing	18
2.1.5	Community Support	18
2.2	Subsystem Relationships	19
2.3	Common Functions within Common Subsystems	20
3	ENVRI REFERENCE MODEL	23
3.1	Science Viewpoint	23
3.1.1	Common Communities	23
3.1.2	Common Community Roles	24
3.1.3	Common Community Behaviours	29
3.2	Information Viewpoint	32
3.2.1	Components	33
3.2.2	Dynamic Schemata	47
3.2.3	Static Schemata	53
3.2.4	Subsystems	55
3.3	Computational Viewpoint	58
3.3.1	Data Acquisition	62
3.3.2	Data Curation	64
3.3.3	Data Access	68
3.3.4	Data Processing	70
3.3.5	Community Support	71
3.3.6	Brokered Data Export	74
3.3.7	Brokered Data Import	74
3.3.8	Brokered Data Query	75
3.3.9	Citation	76
3.3.10	Internal Data Staging	76
3.3.11	Processed Data Import	77
3.3.12	Raw Data Collection	77
3.3.13	Instrument Integration	78
4	CONCLUSION AND FUTURE WORK	80



ENVRI Common Operations of Environmental Research Infrastructures

5 REFERENCES	81
APPENDIXES	83
A. Terminology and Glossary	83
A.1 Acronyms and Abbreviations	83
A.2 Terminology.....	84
B. Common Requirements of ENVRI Research Infrastructures	93
C. Dynamic Schemata in Details	97



TABLE OF FIGURES

Figure 2.1: Illustration of the major points-of-reference between different subsystems of the ENVRI-RM.....	19
Figure 2.2: Radial depiction of ENVRI-RM requirements with the minimal model highlighted.....	20
Figure 3.1: Common Communities.....	24
Figure 3.2: Roles in the Data Acquisition Community.....	25
Figure 3.3: Roles in the Data Curation Community.....	26
Figure 3.4: Roles in the Data Publication Community.....	27
Figure 3.5: Roles in the Data Service Provision Community.....	28
Figure 3.6: Roles in the Data Usage Community.....	28
Figure 3.7: Behaviours of the Data Acquisition Community.....	29
Figure 3.8: Behaviours of the Data Curation Community.....	30
Figure 3.9: Behaviours of the Data Publication Community.....	30
Figure 3.10: Behaviours of the Data Service Provision Community.....	31
Figure 3.11: Behaviours of the Data Usage Community.....	32
Figure 3.12: Information Objects.....	34
Figure 3.13: Information Objects Action Types.....	40
Figure 3.14: Instances of Information Objects.....	45
Figure 3.15: An Example of Data States Changes As Effects of Actions.....	46
Figure 3.16: Dynamic Schemata Overview.....	51
Figure 3.17: Tracing of Provenance.....	52
Figure 3.18: Constraints for Data Collection.....	53
Figure 3.19: Constraints for Data Integration.....	54
Figure 3.20: Constraints for Data Publication.....	55
Figure 3.21: Information Specification of Data Acquisition Subsystem.....	56
Figure 3.22: Information Specification of Data Curation Subsystem.....	57
Figure 3.23: Information Specification of Data Access Subsystem.....	58
Figure 3.24: Computational Specification of Data Acquisition Subsystem.....	62
Figure 3.25: Computational Specification of Data Curation Subsystem.....	64
Figure 3.26: Computational Specification of Data Access Subsystem.....	68
Figure 3.27: Computational Specification of Data Processing Subsystem.....	70
Figure 3.28: Computational Specification of Community Support Subsystem.....	72
Figure 3.29: Brokered Data Export.....	74
Figure 3.30: Brokered Data Import.....	75
Figure 3.31: Brokered Data Query.....	75
Figure 3.32: Citation.....	76
Figure 3.33: Internal Data Staging.....	76
Figure 3.34: Processed Data Import.....	77
Figure 3.35: Raw Data Collection.....	78
Figure 3.36: Instrument Integration.....	79



1 INTRODUCTION

1.1 Purpose and Scope

It has been recognised that all ENVRI research infrastructures, although very diverse, have some common characteristics, enabling them potentially to achieve a greater level of interoperability through the use of common standards for various functions. The objective of the ENVRI Reference Model is to develop a common ontological framework and standards for the description and characterisation of computational and storage infrastructures in order to achieve seamless interoperability between the heterogeneous resources of different infrastructures.

The ENVRI Reference Model serves the following purposes [1]:

- to provide a way for structuring thinking which helps the community to reach a common vision;
- to provide a common language which can be used to communicate concepts concisely;
- to help discover existing solutions to common problems;
- to provide a framework into which different functional components of research infrastructures can be placed, in order to draw comparisons and identify missing functionality.

This document describes the ENVRI Reference Model which:

- captures computational characteristics of data and operations that are common in ENVRI Research Infrastructures;
- establishes a taxonomy of terms, concepts and definitions to be used by the ENVRI community.

The Reference Model provides an abstract conceptual model; it does not impose a specific architecture nor does it impose any specific design decisions on the design of an infrastructure.

The *initial* model focuses on the urgent and important issues prioritised for ENVRI research infrastructures including data preservation, discovery and access, and publication. It defines a minimal set of computational functionalities to support these requirements. The core set will be extended incrementally over the course of the ENVRI project. The initial model does not cover engineering mechanisms or the applicability of existing standards or technologies.

1.2 Rationale

Environmental issues will dominate the 21st century [2]. Research infrastructures which provide advanced capabilities for data sharing, processing and analysis enable excellent research and play an ever-increasing role in the environmental sciences. The ENVRI project gathers 6 EU ESFRI¹ environmental infrastructures (ICOS², EURO-Argo³, EISCAT-3D⁴, LifeWatch⁵, EPOS⁶, and EMSO⁷)

¹ ESFRI, the European Strategy Forum on Research Infrastructures, is a strategic instrument to develop the scientific integration of Europe and to strengthen its international outreach.

² ICOS, <http://www.icos-infrastructure.eu/>, is a European distributed infrastructure dedicated to the monitoring of greenhouse gases (GHG) through its atmospheric, ecosystem and ocean networks.

³ EURO-Argo, <http://www.euro-argo.eu/>, is the European contribution to Argo, which is a global ocean observing system.

⁴ EISCAT-3D, <http://www.eiscat3d.se/>, is a European new-generation incoherent-scatter research radar for upper atmospheric science.

⁵ LifeWatch, <http://www.lifewatch.com/>, is an e-science Infrastructure for biodiversity and ecosystem research.

in order to develop common data and software services. The results will accelerate the construction of these infrastructures and improve interoperability among them by encouraging the adoption of the reference model. The experiences gained from this endeavour will also benefit the building of other advanced research infrastructures.

The primary objective of ENVRI is to agree on a reference model for joint operations. This will enable greater understanding and cooperation between users since fundamentally the model will serve to provide a universal reference framework for discussing many common technical challenges facing all of the ESFRI-ENV infrastructures. By drawing analogies between the reference components of the model and the actual elements of the infrastructures (or their proposed designs) as they exist now, various gaps and points of overlap can be identified [3].

The ENVRI Reference Model is based on the design experiences of the state-of-the-art environmental research infrastructures, with a view of informing future implementation. It tackles multiple challenging issues encountered by existing initiatives, such as data streaming and storage management; data discovery and access to distributed data archives; linked computational, network and storage infrastructure; data curation, data integration, harmonisation and publication; data mining and visualisation, and scientific workflow management and execution. It uses Open Distributed Processing (ODP), a standard framework for distributed system specification, to describe the model.

To our best knowledge there is no existing reference model for environmental science research infrastructures. This work intends to make a first attempt, which can serve as a basis to inspire future research explorations.

There is an urgent need to create such a model, as we are at the beginning of a new era. The advances in automation, communication, sensing and computation enable experimental scientific processes to generate data and digital objects at unprecedentedly great speeds and volumes. Many infrastructures are starting to be built to exploit the growing wealth of scientific data and enable multi-disciplinary knowledge sharing. In the case of ENVRI, most investigated RIs are in their planning / construction phase. The high cost attached to the construction of environmental infrastructures require cooperation on the sharing of experiences and technologies, solving crucial common e-science issues and challenges together. Only by adopting a good reference model can the community secure interoperability between infrastructures, enable reuse, share resources and experiences, and avoid unnecessary duplication of effort.

The contribution of this work is threefold:

- The model captures the computational requirements and the state-of-the-art design experiences of a collection of representative research infrastructures for environmental sciences. It is the first reference model of this kind which can be used as a basis to inspire future research.
- It provides a common language for communication to unify understanding. It serves as a community standard to secure interoperability.
- It can be used as a base to drive design and implementation. Common services can be provided which can be widely applicable to various environmental research infrastructures and beyond.

⁶ EPOS, <http://www.epos-eu.org/>, is a European Research Infrastructure on earthquakes, volcanoes, surface dynamics and tectonics.

⁷ EMSO, <http://www.emso-eu.org/management/>, is a European network of seafloor observatories for the long-term monitoring of environmental processes related to ecosystems, climate change and geo-hazards.



ENVRI Common Operations of Environmental Research Infrastructures

1.3 Basis

The ENVRI Reference Model is built on top of the Open Distributed Processing (ODP) framework [4-7]. ODP is an international standard for architecting open, distributed processing systems. It provides an overall conceptual framework for building distributed systems in an incremental manner.

The reasons for adopting the ODP framework in the ENVRI project come from three aspects:

- It enables large collaborative design activities;
- It provides a framework for specifying and building large or complex system which consists of a set of guiding concepts and terminology. This provides a way of thinking about architectural issues in terms of fundamental patterns or organising principles; and
- Being an international standard, ODP offers authority and stability.

ODP adopts the *object modelling* approach to system specification. ISO/IEC 10746-2 [5] includes the formal definitions of the concepts and terminology adopted from object models, which provide the foundation for expressing the architecture of ODP systems. The modelling concepts fall into three categories [4, 5]:

- Basic modelling concepts for a general object-based model;
- Specification concepts to allow designers to describe and reason about ODP system specifications;
- Structuring concepts, including organisation, the properties of systems and objects, management, that correspond to notions and structures that are generally applicable in the design and description of distributed systems.

ODP is best known for its use of viewpoints. A *viewpoint* (on a system) is an abstraction that yields a specification of the whole system related to a particular set of concerns. The ODP reference model defines five specific viewpoints as follows [4, 6]:

- The *Enterprise Viewpoint*, which concerns the organisational situation in which business (research activity in the current case) is to take place. In order to better communicate with the ENVRI community, in this document, we rename it as **Science Viewpoint**;
- The *Information Viewpoint*, which concerns modelling of the shared information manipulated within the system of interest;
- The *Computational Viewpoint*, which concerns the design of the analytical, modelling and simulation processes and applications provided by the system;
- The *Engineering Viewpoint*, which tackles the problems of diversity in infrastructure provision; it gives the prescriptions for supporting the necessary abstract computational interactions in a range of different concrete situations;
- The *Technology Viewpoint*, which concerns real-world constraints (such as restrictions on the facilities and technologies available to implement the system) applied to the existing computing platforms on which the computational processes must execute.

This version of the ENVRI Reference Model covers 3 ODP viewpoints: the science, information, and computational viewpoints.



ENVRI Common Operations of Environmental Research Infrastructures

1.4 Approaches

The approach leading to the creation of the ENVRI Reference Model is based on the analysis of the requirements of a collection of representative environmental research infrastructures, which are reported in two ENVRI deliverable:

- D3.1: Assessment of the State of the Art
- D3.3: Analysis of Common Requirements for ENVRI Research Infrastructures

The ODP standard is used as the modelling and specification framework, which enables the designers from different organisations to work independently and collaboratively.

The development starts from a core model and will be incrementally extended based on the community common requirements and interests.

The reference model will be evaluated by examining the feasibilities in implementations, and the refinement of the model will be based on community feedback.

1.5 Conformance

A conforming environmental research infrastructure should support the common subsystems described in section 3 and the functional and information model described in section 4.

The ENVRI Reference Model does not define or require any particular method of implementation of these concepts. It is assumed that implementers will use this reference model as a guide while developing a specific implementation to provide identified services and content. A conforming environmental research infrastructure may provide additional services to users beyond those minimally required computations defined in this document.

Any descriptive (or prescriptive) documents that claim to be conformant to the ENVRI Reference Model should use the terms and concepts defined herein in a similar way.

1.6 Related Work

1.6.1 Related concepts

A **reference model** is an abstract framework for understanding significant relationships among the entities of some environment. It consists of a minimal set of unifying concepts, axioms and relationships within a particular problem domain. [8]

A reference model is not a reference architecture. A **reference architecture** is an architectural design pattern indicating an abstract solution that implements the concepts and relationships identified in the reference model [8]. Different from a reference architecture, a reference model is independent from specific standards, technologies, implementations or other concrete details. A reference model can drive the development of a reference architecture or more than one of them [9].

It could be argued that a reference model is, at its core, an **ontology**. Conventional reference models e.g., OSI[10], RM-ODP [4], OAIS[11], are built upon modelling disciplines. Many recent works, such as the DL.org Digital Library Reference Model [9], are more ontology-like.

Both models and ontologies are technologies for information representation, but have been developed separately in different domains. Modelling approaches have risen to prominence in the software engineering domain over the last ten to fifteen years [12]. Traditionally, software engineers have taken very pragmatic approaches to data representation, encoding only the information needed to solve the problem in hand, usually in the form of language data structures or database tables. Modelling approaches are meant to increase the productivity by maximising compatibility between systems (by reuse of standardised models), simplifying the process of design (by models of recurring design patterns in the application domain), and promoting communication between individuals and teams working on the system (by a standardisation of the terminology and the best practices used in the application domain) [13]. On the other hand, ontologies have been developed by the Artificial Intelligence community since the 1980s. An ontology is a structuring framework for organising information. It renders shared vocabulary and taxonomies which models a domain with the definition of objects and concepts and their properties and relations. These ideas have been heavily drawn upon in the notion of the Semantic Web. [13]

Traditional views tend to distinguish the two technologies. The main points of argument include but are not limited to:

1. Models usually focus on realisation issues (e.g., the Object-Oriented Modelling approach), while ontologies usually focus on capturing abstract domain concepts and their relationship [14].
2. Ontologies are normally used for run-time knowledge exploitation (e.g., for knowledge discovery in a knowledgebase), but models normally do not [15].
3. Ontologies can support reasoning while models cannot (or do not) [13].
4. Finally, models are often based on the Closed World Assumption while ontologies are based on the Open World Assumption [13].

However, these separations between the two technologies are rapidly disappearing in recent developments. Study [13] shows that ‘all ontologies are models’, and ‘almost all models used in modern software engineering qualify as ontologies.’ As evidenced by the growing number of research workshops dealing with the overlap of the two disciplines (e.g., SEKE [16], VORTE [17], MDSW [18], SWESE [19], ONTOSE [20], WoMM [21]), there has been considerable interests in the integration of software engineering and artificial intelligence technologies in both research and practical software engineering projects.[13]

We tend to take this point of view and regard the ENVRI Reference Model as both a model and an ontology. The important consequence is that we can explore further in both directions, e.g., the reference model can be expressed using a modelling language, such as UML (UML4ODP). It can then be built into a tool chain, e.g., to plugin to an integrated development environment such as Eclipse, which makes it possible to reuse many existing UML code and software. On the other hand, the reference model can also be expressed using an ontology language such as RDF or OWL which can then be used in a knowledge base. In this document we explore principally from modelling aspects. In another ENVRI task, T3.4, the ontological aspect of the reference model will be exploited.

Finally, a reference model is a **standard**. Created by ISO in 1970, OSI is probably among the earliest reference models, which defines the well-known 7-layered network communication. As one of the ISO standard types, the reference model normally describes the overall requirements for standardisation and the fundamental principles that apply in implementation. It often serves as a framework for more specific standards [22]. This type of standard has been rapidly adopted, and many reference models

exist today, which can be grouped into 3 categories, based on the type of agreement and the number of people, organisations or countries who were involved in making the agreement:

- **Committee reference model** – a widely-based group of experts nominated by organizations who have an interest in the content and application of the standard build the standard.
- **Consensus reference model** – the principle that the content of the standard is decided by general agreement of as many as possible of the committee members, rather than by majority voting. The ENVRI Reference Model falls into this group.
- **Consultation reference model** – making a draft available for scrutiny and comment to anyone who might be interested in it.

Some examples from each of the categories are discussed below, with emphasis on approaches and technologies.

1.6.2 Related reference models

1.6.2.1 Committee Reference Models

In this category, we look at those defined by international organisations, such as the Advancing Open Standards for the Information Society (OASIS), the Consultative Committee for Space Data Systems (CCSDS), and the Open Geospatial Consortium (OGC).

The Open Archival Information System (OAIS) Reference Model [11] is an international standard created by CCSDS and ISO which provides a framework, including terminology and concepts for archival concept needed for Long-Term digital information preservation and access.

The OASIS Reference Model for Service Oriented Architecture (SOA-RM) [8] defines the essence of service oriented architecture emerging with a vocabulary and a common understanding of SOA. It provides a normative reference that remains relevant to SOA as an abstract model, irrespective of the various and inevitable technology evolutions that will influence SOA deployment.

The OGC Reference Model (ORM) [23], describes the OGC Standards Baseline, and the current state of the work of the OGC. It provides an overview of the results of extensive development by OGC Member Organisations and individuals. Based on RM-ODP's 5 viewpoints, ORM captures business requirements and processes, geospatial information and services, reusable patterns for deployment, and provides a guide for implementations.

The Reference Model for the ORCHESTRA Architecture (RM-OA) [24] is another OGC standard. The goal of the integrated project ORCHESTRA (Open Architecture and Spatial Data Infrastructure for Risk Management) is the design and implementation of an open, service-oriented software architecture to overcome the interoperability problems in the domain of multi-risk management. The development approach of RM-OA is standard-based which is built on the integration of various international standards. Also using RM-ODP standard as the specification framework, RM-OA describes a platform neutral (abstract) model consisting of the informational and functional aspects of service networks combining architectural and service specification defined by ISO, OGC, W3C, and OASIS. [24]

1.6.2.2 Consensus Reference Models

In this category, we discuss those created by non-formal standard organisations.

The LifeWatch Reference Model [25], developed by the EU LifeWatch consortium, is a specialisation of the RM-OA standard which provides the guidelines for the specification and implementation of a biodiversity research infrastructure. Inherited from RM-OA, the reference model uses the ODP standard as the specification framework.

The Digital Library Reference Model [9] developed by DL.org consortium introduces the main notations characterising the whole digital library domain, in particular, it defines 3 different types of systems: (1) Digital Library, (2) Digital Library System, and (3) Digital Library Management System; 7 core concepts characterising the digital library universe: (1) Organisation, (2) Content, (3) Functionality, (4) User, (5) Policy, (6) Quality, and (7) Architecture; and 3 categories of actors: (1) DL End-Users (including, Content Creators, Content Consumers, and Digital Librarians), (2) DL Managers (including, DL Designer, and DL System Administrators) , and (3) DL Software Developers.

The Workflow Reference Model [26] provides a common framework for workflow management systems, identifying their characteristics, terminology and components. The development of the model is based on the analysis of various workflow products in the market. The workflow Reference Model firstly introduces a top level architecture and various interfaces it has which may be used to support interoperability between different system components and integration with other major IT infrastructure components. This maps to the ODP Computational Viewpoint. In the second part, it provides an overview of the workflow application program interface, comments on the necessary protocol support for open interworking and discusses the principles of conformance to the specifications. This maps to the ODP Technology Viewpoint.

The Agent System Reference Model [27] provides a technical recommendation for developing agent systems, which captures the features, functions and data elements in the set of existing agent frameworks. Different from conventional methods, a reverse engineering method has been used to develop the reference model, which starts by identifying or creating an implementation-specific design of the abstracted system; secondly, identifying software modules and grouping them into the concepts and components; and finally, capturing the essence of the abstracted system via concepts and components.

1.6.2.3 Consultation Reference Models

The Data State Reference Model [28] provides an operator interaction framework for visualisation systems. It breaks the visualisation pipeline (from data to view) into 4 data stages (Value, Analytical Abstraction, Visualisation Abstraction, and View), and 3 types of transforming operations (Data Transformation, Visualisation Transformation and Visual Mapping Transformation). Using the data state model, the study [29] analyses 10 existing visualisation techniques including, 1) scientific visualisations, 2) GIS, 3) 2D, 4) multi-dimensional plots, 5) trees, 6) network, 7) web visualisation, 8) text, 9) information landscapes and spaces, and 10) visualisation spread sheets. The analysis results in a taxonomy of existing information visualisation techniques which help to improve the understanding of the design space of visualisation techniques.



ENVRI Common Operations of Environmental Research Infrastructures

The Munich Reference Model [30] is created for adaptive hypermedia applications which is a set of nodes and links that allows one to navigate through the hypermedia structure and that dynamically “adapts” (personalise) various visible aspects of the system to individual user’s needs. The Munich Reference Model uses an object-oriented formalisation and a graphical representation. It is built on top of the Dexter Model layered structure, and extends the functionality of each layer to include the user modelling and adaptation aspects. The model is visually represented using in UML notation and is formally specified in Object Constraint Language (which is part of the UML).

While these works use a similar approach to the development of the reference model as the ENVRI-RM, which is based on the analysis of existing systems and abstracts to obtain the ‘essence’ of those systems, a major difference is that these works have not normally met with significant feedback or been formally approved by an existing community, with the consequence that they express less authority as a standard.

1.6.3 Other Related standards

Data Distribution Service for Real-Time Systems (DDS) [31], an Object Management Group (OMG) standard, is created to enable scalable, real-time, dependable, high performance, interoperable data exchanges between publishers and subscribers. DDS defines a high-level conceptual model as well as a platform-specific model. UML notations are used for specification. While DDS and the ENVRI share many similar views in design and modelling, DDS focuses on only one specific issue, i.e., to model the communication patterns for real-time applications; while ENVRI aims to capture a overall picture of requirements for environmental research infrastructures.

Published by the web standards consortium OASIS in 2010, the Content Management Interoperability Services (CMIS) [32] is an open standard that allows different content management systems to inter-operate over the Internet. Specially, CMIS defines an abstraction layer for controlling diverse document management systems and repositories using web protocols. It defines a domain model plus web services and Restful AtomPub bindings that can be used by applications to work with one or more Content Management repositories/systems. However as many other OASIS standards, CMIS is not a conceptual model and is highly technology dependent [32].

In the next, we introduce the ENVRI Reference Model.



2 MODEL OVERVIEW

2.1 Subsystems of Environmental Infrastructures

In a pre-study, we have investigated a collection of representative environmental research infrastructures. By examining their computational characteristics, we have identified 5 common subsystems: *Data Acquisition*, *Data Curation*, *Data Access*, *Data Processing* and *Community Support*. The fundamental reason of the division of the 5 subsystems is based on the observation that all applications, services and software tools are designed and implemented around 5 major physical resources: the sensor network, the storage, the (internet) communication network, application servers and client devices.

The ENVRI Reference Model is partitioned into the five subsystems. The partitioning of the reference model into subsystems is based broadly on a notion of data life-cycle evident in all existing research infrastructures investigated plus a generic one dedicated to community management.

This lifecycle begins with the acquisition of raw data from a network of integrated data collecting instruments (seismographs, weather stations, robotic buoys, human observations, etc.) which is then pre-processed and curated within a number of data stores belonging to an infrastructure or one of its delegate infrastructures. This data is then made accessible to authorised requests by parties outwith the infrastructure, as well as to services within the infrastructure. This results in a natural partitioning into data acquisition, curation and access. In addition, data can be extracted from parts of the infrastructure and made subject to data processing, the results of which can then be situated against within the infrastructure. Finally, the community support subsystem provides tools and services required to handle data outside of the core infrastructure and reintegrate it when necessary.

Each subsystem should provide a set of capabilities via interfaces invoked by the other subsystems. In ODP, an interface is simply an abstraction of the behaviour of an object that consists of a subset of the interactions expected of that object together with the constraints imposed on their occurrence.

2.1.1 Data Acquisition

The *data acquisition subsystem* of a research infrastructure collects raw data from registered sources to be stored and made accessible within the infrastructure.

The data acquisition subsystem collects raw data from sensor arrays and other instruments, as well as from human observers, and brings those data into the system. Within the ENVRI-RM, the acquisition subsystem is considered to begin upon point of data entry into the modelled system, the general maintenance and deployment of sensor stations and human observers being outside the scope of ENVRI. Acquisition is typically distributed across a network of observatories and stations. Data acquired is generally assumed to be non-reproducible, being associated with a specific (possibly continuous) event in time and place; as such, the assignment of provenance (particularly data source and timestamp) is essential. Real-time data streams may be temporarily stored, sampled, filtered and processed (e.g., based on applied quality control criteria) before being passed on for curation. Control software is often deployed to manage and schedule the execution and monitoring of data flows. Data collected by the acquisition subsystem must ultimately be transferred to the data curation subsystem for preservation, usually within a specific time period.

2.1.2 Data Curation

The *data curation subsystem* of a research infrastructure stores, manages and ensures access to all persistent data-sets produced within the infrastructure.

The data curation subsystem facilitates quality control and preservation of scientific data. The subsystem is typically implemented across one or more dedicated data centres. Data handled by the subsystem include raw data products, metadata and processed data; where possible, processed data should be reproducible by executing the same process on the same source data-sets. Operations such as data quality verification, identification, annotation, cataloguing, replication and archival are often provided. Access to curated data from outside the infrastructure is brokered through the data access subsystem. There is usually an emphasis on non-functional requirements for data curation satisfying availability, reliability, utility, throughput, responsiveness, security and scalability criteria.

2.1.3 Data Access

The *data access subsystem* of a research infrastructure enables discovery and retrieval of scientific data subject to authorisation.

The data access subsystem enables discovery and retrieval of data housed in data resources managed by the data curation subsystem. Data access subsystems often provide gateways for presenting or delivering data products. Query and search tools may be provided which allow users or upstream services to discover data based on metadata or semantic linkages. Data handled by the access subsystem need not be homogeneous. When supporting heterogeneous data, different types of data (often pulled from a variety of distributed data resources) may be converted into uniform representations with uniform semantics resolved by a data discovery service. The subsystem may provide services for harvesting, compressing and packaging (meta)data as well as encoding services for secure data transfer. Data access is controlled using authentication and authorisation policies. Despite the name, the access subsystem may also provide services for importing data into the infrastructure.

2.1.4 Data Processing

The *data processing subsystem* of a research infrastructure provides a toolbox of services for performing a variety of data processing tasks.

The data processing subsystem is able to aggregate data from various sources and conduct a range of experiments and analyses upon that data. Data handled by the subsystem are typically derived and recombined via the data access subsystem. The data processing subsystem is expected to offer operations for statistical analysis and data mining as well as facilities for conducting scientific experiments, modelling and simulation, and scientific visualisation. Performance requirements for processing scientific data tend to be concerned with scalability which may be addressable at the level of engineering (e.g., by making use of Grid or Cloud services).

2.1.5 Community Support

The *community support subsystem* of a research infrastructure exists to support users of an infrastructure in their interactions with that infrastructure.



ENVRI Common Operations of Environmental Research Infrastructures

The community support subsystem manages, controls and tracks users' activities and supports users to conduct their roles in their communities. Data 'handled' within the subsystem are typically user-generated data and communications. The community support subsystem may support interactive visualisation, standardised authentication, authorisation and accounting protocols, and the use of virtual organisations. The subsystem is considered to encircle the other four subsystems, describing the interface between the research infrastructure and the wider world in which it exists.

2.2 Subsystem Relationships

As shown in Figure 2.1, amongst the five subsystems can be identified seven major points-of-reference wherein interfaces between subsystems can be implemented.

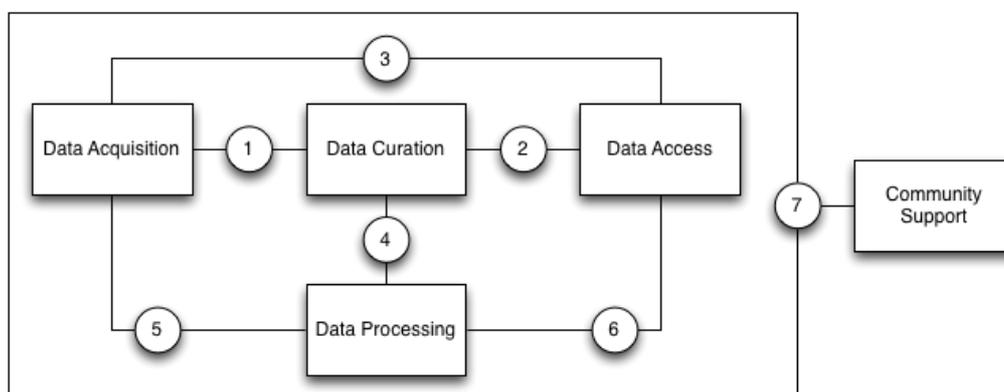


Figure 2.1: Illustration of the major points-of-reference between different subsystems of the ENVRI-RM.

These points-of-reference are as follows:

1. **Acquisition/Curation** by which the collection of raw data is managed.
2. **Curation/Access** by which the retrieval of curated data products is arranged.
3. **Acquisition/Access** by which the status of the data acquisition network can be monitored externally.
4. **Curation/Processing** by which analyses of curated data is coordinated.
5. **Acquisition/Processing** by which acquisition events are listened for and responded to.
6. **Processing/Access** by which data processes are scheduled and reported.
7. **Community/All** by which the outside world interacts with the infrastructure in many different roles.

Depending on the distribution of resources in an implemented infrastructure, some of these reference points may not be present in the infrastructure. They take particular importance however when considering scenarios where a research infrastructure delegates subsystems to other client infrastructures. For example, EPOS and LifeWatch both delegate data acquisition and some data curation activities to client national or domain-specific infrastructures, but provide data processing services over the data held by those client infrastructures. Thus reference points 4 and 5 become of great importance to the construction of those projects.



2.3 Common Functions within Common Subsystems

Analysis of the common requirements of the six ESFRI environmental infrastructures affiliated with the ENVRI project has resulted in the identification of a number of common functionalities. These functionalities can be partitioned amongst the five subsystems of the ENVRI-RM and presented as interfaces of each subsystem. They encompass a range of concerns, from the fundamental (e.g. data collection and storage, data discovery and access and data security) to more specific challenges (e.g. data versioning, instrument monitoring and interactive visualisation).

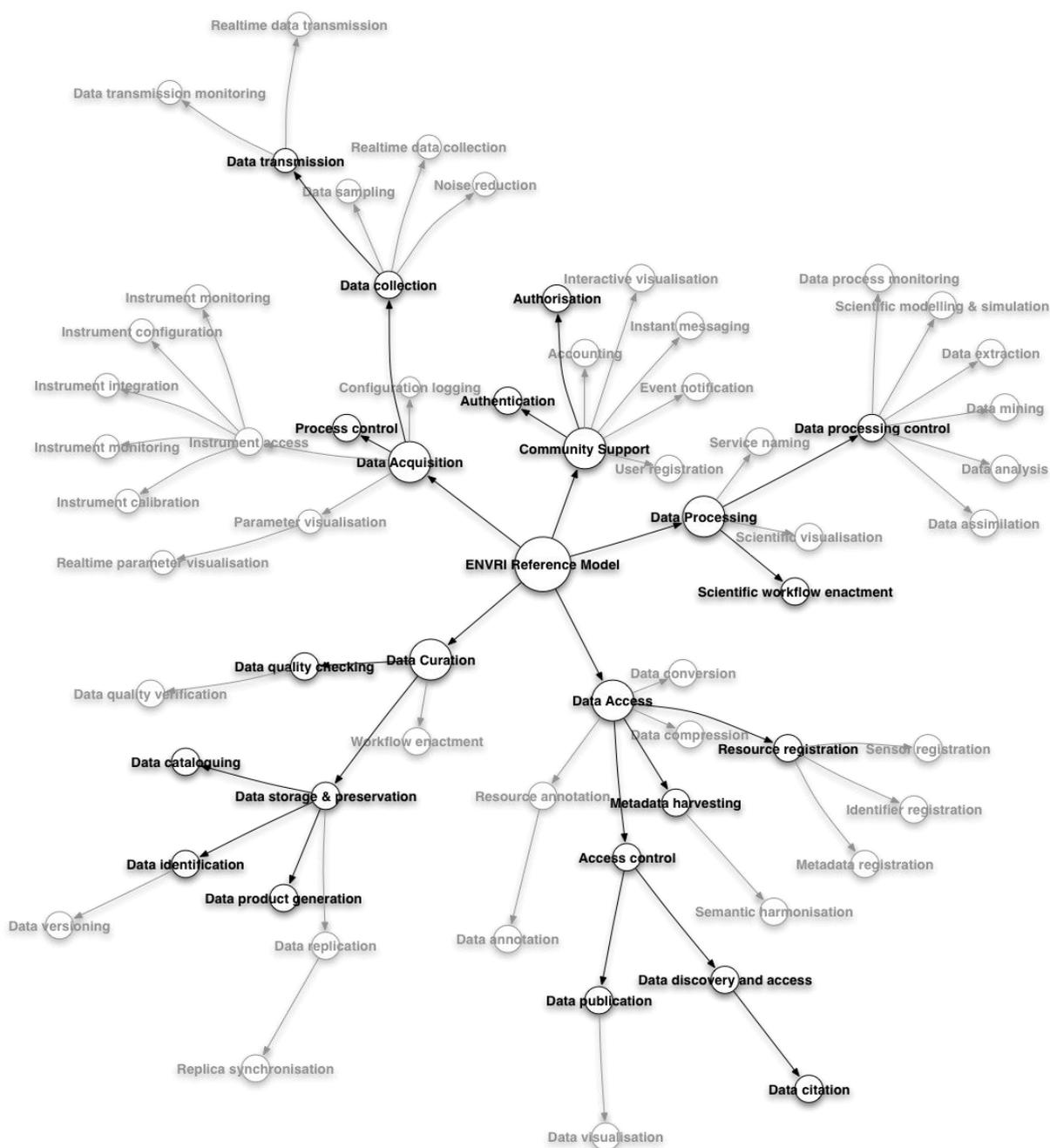


Figure 2.2: Radial depiction of ENVRI-RM requirements with the minimal model highlighted.

In order to better manage the range of requirements, and in order to ensure rapid publication of incremental refinements to the ENVRI-RM, as highlighted in Figure 2.2, a *minimal model* has been identified which describes the fundamental functionality necessary to describe a functional environmental research infrastructure. By initially focusing on this minimal model, it then becomes practical to produce a partial specification of the ENVRI-RM which nonetheless reflects the final shape of the ENVRI-RM without the need for significant refactoring. Further development of the ENVRI-RM will focus on designated priority areas based on feedback from the contributing ESFRI representatives.

The definitions of the minimal set of functions are given as follows. The definition of the full list of common functions are provided in Appendix A.

(A) Data Acquisition Subsystem

Process Control: A functionality that receives input status, applies a set of logic statements or control algorithms, and generates a set of analogue / digital outputs to change the logic states of devices.

Data Collection: A functionality that obtains digital values from a sensor instrument, associating consistent timestamps and necessary metadata.

Data Transmission: A functionality that transfers data over a communication channel using specified network protocols.

(B) Data Curation Subsystem

Data Quality Checking: A functionality that detects and corrects (or remove) corrupt, inconsistent or inaccurate records from datasets.

Data Identification: A functionality that assigns (global) unique identifiers to data contents.

Data Cataloguing: A functionality that associates a data object with one or more metadata objects which contain data descriptions.

Data Product Generation: A functionality that processes data against requirement specifications and standardised formats and descriptions.

Data Storage & Preservation: A functionality that deposits (over the long-term) data and metadata or other supplementary data and methods according to specified policies, and then to make them accessible on request.

(C) Data Access Subsystem

Access Control: A functionality that approves or disapproves of access requests based on specified access policies.

Metadata Harvesting: A functionality that (regularly) collects metadata in agreed formats from different sources.

Resource Registration: A functionality that creates an entry in a resource registry and inserts a resource object or a reference to a resource object with specified representation and semantics.

Data Publication: A functionality that provides clean, well-annotated, anonymity-preserving datasets in a suitable format, and by following specified data-publication and sharing policies to make the datasets publically accessible or to those who agree to certain conditions of use, and to individuals who meet certain professional criteria.

Data Citation: A functionality that assigns an accurate, consistent and standardised reference to a data object, which can be cited in scientific publications.

Data Discovery and Access: A functionality that retrieves requested data from a data resource by using suitable search technology.

(D). Data Processing Subsystem

Data Assimilation: A functionality that combines observational data with output from a numerical model to produce an optimal estimate of the evolving state of the system.

Data Analysis: A functionality that inspects, cleans, transforms data, and to provide data models with the goal of highlighting useful information, suggesting conclusions, and supporting decision making.

Data Mining: A functionality that supports the discovery of patterns in large datasets.

Data Extraction: A functionality that retrieves data out of (unstructured) data sources, including web pages, emails, documents, PDFs, scanned text, mainframe reports, and spool files.

Scientific Modelling and Simulation: A functionality that supports of the generation of abstract, conceptual, graphical or mathematical models, and to run an instance of the model.

(Scientific) Workflow Enactment: A specialisation of Workflow Enactment, which support of composition and execution a series of computational or data manipulation steps, or a workflow, in a scientific application. Important processes should be recorded for provenance purposes.

Data Processing Control: A functionality that initiates the calculation and manage the outputs to be returned to the client.

(E) Community Support Subsystem

Authentication: A functionality that verifies the credentials of a user.

Authorisation: A functionality that specifies access rights to resources.



3 ENVRI REFERENCE MODEL

The ENVRI Reference Model is structured according to the Open Distributed Processing (ODP) standard. As such, the Reference Model is defined from five different perspectives. In the context of ENVRI, which uses ODP to define an 'archetypical' environmental research infrastructure rather than a specific (implemented) infrastructure, three viewpoints take particular priority – the *Science*, *Information* and *Computational* viewpoints.

The remaining two viewpoints (*Engineering* and *Technology*) are more relevant to specific instances of research infrastructure. Nevertheless, the ENVRI Reference Model will address these viewpoints to some extent in future revisions.

3.1 Science Viewpoint

The Science Viewpoint of the ENVRI-RM intends to capture the requirements for an environmental research infrastructure from the perspective of the people who perform their tasks and achieve their goals as mediated by the infrastructure. Modelling in this viewpoint uses a reverse engineering method, which derives the principles and properties of model objects through the analysis of the structure and functionality of the real-world systems.

In a pre-study, we have observed 5 subsystems commonly exist in environmental science research infrastructures: *Data Acquisition*, *Data Curation*, *Data Access*, *Data Processing* and *Community Support*. Correspondingly, human activities which interact with the 5 subsystems in order to collaboratively conduct scientific research, from data collection to the delivery of scientific results, can also be grouped in the same way. Such groups are the so-called *communities* in ODP. In this viewpoint, we examine what those communities are, what kind of roles they have, and what main behaviours they act out.

3.1.1 Common Communities

A **community** is a collaboration which consists of a set of *roles* agreeing their objective to achieve a stated business purpose.

In the ENVRI-RM, we distinguish 5 activities, seen as communities in accordance to the 5 common sub-systems. As shown in Figure 3.1, the 5 communities are, *data acquisition*, *data curation*, *data publication*, *data service provision*, and *data usage* community. The definition of the communities are based on their objectives.

- **Data Acquisition Community**, who collects raw data and brings (streams of) measurements into a system;
- **Data Curation Community**, who curates the scientific data, maintains and archives them, and produces various data products with metadata;
- **Data Publication Community**, who assists data publication, discovery and access;
- **Data Service Provision Community**, who provides various services, applications and software/tools to link and recombine data and information in order to derive knowledge;
- **Data Usage Community**, who make use of data and service products, and transfer knowledge into understanding.

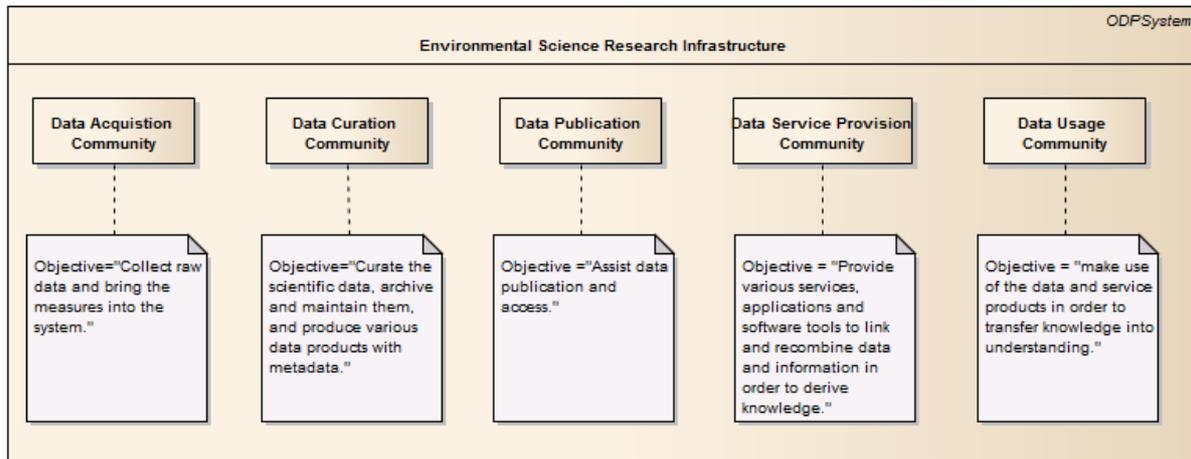


Figure 3.1: Common Communities

3.1.2 Common Community Roles

A **role** in a community is a prescribing behaviour that can be performed any number of times concurrently or successively. A role can be either *active* (typically associated with a human actor) or *passive* (typically associated with a non-human actor).

In the following, we identify *active roles* in relation to people associated with a research infrastructure:

- those who use the research infrastructure to do science;
- those who work on resources to build, maintain and operate the research infrastructure; and
- those who govern, manage and administer the research infrastructure

Note, an individual may be a member of more than one community.

A system (or part of it) and the hardware facilities which *active roles* interact with are modelled as *passive roles*.

3.1.2.1 Roles in the Data Acquisition Community

The main objectives of the data acquisition community is to bring measurements into the system. The measurement and monitoring models are designed by *model designer* based on the requirements of *environmental scientists*. Such a design decides what data is to be collected and what metadata is to be associated with it, such as experimental information and instrument conditions. *Technicians* configure and calibrate a *sensor* or a *sensor network* to satisfy the experiment specifications. In the case where human sensors are to be used, *observers* or *measurers* input the measures to the system, e.g., by using mobile devices. *Data collectors* interact with a *data acquisition system* to prepare the data or control the flow of data and automatically collect and transmit the data.

As shown in Figure 3.2, the following roles are identified in the data acquisition community:

- **Environmental Scientist:** An active role, which is a person who conducts research or performs investigation for the purpose of identifying, abating, or eliminating sources of pollutants or hazards that affect either the environment or the health of the population. Using knowledge of

various scientific disciplines, they may collect, synthesize, study, report, and recommend action based on data derived from measurements or observations of air, food, soil, water, and other sources.

- **(Measurement Model) Designer:** An active role, which is a person who designs the measurements and monitoring models based on the requirements of environmental scientists.
- **Sensor:** A passive role, which is a converter that measures a physical quantity and converts it into a signal which can be read by an observer or by an (electronic) instrument.
- **Sensor network:** A passive role, which is a network consisting of distributed autonomous sensors to monitor physical or environmental conditions.
- **Technician:** An active role, which is a person who develops and deploys sensor instruments, establishing and testing the sensor network, operating, maintaining, monitoring and repairing the observatory hardware.
- **Measurer:** An active role, which is a person who determines the ratio of a physical quantity, such as a length, time, temperature etc., to a unit of measurement, such as the meter, second or degree Celsius.
- **Observer:** An active role, which is a person who receives knowledge of the outside world through his/her senses, or records data using scientific instruments.
- **Data collector:** An active role, which is a person who prepares and collects data. The purpose of data collection is to obtain information to keep on record, to make decisions about important issues, or to pass information on to others.
- **Data Acquisition Subsystem:** In Science Viewpoint, data acquisition subsystem represents a passive role of the data acquisition community. As defined in Section 2 Model Overview, it is research infrastructure which provides functionalities to automate the process of data acquisition.

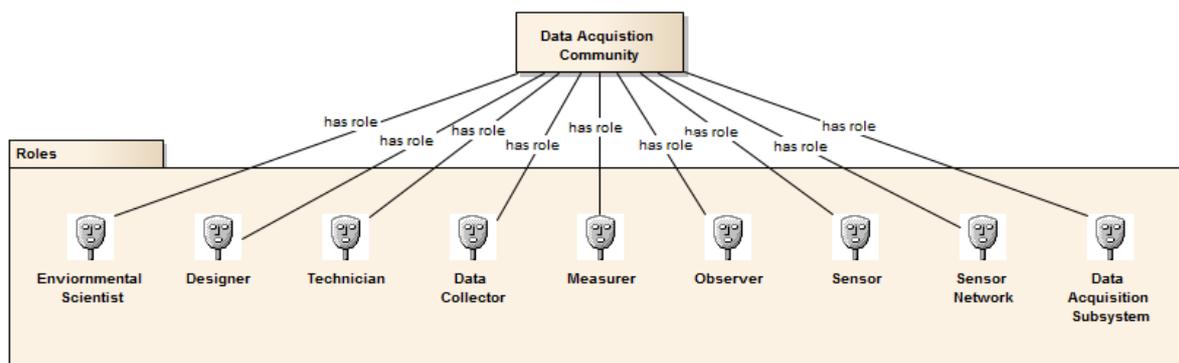


Figure 3.2: Roles in the Data Acquisition Community

3.1.2.2 Roles in the Data Curation Community

The data curation community responds to provide quality data products and maintain the data resources. Consider a typical data curation scenario: when data is being imported into a curation subsystem, a curator will perform the quality checking of the scientific data. Unique identifiers will be assigned to the qualified data, which will then be properly catalogued by associating necessary metadata, and stored or archived. The main human roles interacting with or maintaining a data curation subsystem are data curators who manage the data and storage administrators who manage the storage facilities.

As shown in Figure 3.3, we identified the following roles in this community:

- **Data Curator:** An active role, which is a person who verifies the quality of the data, preserves and maintain the data as a resource, and prepares various required data products.
- **Data Curation System:** In Science Viewpoint, data curation subsystem represents a passive role of the data curation community. The definition is given in Section 2 Model Overview, which a research infrastructure stores, manages and ensures access to all persistent data-sets produced within the infrastructure.
- **Storage Administrator:** An active role, which is a person who has the responsibilities to design data storage, tune queries, perform backup and recovery operations, set up RAID mirrored arrays, and make sure drive space is available for the network.
- **Storage:** A passive role, which includes memory, components, devices and media that retain digital computer data used for computing for some interval of time.

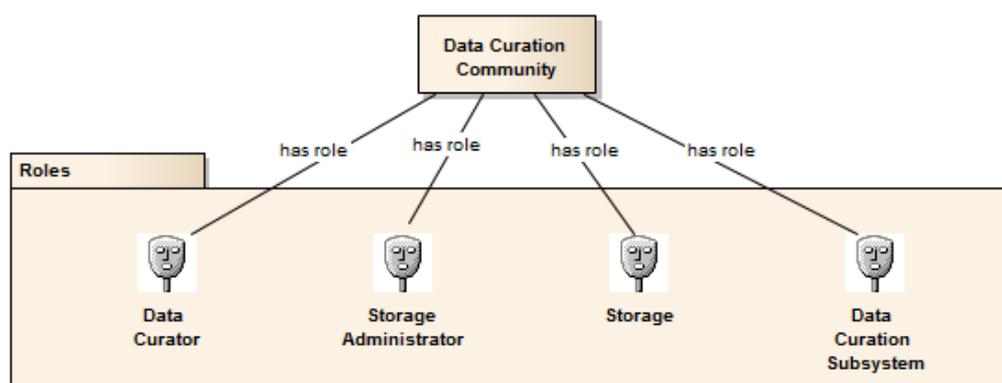


Figure 3.3: Roles in the Data Curation Community

3.1.2.3 Roles in the Data Publication Community

The objectives of the data publication community are to publish data and assist discovery and access. We consider the scenarios described by Kahn's data publication model [34]: an *originator*, i.e., a user with digital material to be made available for public access, makes the material into a digital object. A digital object is a data structure whose principal components are digital material, or data, plus a unique identifier for this material, called a *handle* (and, perhaps, other material). To get a *handle*, the user requests one from an authorized *handle generator*. A user may then deposit the digital object in one or more *repositories*, from which it may be made available to others (subject, to the particular item's terms and conditions, etc.). Upon depositing a digital object in a repository, its *handle* and the *repository* name or IP address is registered with a globally available system of *handle servers*. Users may subsequently present a *handle* to a *handle server* to learn the network names or addresses of repositories in which the corresponding digital object is stored. We use a more general term "PID" instead of "*handle*" (thus, "*PID registry*" instead of "*handle servers*"), and identify the key roles involved in the data publication process including, a data originator, a PID generator, a repository, and a PID registry.

The published data are to be discovered and accessed by data consumers. A semantic mediator is used to facilitate the heterogeneous data discovery.

In summary, as shown in Figure 3.4, the following roles are involved in the data publication community:

- **Data Originator:** Either an active or a passive role, which provides the digital material to be made available for public access.
- **PID Generator:** A passive role, a system which assigns persist global unique identifiers to a (set of) digital object.
- **PID Registry:** A passive role, which is an information system for registering PIDs.
- **(Data Publication) Repository:** A passive role, which is a facility for the deposition of published data.
- **Semantic Mediator:** A passive role, which is a system or middleware facilitating semantic mapping discovery and integration of heterogeneous data.
- **Data Access Subsystem:** In Science Viewpoint, data access subsystem represents a passive role of the data publication community. The definition is given in Section 2 Model Overview, which is a research infrastructure enables discovery and retrieval of scientific data subject to authorisation. Data access may include authorization or may be open.
- **Data Consumer:** Either an active or a passive role, which is an entity who receives and uses the data.

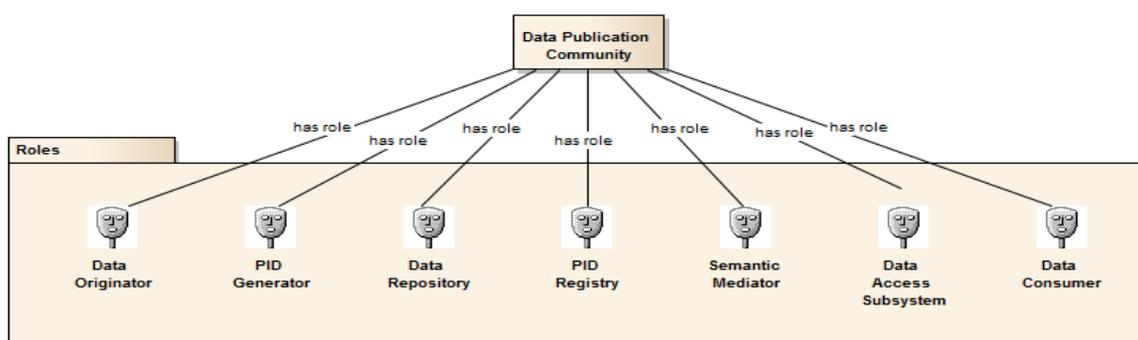


Figure 3.4: Roles in the Data Publication Community

3.1.2.4 Roles in the Data Service Provision Community

The data service provision community provides various application services such as data analysis, mining, simulation and modelling, visualisation, and experimental software tools, in order to facilitate the use of the data. We consider scenarios of service oriented computing paradigm which is adopted by the ENVRI implementation model, and identify the key roles as below. These concepts are along the lines of the existing standards such as OASIS Reference Model for Service Oriented Architecture.

As shown in Figure 3.5, roles in the data service provision community include:

- **Data Provider:** Either an active or a passive role, which is an entity providing the data to be used.
- **Service Provider:** Either an active or a passive role, which is an entity providing the services to be used.
- **Service Registry:** A passive role, which is an information system for registering services.
- **Capacity Manager:** An active role, which is a person who manages and ensures that the IT capacity meets current and future business requirements in a cost-effective manner.
- **Service Consumer:** Either an active or a passive role, which is an entity using the services provided.

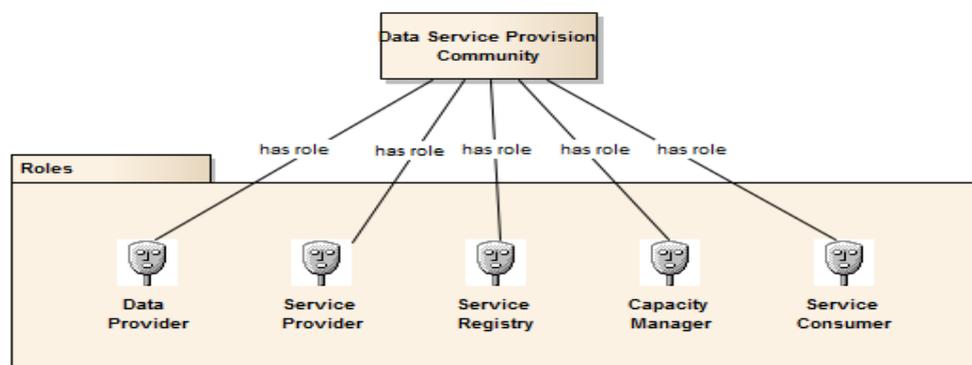


Figure 3.5: Roles in the Data Service Provision Community

3.1.2.5 Roles in the Data Usage Community

The main role in the data usage community is a *user* who is the ultimate consumer of data, applications and services. Depending on the purposes of usage, a user can be one of the following active roles:

- **Scientist or Researcher:** An active role, which is a person who makes use of the data and application services to conduct scientific research.
- **Technologist or Engineer:** An active role, which is a person who develops and maintains the research infrastructure.
- **Education or Trainee:** An active role, which is a person who makes use of the data and application services for education and training purposes.
- **Policy or Decision Maker:** An active role, which is a person who makes decisions based on the data evidences.
- **Private Sector (Industry investor or consultant):** An active role, which is a person who makes use of the data and application service for predicting markets so as to make business decisions on producing related commercial products.
- **General Public, Media or Citizen (Scientist):** An active role, which is a person who is interested in understanding the knowledge delivered by an environmental science research infrastructure, or discovering and exploring the knowledgebase enabled by the research infrastructure.

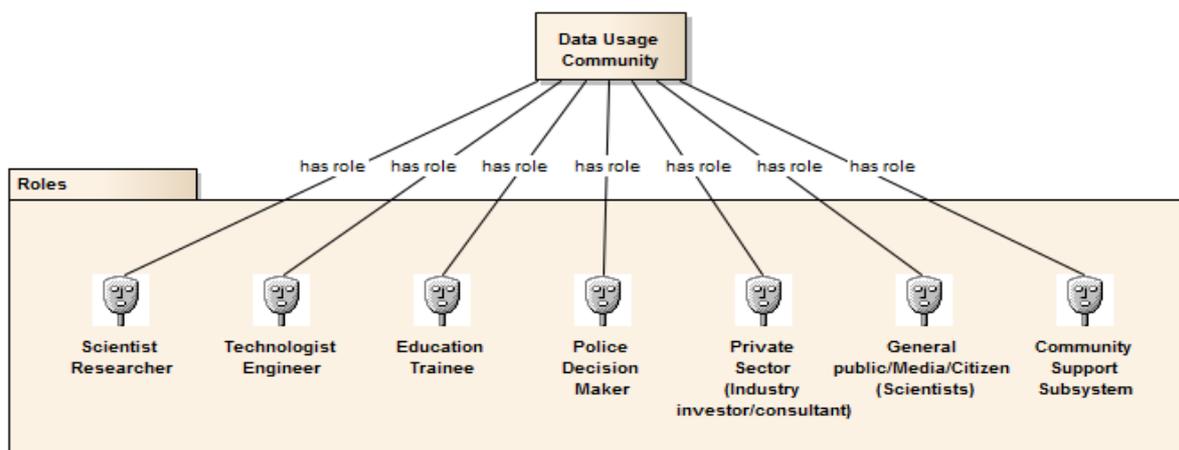


Figure 3.6: Roles in the Data Usage Community

3.1.3 Common Community Behaviours

A **behaviour** of a community is a composition of actions performed by roles normally addressing separate business requirements. In the ENVRI-RM, the modelling of community behaviours is based on analysis of the common requirements of the ENVRI research infrastructure which has resulted in a list of common functions. The initial model focuses on the minimal set of requirements. A community behaviour can be either a single function or a composition of several functions from the function list.

3.1.3.1 Behaviours of the Data Acquisition Community

Figure 3.7 depicts the main behaviours of the data acquisition community including:

- **Design of Measurement Model:** A behaviour performed by a *Measurement Model Designer* that designs the measurement or monitoring model based on scientific requirements.
- **Instrument Configuration:** A behaviour performed by a *Technician* that sets up a *sensor* or a *sensor network*.
- **Instrument Calibration:** A behaviour performed by a *Technician* that controls and records the process of aligning or testing a *sensor* against dependable standards or specified verification processes.
- **Data Collection:** A behaviour performed by a *Data Collector* that obtains digital values from a *sensor* instrument (or a human sensor such as a *Measurer* or an *Observer*), associating consistent timestamps and necessary metadata

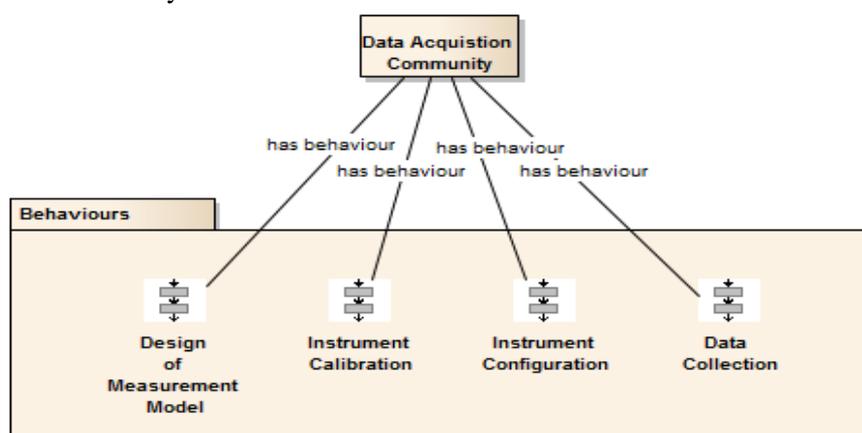


Figure 3.7: Behaviours of the Data Acquisition Community

3.1.3.2 Behaviours of the Data Curation Community

The main behaviours of the data curation community are depicted in Figure 3.8 which include:

- **Data Quality Checking:** A behaviour performed by a *Data Curator* that detects and corrects (or removes) corrupt, inconsistent or inaccurate records from data sets.
- **Data Preservation:** A behaviour performed by a *Data Curator* that deposits (over long-term) the data and metadata or other supplementary data and methods according to specified policies, and makes them accessible on request.
- **Data Product Generation:** A behaviour performed by a *Data Curator* that processes data against requirement specifications and standardised formats and descriptions.
- **Data Replication:** A behaviour performed by a *Storage Administrator* that creates, deletes and maintains the consistency of copies of a data set on multiple storage devices.

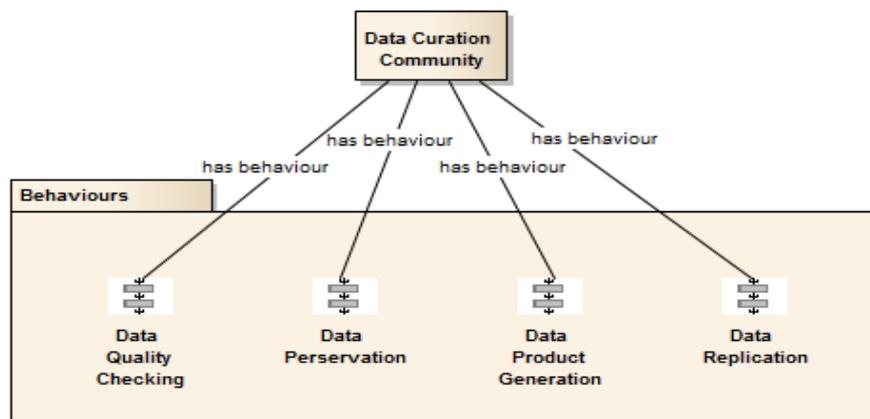


Figure 3.8: Behaviours of the Data Curation Community

3.1.3.3 Behaviours of the Data Publication Community

As shown in Figure 3.9, a data publication community may perform the following behaviours:

- **Data Publication:** A behaviour that provides clean, well-annotated, anonymity-preserving datasets in a suitable format, and by following specified data-publication and sharing policies, to make the datasets accessible publicly or to those who agree to certain conditions of use, and to individuals who meet certain professional criteria.
- **Semantic Harmonisation:** A behaviour enabled by a *Semantic Mediator* that unifies similar data (knowledge) models based on the consensus of collaborative domain experts to achieve better data (knowledge) reuse and semantic interoperability.
- **Data Discovery and Access:** A behaviour enabled by a *Data Discovery and Access system* that retrieves requested data from a data resource by using suitable search technology.
- **Data Citation:** A behaviour performed by a *Data Consumer* that assigns an accurate, consistent and standardised reference to a data object, which can be cited in scientific publications.

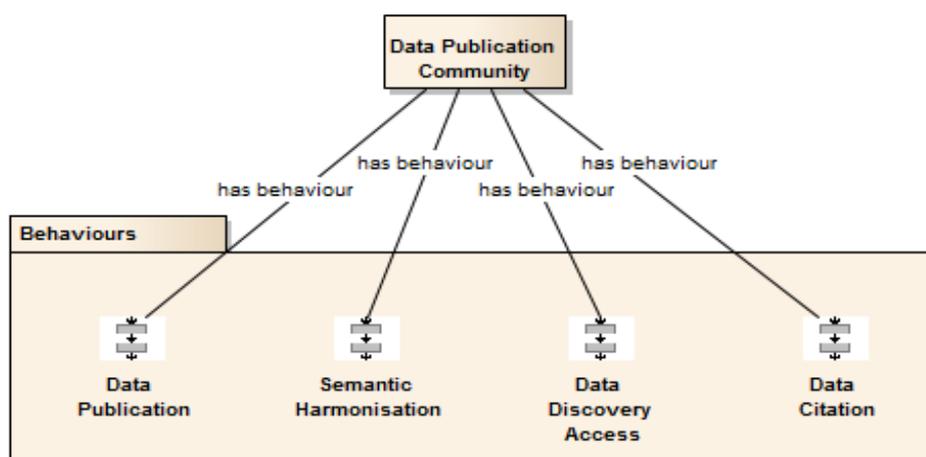


Figure 3.9: Behaviours of the Data Publication Community



3.1.3.4 Behaviours of the Data Service Provision Community

Figure 3.10 depicts the behaviours modelled for the data service provision community, which include:

- **Service Description:** A behaviour performed by a *Service Provider* to provide the information needed in order to use a service [8].
- **Service Registration:** A behaviour performed by a *Service Provider* to make the service visible to *Service Consumers* by registering it in a service registry [8].
- **Service Coordination:** A behaviour performed by a *Service Provider* to coordinate the actions of distributed applications in order to reach consistent agreement on the outcome of distributed transactions.
- **Service Composition:** A behaviour performed by a *Service Provider* to combine multiple services which can be achieved by either *Choreography* or *Orchestration*. **Service Choreography** is a collaboration between *Service Providers* and *Service Consumers*. **Service Orchestration** is the behaviour that a *Service Provider* performs internally to realise a service that it provides [35].

These are general behaviours of a service-oriented computing model. In the context of environmental science research infrastructures, a data service provision community will focus on the implementation of domain special services, in particular those supporting **Data Assimilation, Data Analysis, Data Mining, Data Extraction, Scientific Modelling and Simulation, (Scientific) Workflow Enactment**. (See Chapter 2, Terminology and Glossary, for the definitions of these functionalities.)

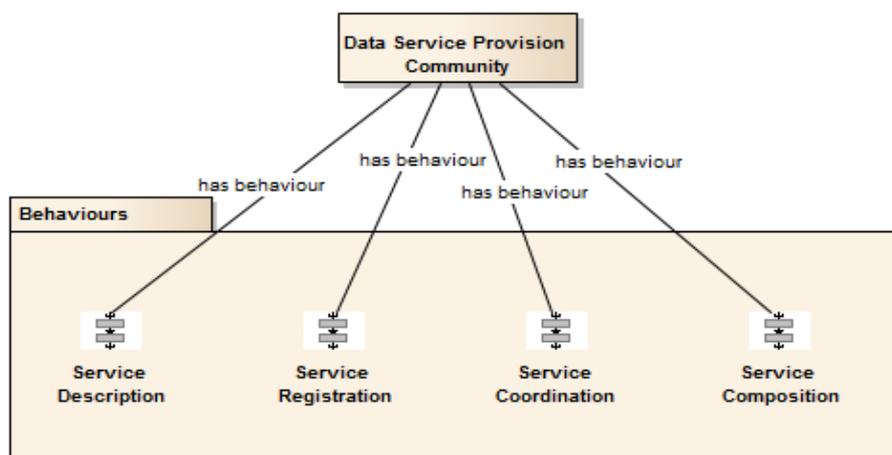


Figure 3.10: Behaviours of the Data Service Provision Community

3.1.3.5 Behaviours of the Data Usage Community

Finally, a data usage community may have the following behaviours, which is depicted in Figure 3.11:

- **User Behaviour Tracking:** A behaviour enabled by a *Community Support System* to track the *Users*. If the research infrastructure has identity management, authorisation mechanisms, accounting mechanisms, for example, a Data Access Sub-System is provided, then the Community Support System either include these or work well with them.
- **User Profile Management:** A behaviour enabled by a *Community Support System* to support persistent and mobile profiles, where profiles will include preferred interaction settings, preferred computational resource settings, and so on.



ENVRI Common Operations of Environmental Research Infrastructures

- **User Working Space Management:** A behaviour enabled by a *Community Support System* to support work spaces that allow data, document and code continuity between connection sessions and accessible from multiple sites or mobile smart devices.
- **User Working Relationships Management:** A behaviour enabled by a *Community Support System* to support a record of working relationships, (virtual) group memberships and friends.
- **User Group Work Supporting:** A behaviour enabled by a *Community Support System* to support controlled sharing, collaborative work and publication of results, with persistent and externally citable PIDs.

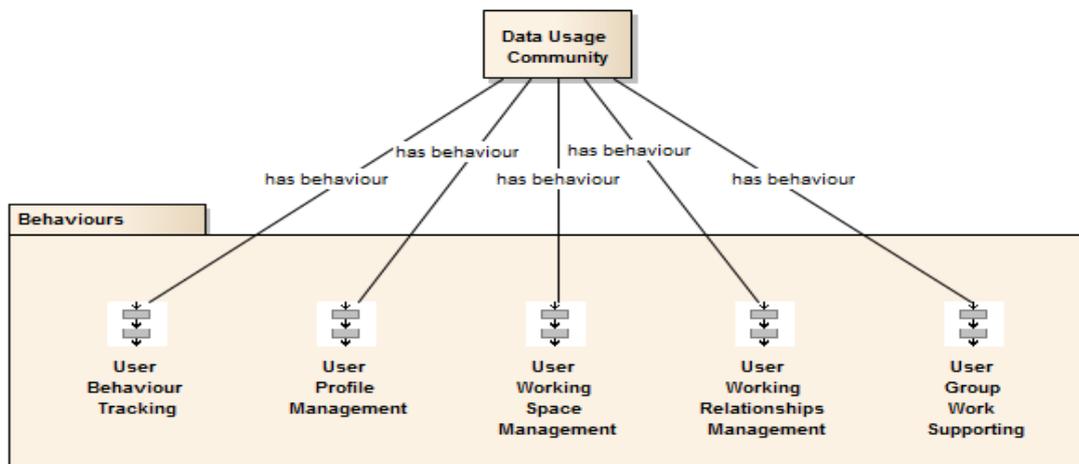


Figure 3.11: Behaviours of the Data Usage Community

3.2 Information Viewpoint

The goal of the information viewpoint is to provide a common abstract model for the shared information handled by the infrastructure. The focus lies on the information itself, without considering any platform-specific or implementation details. It is independent from the computational interfaces and functions that manipulate the information or the nature of technology used to store it. Similar to a high level ontology, it aims to provide a unique and consistent interpretation of the shared information entities of a particular domain. The information viewpoint specifies the types of the information objects and the relationships between those types. It describes how the state of the data evolves as the results of computational operations. It also defines the constraints on the data and the rules governing the processing of such data.

In the information viewpoint, we discuss the following aspects:

- **Components**, which defines a collections of information objects and action types necessary to support the minimal set of required functionalities (See Section 2 Model Overview);
- **Dynamic Schemata**, which specifies how the information objects evolve as the system operates, describing the allowable state changes as the effects of the actions;
- **Static Schemata**, which describes instantaneous views of the information objects at a certain stage of the data lifecycle; and
- **Subsystems**, which regroups the defined information objects into the 5 common Subsystems (as defined in Section 2 Model Overview) for the purpose of easy observation.



ENVRI Common Operations of Environmental Research Infrastructures

3.2.1 Components

The ENVRI information specification defines a configuration of information objects, the behaviour of those objects, the actions that can happen and a set of constraints that should always hold for this collection of elements. The model elements are organised into four groups:

- Information Objects, which defines a collection of information objects manipulated by the system;
- Information Action Types, which defines events that cause state changes of information objects;
- Information Object Instances, which specifies realised variations of defined information objects;
- Data States, which defines data states and their changes as effects of actions.

3.2.1.1 Information Objects

Information objects are used to model various information entities manipulated by the system. In ENVRI, information objects are defined to capture 3 types of information:

- The meta information of data collections, typically those related to the design of observation and measurement models, including:
 - the design specification of the observation and measurement;
 - the description of the measurement procedure;
- The data or information processed by the system, mainly linked with the persisted data, including:
 - the scientific data;
 - the Quality Assurance (QA) annotations;
 - the metadata or concepts from a conceptual model e.g., an ontology;
 - the unique identifiers for the data identification;
 - the various data states as the effects of actions;
- The information used for the management of data, including:
 - the backup (of data)
 - the mapping rules which are used for the model-to-model transformations; and
 - the data provenance which are used to record the state changes of data in their lifecycles

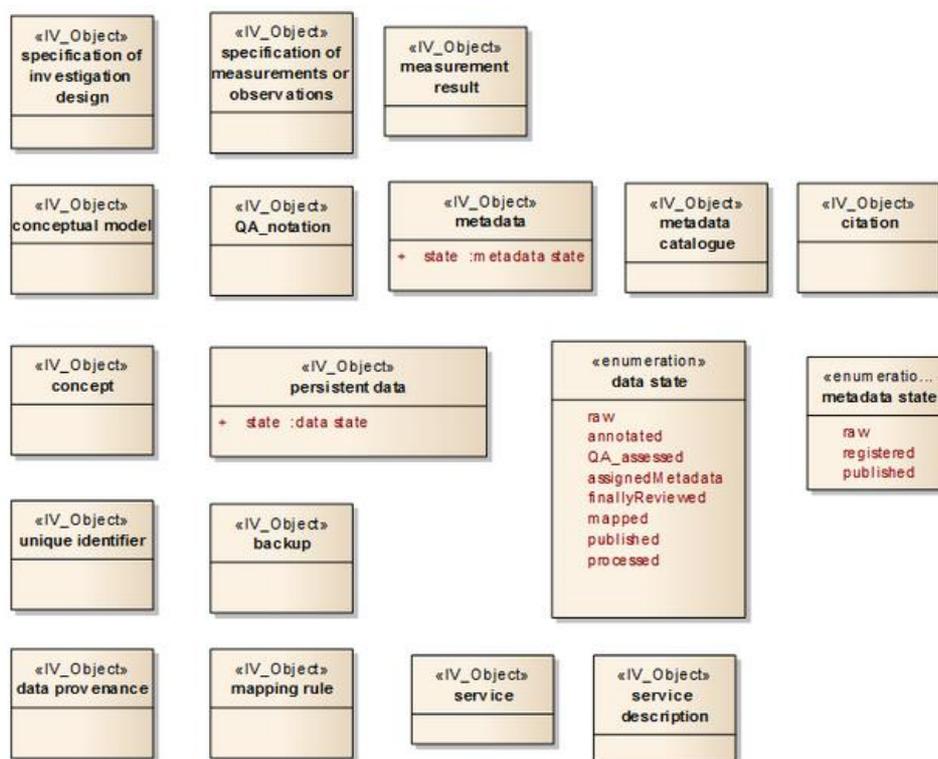


Figure 3.12: Information Objects

The definitions of the information objects given in Figure 3.12 are given as follows:

specification of investigation design

This is the background information 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 information for the allowed evaluations of data. (E.g. the question whether a sampling design was done randomly or by strategy determines which statistical methods that can be applied or not).

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;
- by whom it is measured/observed; and
- what the temporal design is (single /multiple measurements / interval of measurement etc.)



ENVRI Common Operations of Environmental Research Infrastructures

Note, this specification can be included as metadata or the semantic annotations 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.

measurement result

Quantitative determinations of magnitude, dimension and uncertainty to the outputs of observation instruments, sensors (including human observers) and sensor networks.

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 practices, this can be:

- simple flags like "valid" / "invalid" up to comprehensive descriptions like
- "data set to invalid by xxxxxx on ddmmy 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.

unique identifier (UID)

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

There are 3 main generation strategy:

- 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 sub-community

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.



ENVRI Common Operations of Environmental Research Infrastructures

metadata

Data about data, in scientific applications is used to describe, explain, locate, or make it easier to retrieve, use, or manage an information 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 syntax used in scientific applications include Some popular syntax includes:

- 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/
- ISO 11179: metadata-stds.org/11179/
- FGDC (The Federal Geographic Data Committee): www.fgdc.gov/standards
- DDI (Data Documentation Initiative): www.ddialliance.org/
- INSPIRE: <http://inspire.jrc.ec.europa.eu/>
- 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 as data objects or information resources may move to different phases during their life in a digital environment, thus requiring layers of metadata that can be associated.

Metadata can be fused with the data. However, in many applications, such as a provenance system which tracks the environmental experience workflows, or a distributed satellite image annotation system the metadata and data, metadata and data are often created and stored separately, e.g., they may be generated by different users, in different computing processes, stored at different locations, in different types of storage. Metadata and data may have different lifecycles, e.g., they might have been created at a different time, updated and removed independently. 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. Without efficient software and tools, the management of the linkage between metadata and data becomes onerous. 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 state

- **raw:** are established metadata, which are not yet registered. In general, they are not shareable in this status
- **registered:** are metadata which are inserted into a metadata catalogue.
- **published:** are metadata made available to the public, the outside world. Within some metadata catalogues registered.

metadata catalogue

A collection of metadata, usually established to make the metadata available to a community. A metadata catalogue has an access service.

citation

Citation in the sense of IT is a pointer from published data to:

- the data source(s)
- and / or 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

It is important that the citation is resolvable, which means that at least the meaning of the items above are clear.



concept

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. Concepts within ontologies and their relations can be seen as machine readable sentences. Those sentences can be used to establish a self-description. It is, however, practice today, to have both, the human readable description and the machine readable description. In this sense a conceptual model can also be seen as a collection of human and machine readable sentences. Conceptual models can reside within the persistence layer of a data provider or a community or outside. Conceptual models can be fused with the data (e.g. within a network of triple stores) or kept separately.

data state

Term used as defined in ISO/IEC 10746-2. At a given instant in time, data state is the condition of an object that determines the set of all sequences of actions (or traces) in which the object can participate.

The data states and their changes as effects of actions are specified in subsection 3.2.1.4 Data States. In their lifecycle data may have certain states, e.g.:

- **raw**: the primary results of observations or measurements
- **annotated**: data that are connected to concepts, describing their meaning
- **QA assessed**: data that have undergone checks and are connected with descriptions of the results of those checks.
- **assigned metadata**: data that are connected to metadata which describe those data
- **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)

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

persistent data

Term (data) used as defined in ISO/IEC 10746-2. Data is the representations of information dealt by information systems and users thereof.

backup

A copy of computer 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. They can be:

- transformation rules for arithmetic values (mapping from one unit to another) from linear functions like $k.x + d$ to multivariate functions
- transformation rules for ordinal and nominal values, e.g., transforming classifications according to a classification system A to classification system B
- transformation rules for data descriptions (metadata or semantic annotation or QA annotation)
- transformation rules for Parameter names and descriptions (can be n:m)
- transformation rules for Method names and descriptions
- transformation rules for Sampling descriptions

data provenance

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

An creation of an entrance into the data provenance records triggered by any actions typically contain:

- date of action;
- actor;
- type of action;
- data_id.

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.

service

Service or process, available for reuse.

service description

Services and processes, which are available for reuse, be it within an enterprise architecture, within a research infrastructure or within an open network like the Internet, shall be described to help avoid wrong usage. Usually such descriptions include the accessibility of the service, the description of the interfaces, the description of behaviour 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)

3.2.1.2 Information Action Types

Information actions model the information processing in the system. Every action is associated with at least one object. Actions cause state changes in the objects that participate in them.

Figure 3.13 shows a collection of all action types used to model the information viewpoint.

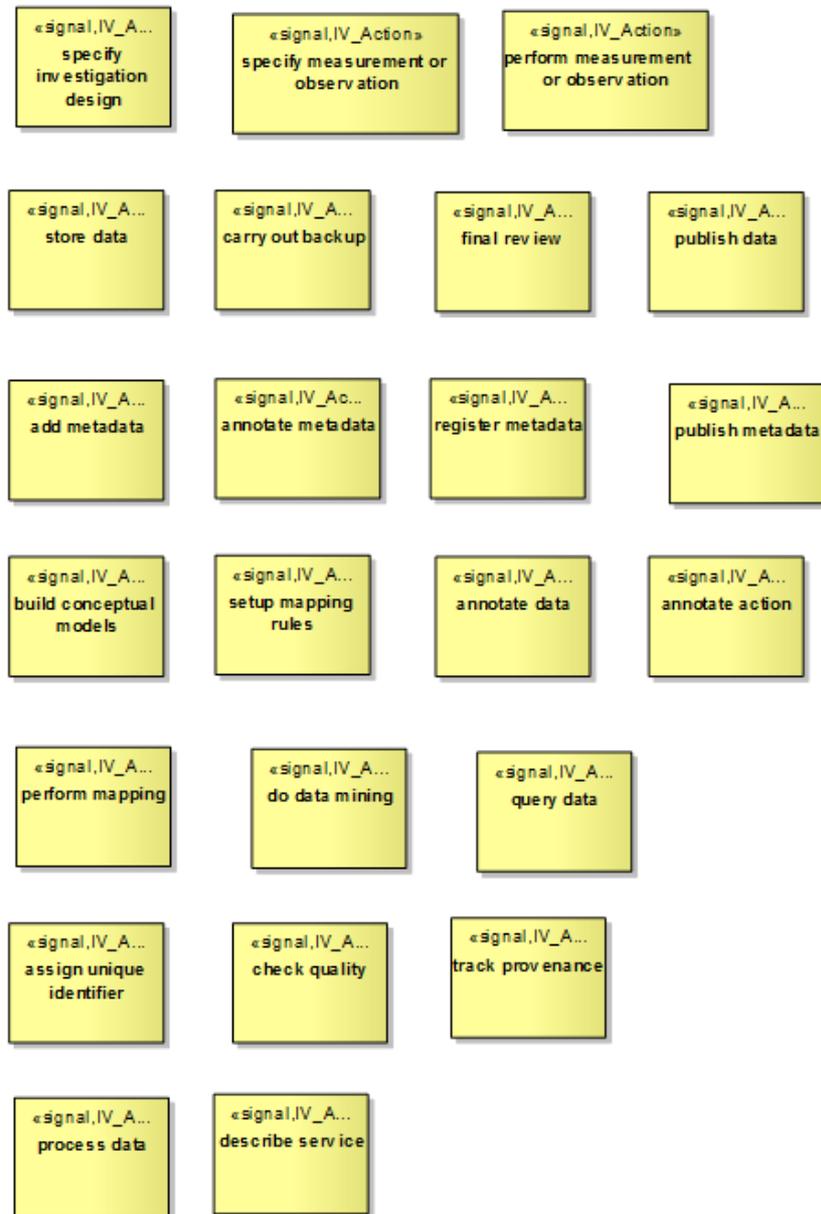


Figure 3.13: Information Objects Action Types

specify investigation design

specify design of investigation, including sampling design:

- geographical position of measurement or observation (site) -- the selections of observations and measurement sites,e.g., can be statistical or stratified by domain knowledge;
- characteristics of site;
- preconditions of measurements.



ENVRI Common Operations of Environmental Research Infrastructures

specify measurement or observation

Specify the details of the method of observations/measurements.

For example, this may include the specification of a measurement device type and its settings, measurement/ observation intervals.

perform measurement observation

Measure parameter(s) or observe an event. The performance of a measurement or observation produces measurement results.

store data

Archive or preserve data in persistent manner to ensure continuing accessible and usable.

carry out backup

Replicate data to an additional data storage so it may be used to restore the original after a data loss event. A special type of backup is a long term preservation.

final review

Review the data to be published, which will not likely be changed again.

The action triggers the change of the data state to be "finally reviewed". In practices, an annotation for such a state change should be recorded for provenance purposes. Usually, this is coupled with archiving and versioning actions.

publish data

Make data public accessible.

For example, this can be done by:

- presenting them in browsable form on the world wide web
- by presenting them via special services:
 - RESTful service
 - SOAP service
 - OPEN GRID service
 - OGC service (web feature service, web map service)
 - SPARQL endpoint

check quality

Actions to verify the quality of data.

For example it may involve:

- remove noise
- remove apparently wrong data
- calculate calibrations

Quality checks can be carried out at different points in the chain of data lifecycle.

Quality checks can be supported by software tools for those processes which can be automated (e.g. statistic tolerance checks).



ENVRI Common Operations of Environmental Research Infrastructures

assign Unique Identifier

Obtain a unique identifier and associate it to the data.

add metadata

Add additional information according to a predefined schema (metadata schema). This partially overlaps with data annotations.

annotate data

Link metadata with meaning (concepts of predefined local or global conceptual models). This can be done by adding tags or a pointer to concepts within a conceptual model to the data. If the concepts are terms e.g., in an SKOS/RDF thesaurus, and published as linked data, then this would mean entering the URL of the term describing the meaning of the data.

register metadata

Enter the metadata into a metadata catalogue.

publish metadata

Make the registered metadata available to the public.

query metadata

Send a request to metadata resources to retrieve metadata of interests.

track provenance

Add information about the actions and the data state changes as data provenances.

build conceptual models

Establish a local or global model of interrelated concepts.

This may involve the following issues:

- commitment: the agreement of a larger group of scientists /data providers / data users should be achieved;
- unambiguousness: the concept model should be unambiguously defined;
- readability: the model should be readably by both human and machine. E.g., ontologies should express the meaning of the concepts with the relations to other concepts. This form is human and machine readable. Recently it has increasingly become important to add definitions in human readable language.
- availability: the conceptual model must be referenceable and dereferenceable for a long time

setup mapping rules

Specify the mapping rules of data and/or concepts.

These rules should be explicitly expressed by a language so that can be processed by software tools.

A minimal set of mapping rule should include the following information:

- source data / concept for which the mapping is valid
- target data / concept, for which the mapping is valid
- mapping process (the translation and or transformation process)
- validity constraints for the mapping (temporal constraints, context constraints, etc.)



ENVRI Common Operations of Environmental Research Infrastructures

annotate data

Annotate data with meaning (concepts of predefined local or global conceptual models).

In practices, this can be done by adding tags or a pointer to concepts within a conceptual model to the data. If the concepts are terms e.g., in an SKOS/RDF thesaurus, and published as linked data, then data annotation would mean to enter the URL of the term describing the meaning of the data.

There is no exact borderline between metadata and semantic annotation.

annotate action

to be defined

resolve annotation

to be defined

perform mapping

Execute transformation rules for values (mapping from one unit to another unit) or translation rules for concepts (translating the meaning from one conceptual model to another conceptual model, e.g. translating code lists).

query data

Send a request to data resources to retrieve data of interests.

In practices, there are two types of data query exist:

- two step approach:
 - step 1: query/search metadata;
 - step 2: access data

For example, when using OGC services, it usually first invokes a web feature service to obtain feature descriptions, then a web map service can be invoked to obtain map images.

- one step approach: to query data e.g., by using SQL services or SPARQL endpoints

Requests can be directly sent to a service or distributed by a broker.

do data mining

Execute a sequence of metadata / data request --> interpret result --> do a new request

Usually this sequence helps to deepen the knowledge about the data. Classically this sequence can:

- lead from data to metadata and semantic annotations
- follow the provenance of data
- can follow data processing

It can be supported by special software that helps to carry out that sequence of data request and interpretation of results.

process data

Process data for the purposes of:

- converting and generating data products
- calculations: e.g., statistical processes, simulation models
- visualisation: e.g., alpha-numerically, graphically, geographically

Data processes should be recorded as provenance.



describe service

Describe the accessibility of a service or processes, which is available for reuse, the interfaces, the description of behaviour and/or implemented algorithms.

3.2.1.3 Information Object Instances

Figure 3.14 shows the collection of instances of information objects which are information objects existing more than once having several instances.

Instances of information objects are needed for two purposes:

1. to show the data state changes as effects of actions;
2. to show the fact that there usually is a "local conceptual model" and a "global conceptual model", both of which are conceptual models.

global concept

Concepts with a commitment of a whole data sharing community. Usually those concepts are part of global conceptual models (global Thesauri like GEMET / EuroVoc / AGROVOC or global ontologies like Gene Ontology , ...).

local concept

A concept of

- person
- institute
- anything else

A concept can be local or global depending on the size of the community which commits to it and only if considered in relation to each other.



ENVRI Common Operations of Environmental Research Infrastructures

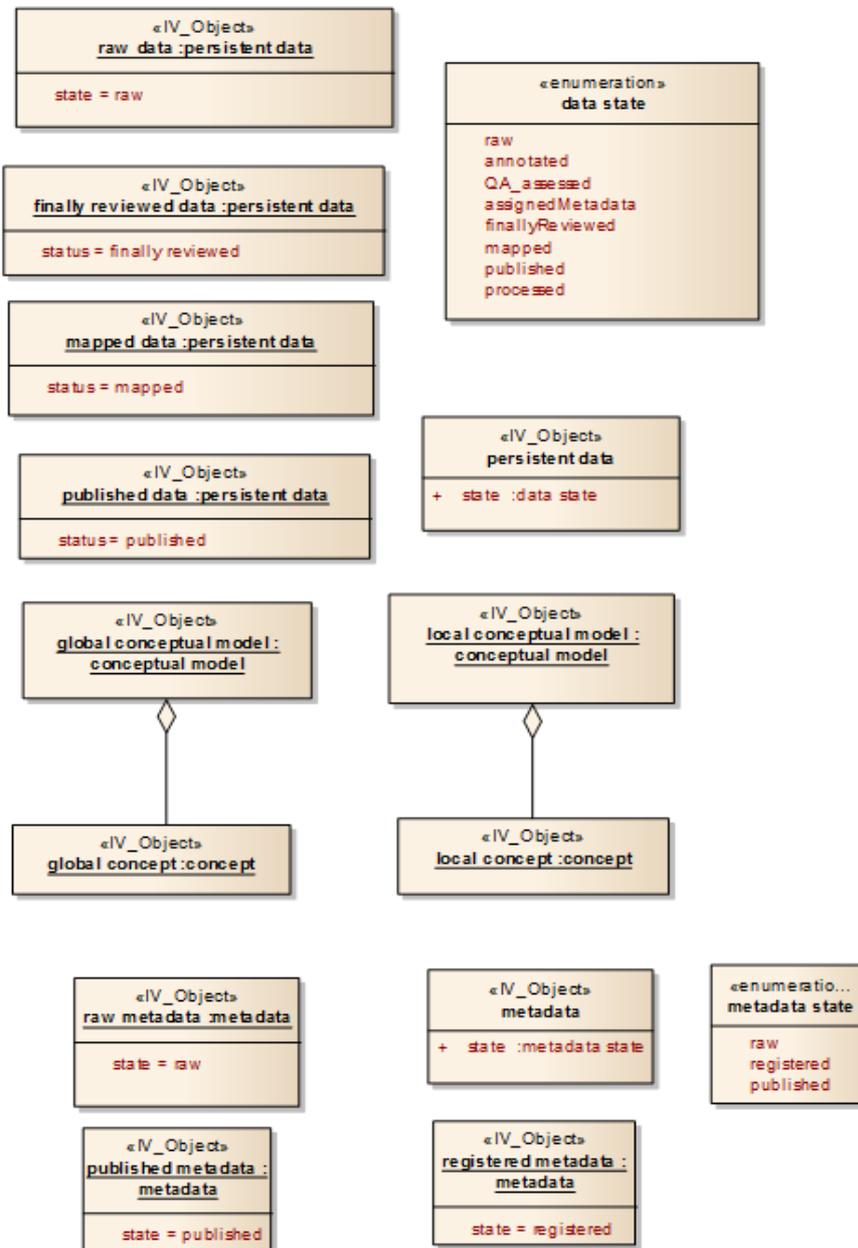


Figure 3.14: Instances of Information Objects



3.2.1.4 Data States

In the ENVRI Information specification, the following data states are defined:

- **raw**: the primary results of observations or measurements;
- **annotated**: data that are connected to concepts, describing their meaning;
- **QA assessed**: data that have undergone checks and are connected with descriptions of the results of those checks;
- **metadata associated**: data that are connected to metadata which describe those data;
- **finally reviewed**: data that have undergone a final review and therefore will not be changed anymore;
- **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).

As an example, the state changes as effects of actions are illustrated in Figure 3.15.

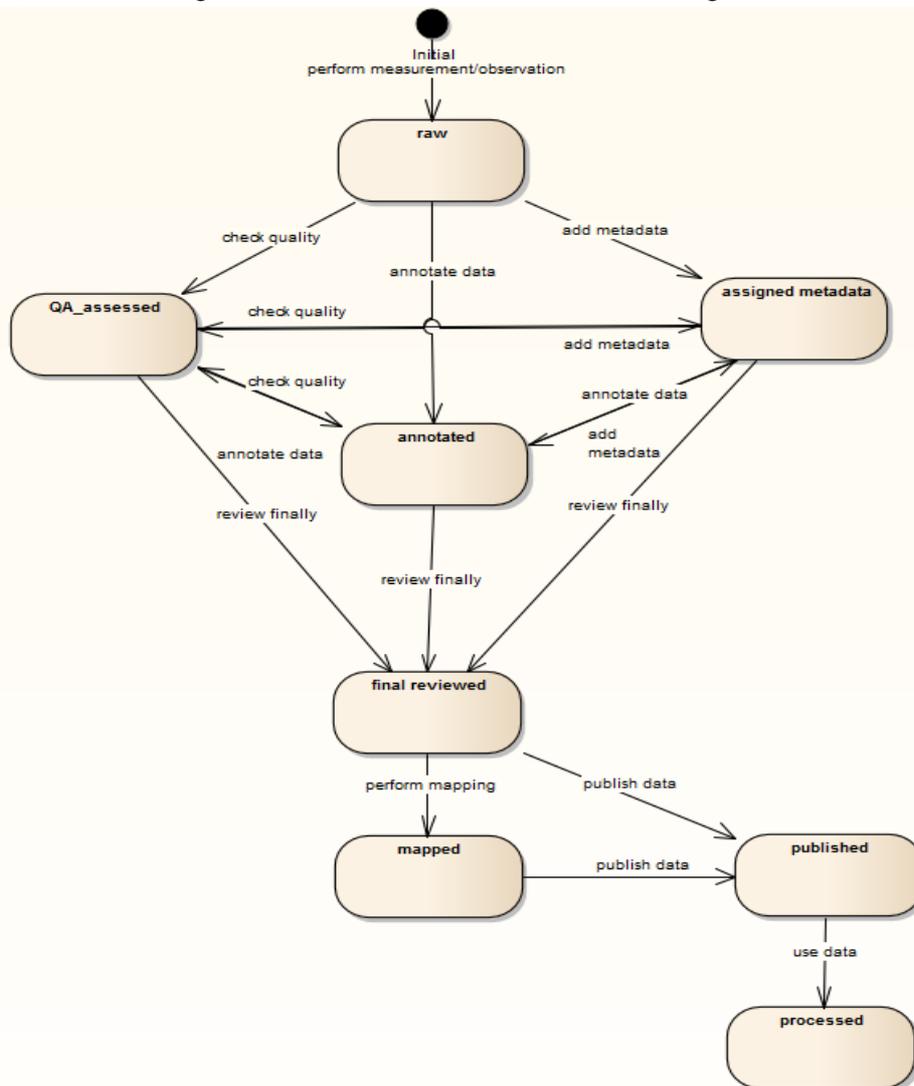


Figure 3.15: An Example of Data States Changes As Effects of Actions



3.2.2 Dynamic Schemata

In information viewpoint, dynamic schemata specify how the information evolves as the system operates, describing the allowable state changes of one or more information objects [37].

The specification consists of 2 parts:

- **Dynamic Schemata Overview**, which is an overview of the schemata and state changes of information objects as effects of actions;
- **Tracing of Provenance**, which summarises the provenance tracking points.

3.2.2.1 Dynamic Schemata Overview

An overview of the specification is illustrated in Figure 3.16. A more detailed specification is given in Appendix C.

Before a measurement or observation can be started the design (or setup) must be defined, including the working hypothesis and scientific question, method of the selection of sites (stratified / random), necessary precision of the observation or measurement, boundary conditions, etc. For correctly using the resulting data, detailed information about that process and its parameters have to be available for people processing the data. (e.g., if a stratified selection of sites according to parameter A is done, the resulting value of parameter A cannot be evaluated in the same way as other results)

After defining the overall design of measurements or observations, the measurement method, complying with the design, including devices which should be used, standards / protocols which should be followed, and other details have to be specified. Information of that process and the parameters resulting of the process have to be stored in order to guarantee correct interpretation of the resulting data. (e.g. when you want to model a dependency of parameter B of a parallel measured wind velocity, the limit of detection of the used anemometer influences the range of values of possible assertions).

When the measurement or observation method is defined, it can be carried out, producing measurement results. The handling of those results, all the actions done, to store the data are pulled together in the action "store data". (This action can be very simple when using a measurement device, which periodically sends the data to the data management system, but this can also be a sophisticated harvesting process or e.g. in case of biodiversity observations a process done by humans). The storage process is the first step in the life cycle of data that makes data accessible in digital form and are persistent.

As soon as data are available for IT purposes a backup can be made, independently of the state of the persisted data. This can be done locally or remote, done by the data owners or by dedicated data curation centres. At any status of the data can be processed for QA-assessments, for readjustment of the measurement or observation design and a lot of other reasons. Evaluations, which lead to answers of the scientific question, however, are usually done on data with a certain status - the status "finally reviewed".

Making data accessible for users outside the Environment of the data owner at least needs two steps: 1) Mapping the data to the "global" semantics, the semantics the data owner shares with the data user. 2) Publish the data. Mapping data to global semantics may include simple conversions like



ENVRI Common Operations of Environmental Research Infrastructures

conversions of units but also need more sophisticated transformations like transformations of code lists and other descriptions like the setup descriptions, measurement descriptions, and data provenance.

It is important to know about published data, whether those data have a status: "finally reviewed" and what such a status means. It can mean, that those data will never change again, the optimum for the outside user. But it might also mean, that only under certain circumstances those data will be changed. In this case it is important to know what "certain circumstances" means. And additionally it is important to know, how and where the used semantics are described. A resolvable pointer to them, of course is the solution which can be handled most easily.

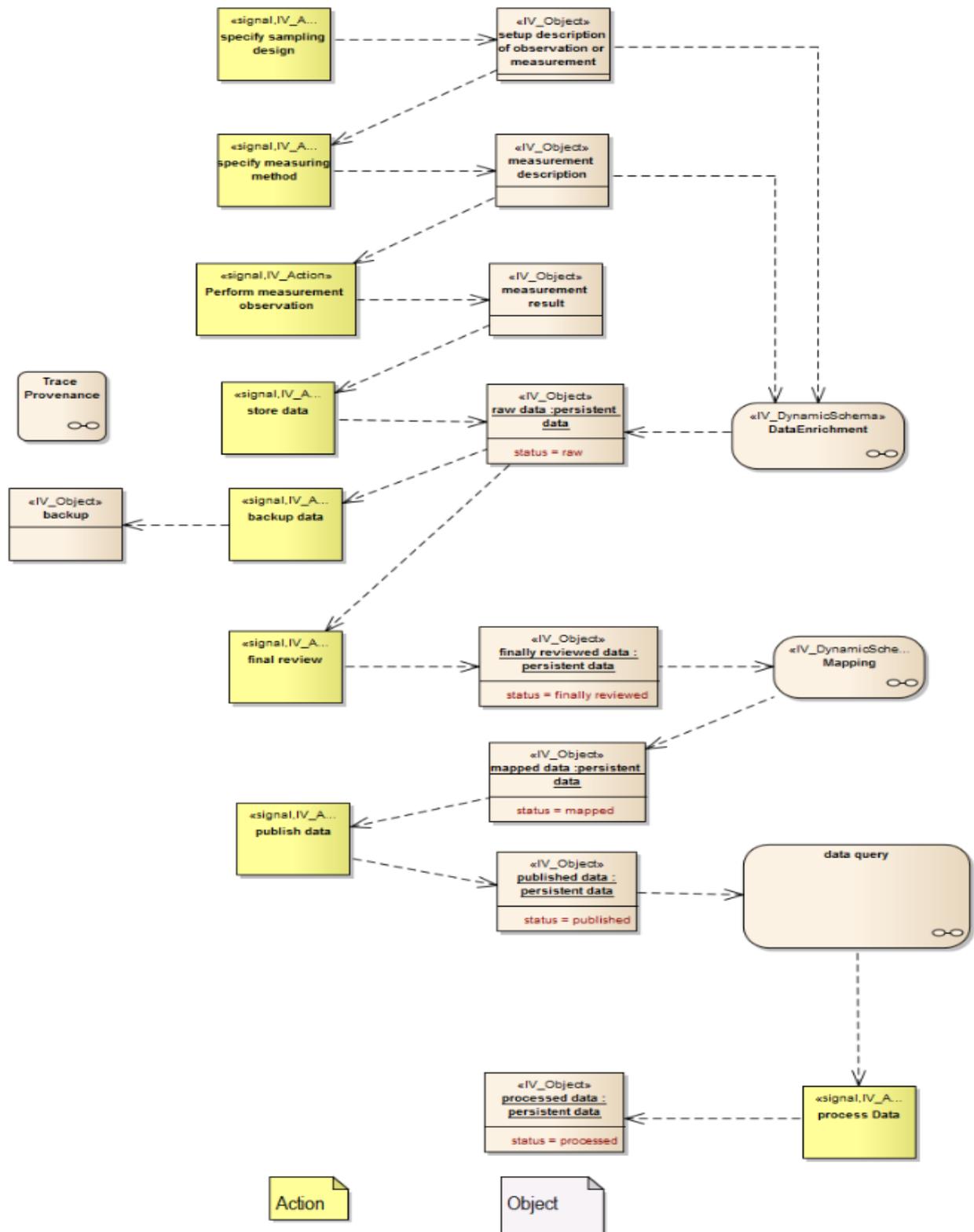
All the steps within the life cycle of data can be stored as data provenance, containing at least information about the used objects, the produced objects and the applied action. There are two important use cases for data provenance: 1.) citation of data and all the actors involved in the production of the data. 2.) correct interpretation of the data.

The states changes of information objects as effects of actions are summarised in the following table, which can be included as provenance information. For example, a provenance tracking service may record information objects being processed, action types applied and resulting objects and some additional data and store that step.

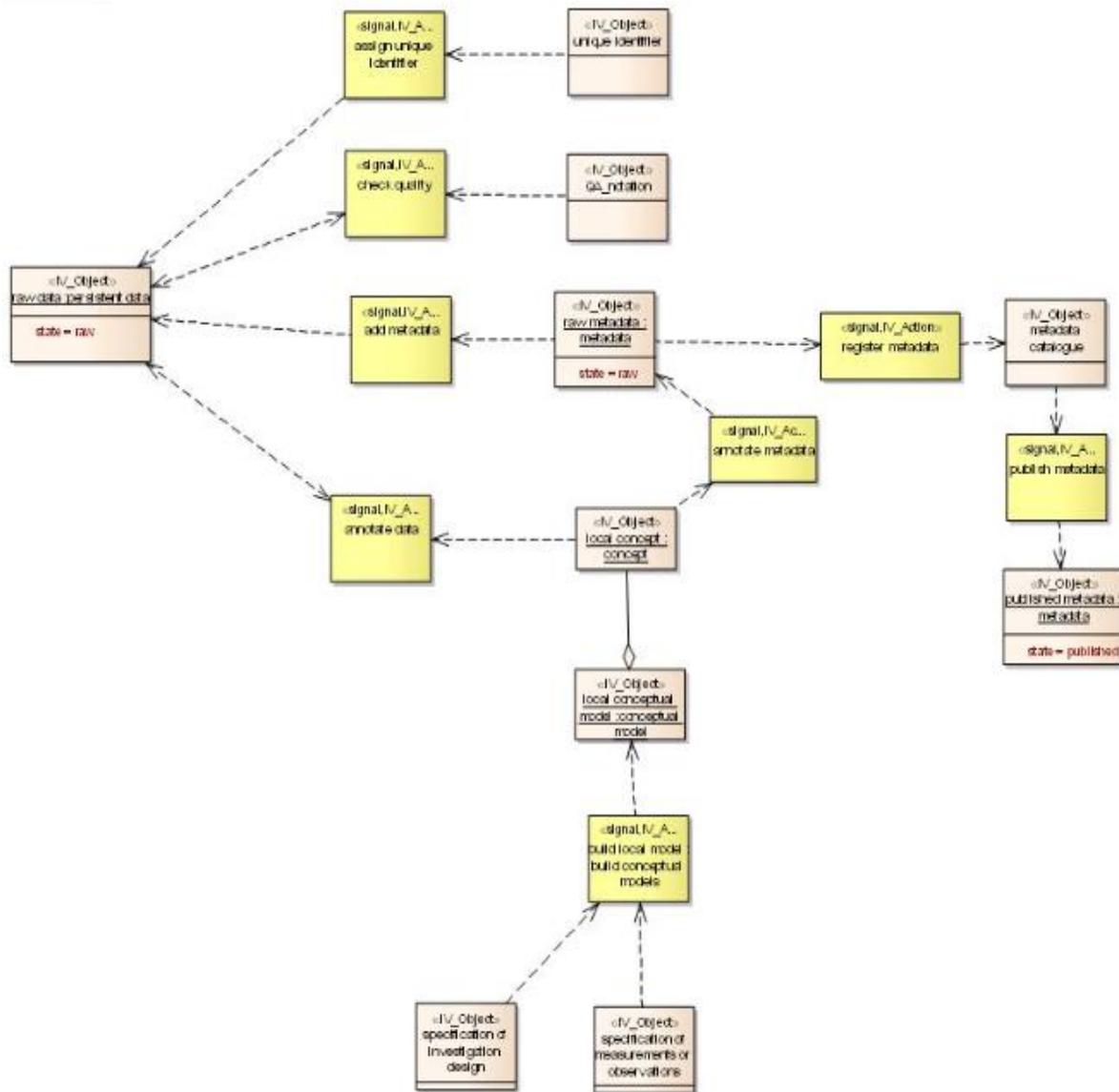
Information Object	Applied Action Types	Resulting Information Objects
setup description of observation or measurement	specify measurement method	measurement description
persistent data (state: raw data)	data enrichment (multiple actions)	persistent data (diverse enriched states)
persistent data (state: mapped)	publish data	persistent data (state: published)
persistent data (state: finally reviewed)	mapping	persistent data (state: mapped)
persistent data (state: published)	process data	persistent data (state: processed)
persistent data (any states)	carry out backup	backup
persistent data (all enrichments)	final review	persistent data (state: finally reviewed)
measurement result	store data	persistent data (state: raw data)
measurement description	perform measurement or observation	measurement result
	specify sampling design	setup description of observation or measurement



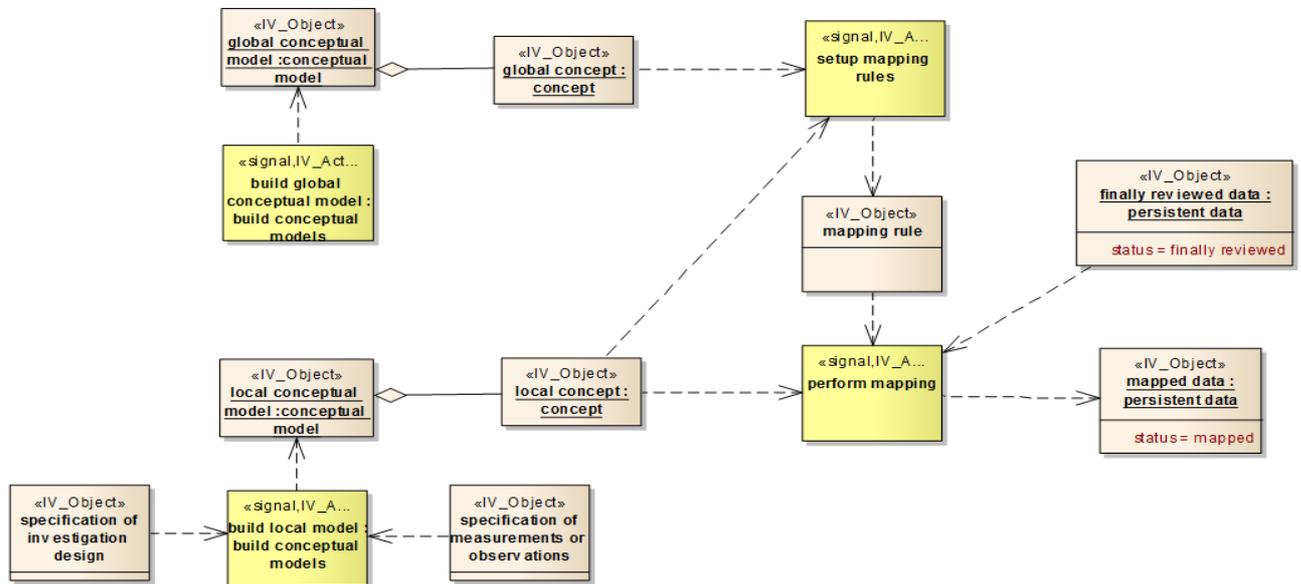
ENVRI Common Operations of Environmental Research Infrastructures



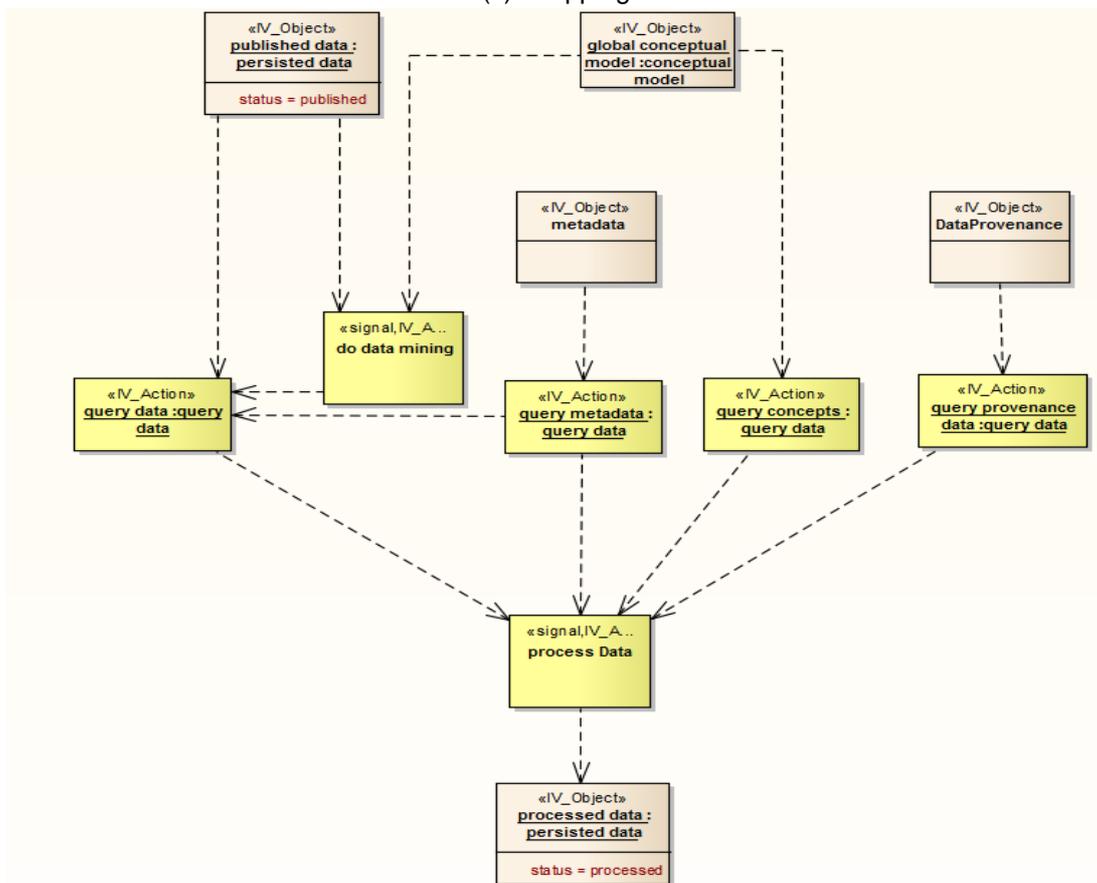
(a): Overview, the details of *Data Enrichment*, *Mapping*, and *Data Query* are depicted in below figures



(b): Data Enrichment



(c): Mapping



(d) Data Query

Figure 3.16: Dynamic Schemata Overview



3.2.2.2 Provenance Tracking

It is important to track state changes of information objects during their lifecycle. As illustrated in Figure 3.17, **track provenances** action is taken place at each point that action applied or any state changes of persistent data.

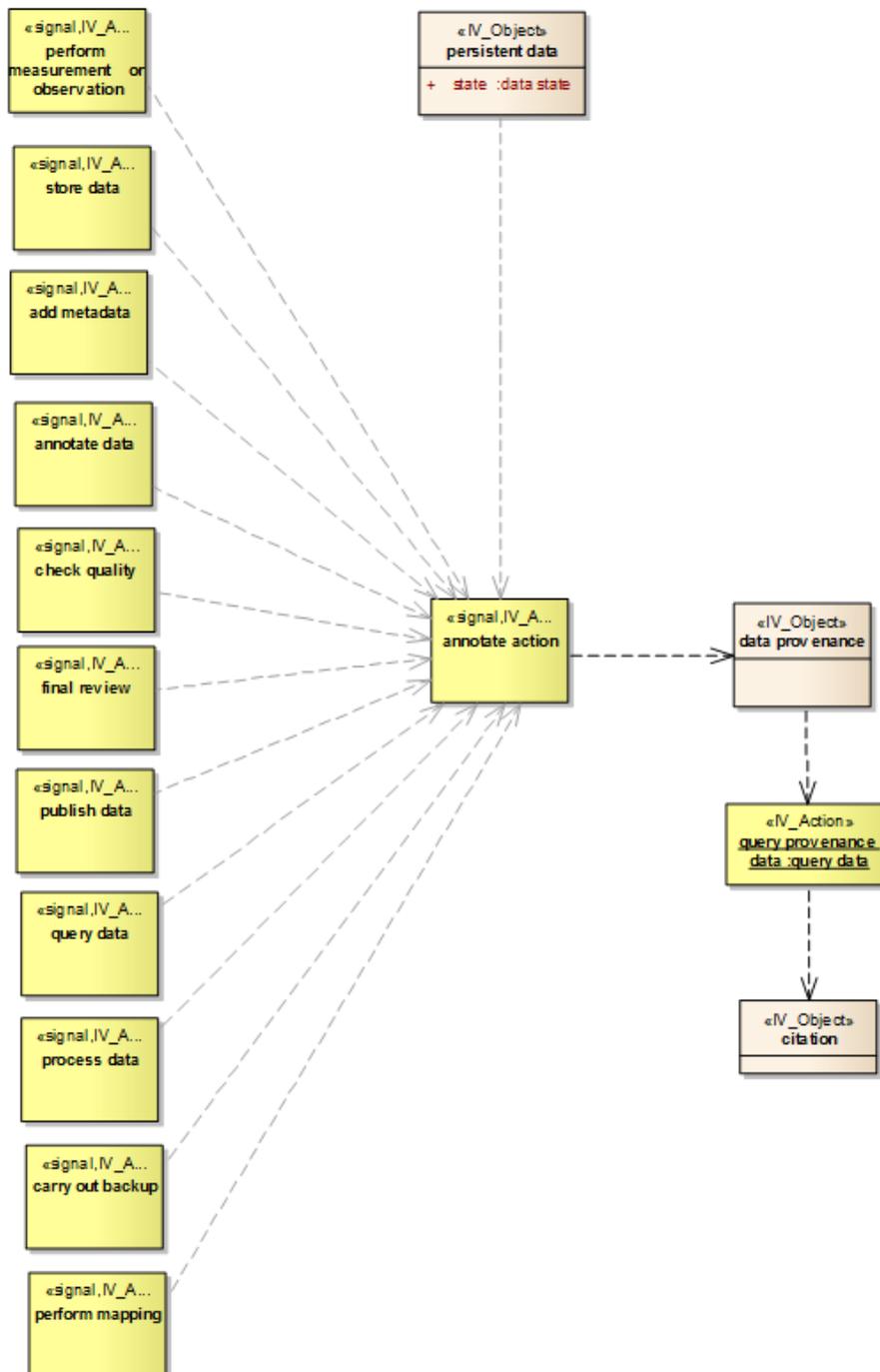


Figure 3.17: Tracing of Provenance



3.2.3 Static Schemata

Static Schemata specifies a minimum set of constraints when sharing data, in order to:

- A. avoid loss of information around measurements and observations;
- B. provide information about the meaning of data; and
- C. make data and additional information available.

Three static schemata are specified, constraints for **data collection**, **data integration** and **data publication**.

3.2.3.1 Constraints for Data Collection

A collection of constraints applied to data collection is illustrated in Figure 3.18, which can help avoid information loss or wrong information to be drawn out.

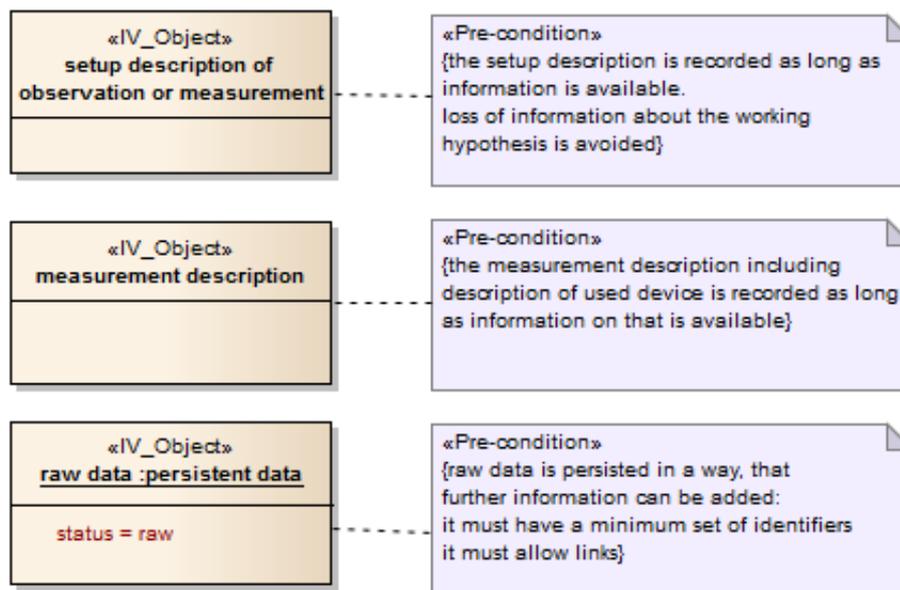


Figure 3.18: Constraints for Data Collection



3.2.3.2 Constraints for Data Integration

Figure 3.19 specifies the constraints for data integration, which is used for interpreting the meaning of data in order to help external data users correctly understand and map the semantics of data.

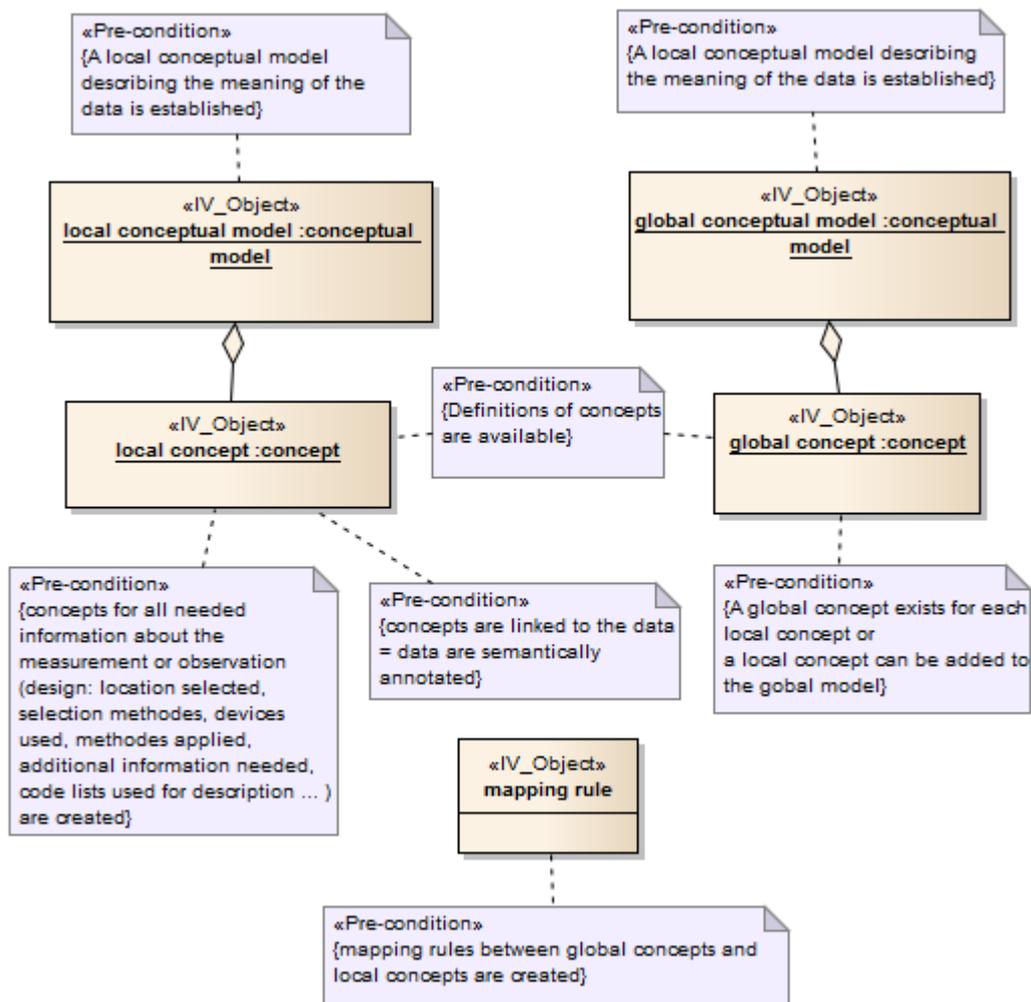


Figure 3.19: Constraints for Data Integration



3.2.3.3 Constraints for Data Publication

Constraints for data publication are described in Figure 3.20 which specifies pre-conditions and constraints necessary for preparing the data to be public accessed.

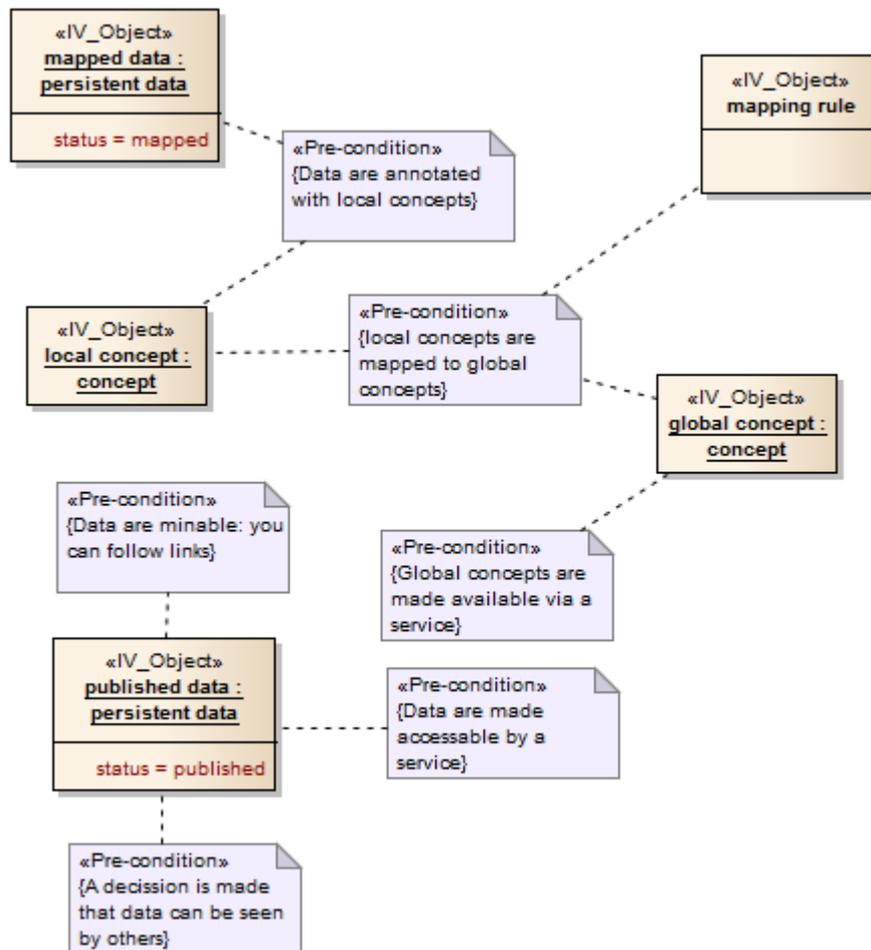


Figure 3.20: Constraints for Data Publication

3.2.4 Subsystems

For the purposes of easy observations, the defined model objects are regrouped into subsystems as defined in Section 2 Model Overview. Only 3 subsystems are discussed here:

- **Data Acquisition**, which consists of a group of information objects, and action types handled in the data acquisition subsystem;
- **Data Curation**, which consists of a group of information objects, and actions types handled in the data curation subsystem;
- **Data Access**, which consists of a group of information objects and actions types handled in the data access subsystems.



3.2.4.1 Data Acquisition

The information objects, and action types involved in the Data Acquisition subsystem are depicted in Figure 3.21, which supports the following functionalities:

- Measurement / Observation Design
- Observation
- Measurement

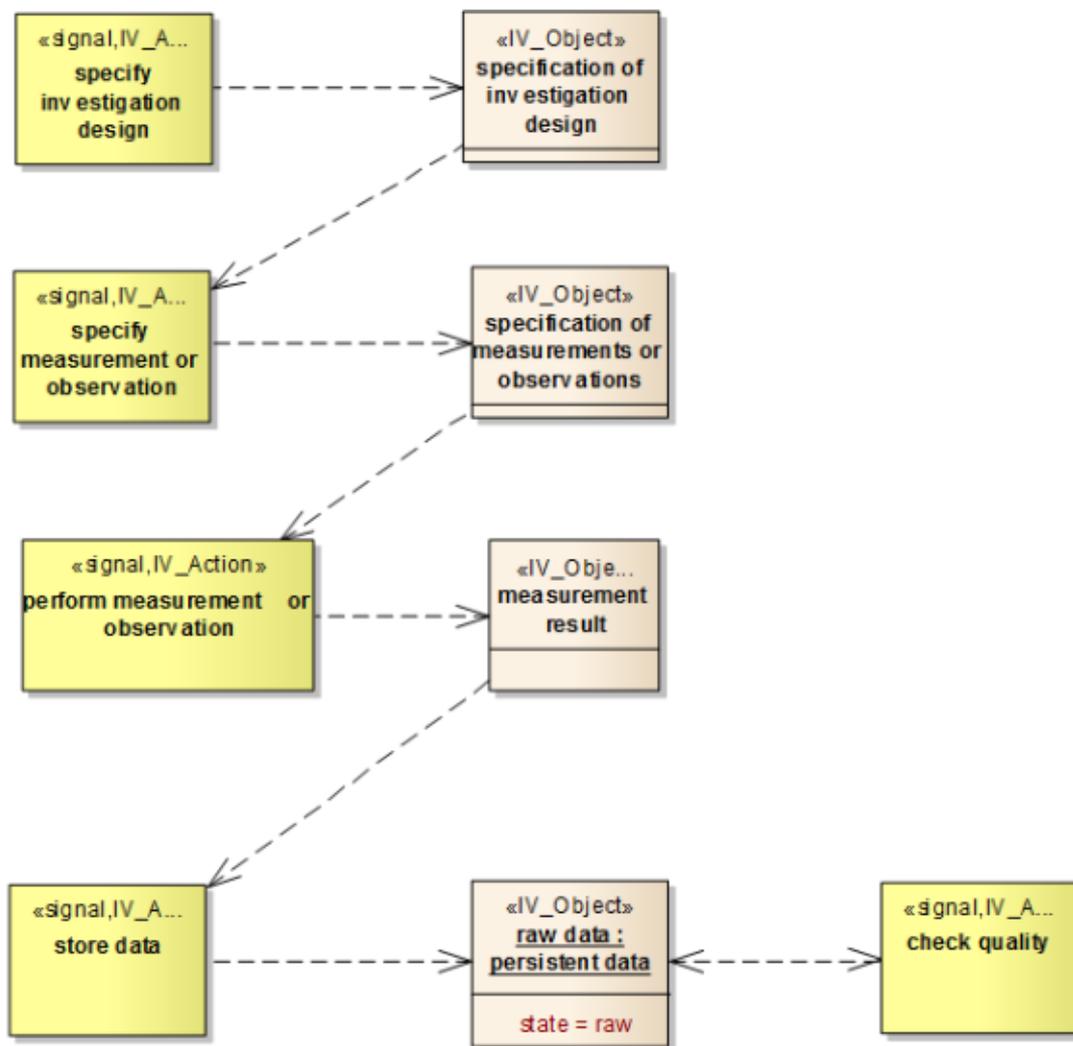


Figure 3.21: Information Specification of Data Acquisition Subsystem

A design for observation / measurements, is established and described in the setup description. Then the measurement method is specified including specifications for the measurement devices. The measurement or observation leads to results which are persisted (stored).



3.2.4.2 Data Curation

Figure 3.22 describes the information objects, and action types involved in the Data Curation subsystem, which support the following functionalities:

- persistent data preservation
- assignment of metadata
- semantics annotation
- quality assurance
- data identification - assignment of persistent identifier
- data backup

Raw data are collected by observations and measurements, and curated and stored in a data curation subsystem.

The following data curation operations are optional:

- QA annotation,
- obtain unique identifiers
- make a backup (mostly for long-term data persistence)
- connect the data to a (machine-readable) description - make semantic annotations - create contextual metadata
- Link the data to metadata, making data searchable

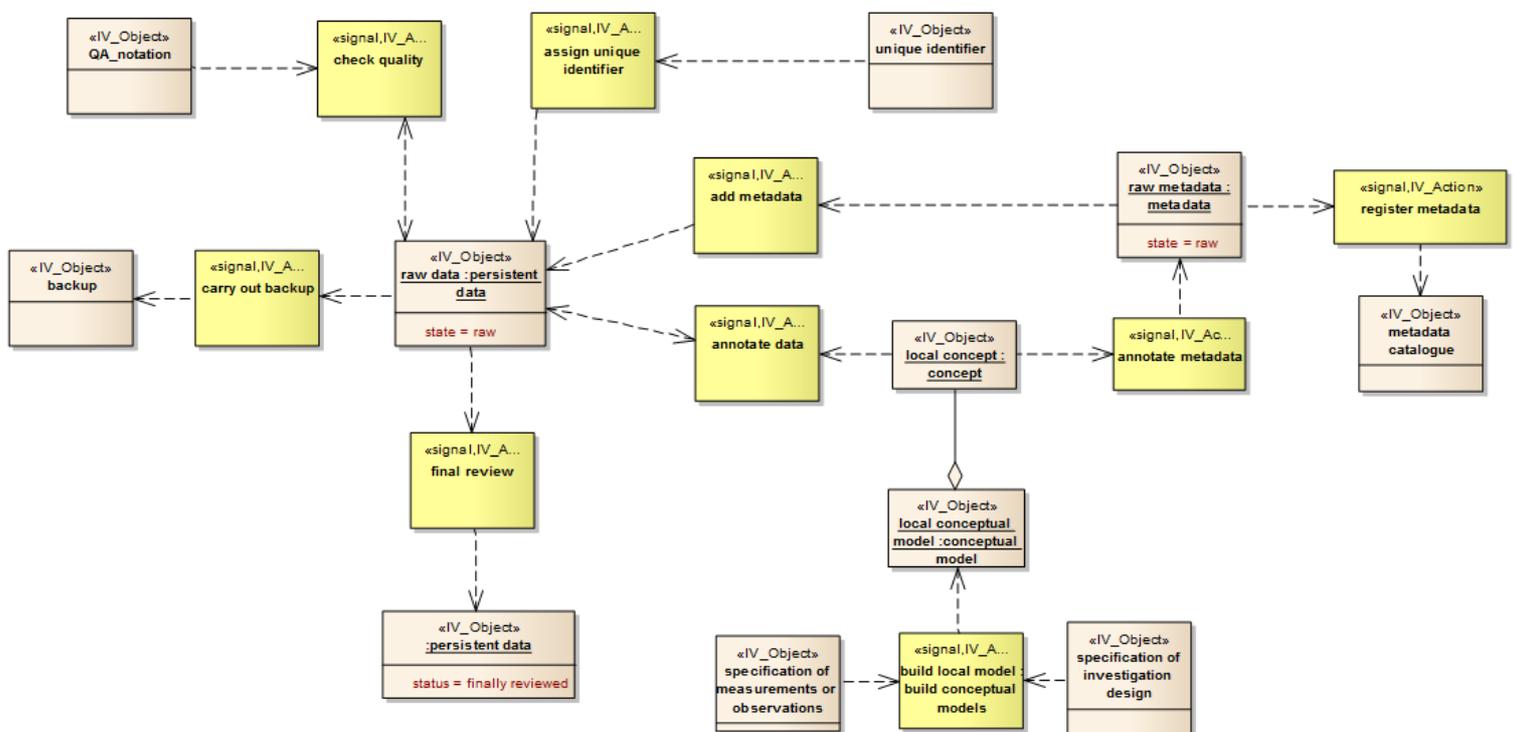


Figure 3.22: Information Specification of Data Curation Subsystem



3.2.4.3 Data Access

The information objects, and action types involved in the Data Access subsystem are shown in Figure 3.23, which supports the following functionalities:

- final review of data to be published
- data mapping (semantic mediation)
- data publication
- data processing

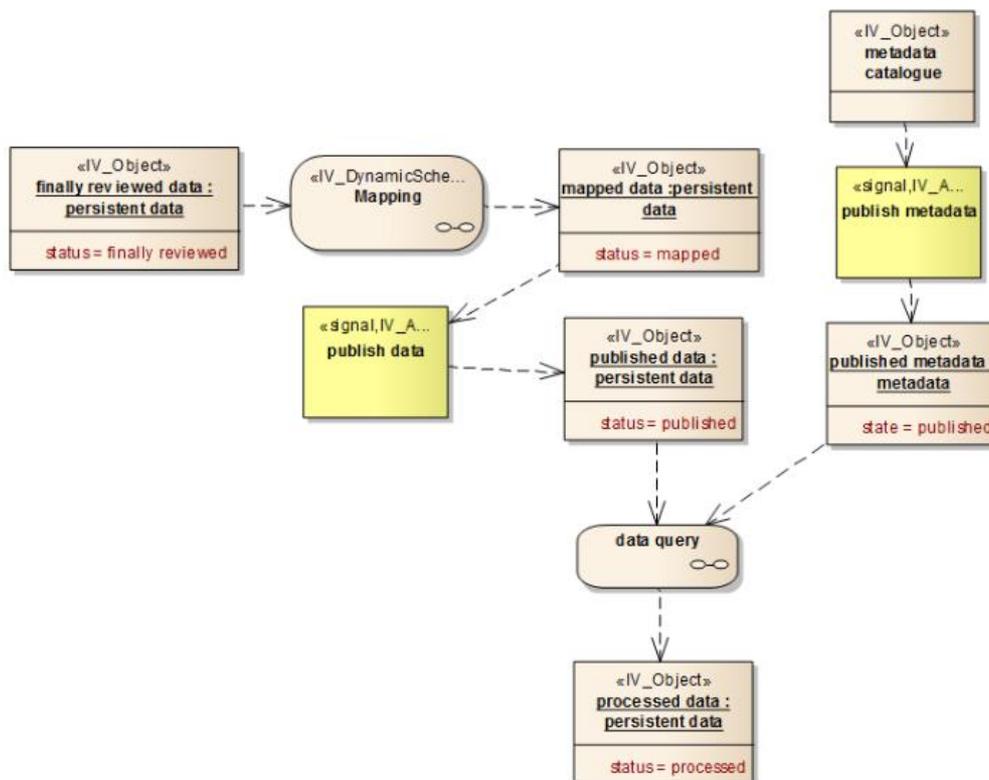


Figure 3.23: Information Specification of Data Access Subsystem

3.3 Computational Viewpoint

A research infrastructure provides a context within which researchers can interact with scientific data. Researcher and data can be thought of as being linked via an extended chain of interactions threaded through a series of intermediary interfaces, where each interface is charged with realising a particular service. Only by providing a sufficient range of services can a research infrastructure support all the different interaction chains that might emerge between users, data and resources.

The **computational viewpoint** of the Model accounts for the major computational objects expected within an environment research infrastructure and the interfaces by which they can be interacted with. Each object encapsulates functionality that needs to be collectively implemented by a service or resource within the infrastructure; each object also provides a number of interfaces by which that functionality can be invoked, or by which the object can invoke the functions of other objects. By binding compatible interfaces together, a network of interactions between objects can be described that demonstrates the computational requirements of a compliant research infrastructure.



ENVRI Common Operations of Environmental Research Infrastructures

Overview

The archetypical environmental research infrastructure is considered here as having a brokered, service-oriented architecture. Core functionality is encapsulated in a number of *service* objects that control various *resources* present in the infrastructure. Access to many of these internal services (particularly when invoked by external agents) is overseen by one or more *brokers* charged with validating requests and providing, where needed, an interoperability layer between heterogeneous components --- this is particularly important for federated infrastructures that may not be able to universally enforce a core set of data interchange and service standards across the entire infrastructure.

Each of the five subsystems of the Model prescribes a number of *computational objects* that should be present in an environmental research infrastructure.

- **Data acquisition**
- **Data curation**
- **Data access**
- **Data processing**
- **Community support**

For each of these objects, necessary *interfaces* are identified and the core *interactions* involving them are described. Particular attention is paid to interactions between subsystems, as many critical functions intercede in the movement of data from one subsystem to another.

A number of reference interactions describing the interaction between objects in different subsystems have been provided:

- **Brokered data export** (the export of user-requested data);
- **Brokered data import** (the import of user-provided data);
- **Brokered data query** (the querying of curated data by users);
- **Citation** (the resolution of data and resources cited in publications);
- **Instrument integration** (the integration of new instruments for data acquisition into the infrastructure).
- **Internal data staging** (the preparation of curated data for processing);
- **Processed data import** (the curation of results derived from processing);
- **Raw data collection** (the acquisition of raw data from integrated data sources);

The aggregation of these core interactions forms a minimal computational model for environmental science research infrastructures.

Computational modelling

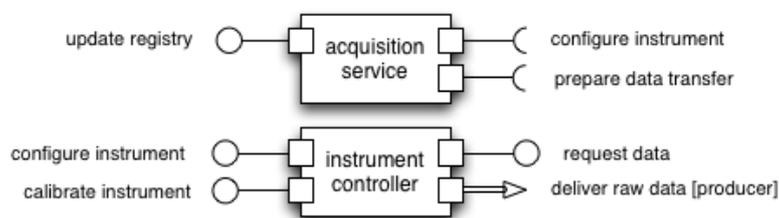
The computational viewpoint prescribed by the Open Distributed Process is concerned with the modelling of computational objects and the interactions between their interfaces. The Reference Model uses a lightweight subset of the full ODP specification to model the abstract computational requirements of the archetypical environmental science research infrastructure.



ENVRI Common Operations of Environmental Research Infrastructures

☞ The encapsulation of computational objects (and interfaces) occurs at a conceptual level rather than the implementation level; it is perfectly admissible for the functions of a given object to be distributed across multiple computational resources in an implemented infrastructure, should that be supported by its architecture, if that distribution does not interfere with the ability to implement all of that object's interfaces (and thus behaviours). Likewise the functionalities of multiple objects can be gathered within a single implemented service, should that be desired.

The first-class entity of the computational viewpoint is the *computational object*:



A computational object encapsulates a set of functions that need to be collectively implemented by a service or resource within an infrastructure. To access these functions, a computational object also provides a number of *operational* interfaces by which that functionality can be invoked; the object also provides a number of operational interfaces by which it can itself invoke functions on other objects. Each computational object may also have *stream* interfaces for ferrying large volumes of data within the infrastructure. In summary:

- **Operation interfaces** are used to pass messages between objects used to coordinate general infrastructure operations such as querying a data resource or configuring a service. A given operation interface must be either a *server* (providing functions to be invoked by other objects) or a *client* (providing functions by which operations can be invoked on other objects).

☞ In diagrams, client and server interfaces are linked using 'ball and socket' notation: clients expose sockets (half-circles) whilst servers expose balls (closed-circles).

- **Stream interfaces** are used to deliver data from one part of the infrastructure to another. A *producer* interface streams data to one or more bound *consumer* interfaces as long as there is data to transfer and the consumer is available to receive that data. Data channels are typically established by operations invoked via operational interfaces, but can persist independently of them.

☞ In diagrams, producer and consumer stream interfaces are linked using a double-arrow notation: the arrow-head points away from producers, towards consumers.

As well as having interfaces by which to interact with other objects, some computational objects possess the right to create other computational objects; this is done typically to deploy transitory services or to demonstrate how an infrastructure might extend its functionality.





ENVRI Common Operations of Environmental Research Infrastructures

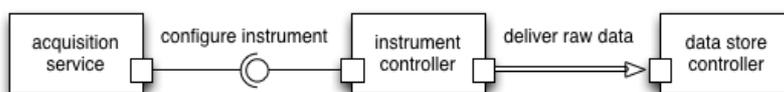
Some objects extend the functionality of other objects; these objects possess all the interfaces of the parent (usually in addition to some of their own) and can be created by the same source object if the capability exists.

- ☞ In diagrams, the ability to create objects is noted by a single filled arrow extending from the creating object to the object being created, with the annotation 'new <object>'. If one object extends another, then this can be illustrated using an unfilled arrow from the sub-object to the parent, with the annotation 'is a'.

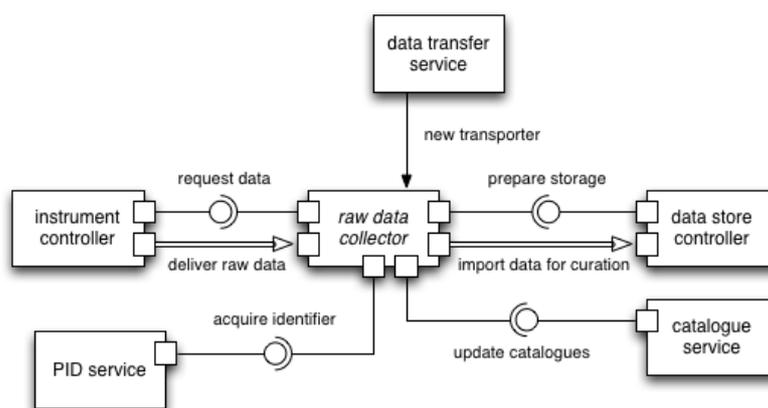
Each interface on a computational object supports a number of either operations or data streams, the logical signatures of which determine the bindings that can be made between them. A *binding* is simply an established connection between two or more interfaces in order to support interaction between the computational objects to which those interfaces belong. A client interface can be bound to any server interface that supports all its prescribed functionality.

- ☞ For simplicity, many client and server interfaces designed to work together in the Reference Model share the same name; when bound in diagrams, the binding itself is identified by that shared name.

Once bound via their corresponding interfaces, two objects can invoke functions on one another to achieve some task (such as configuration of an instrument or establishment of a persistent data movement channel).



Primitive bindings can be established between any client / server pair or producer / consumer pair as appropriate. Compound bindings between three or more interfaces can be realised via the creation of *binding objects*, a special class of transitory computational object that can be used to coordinate complex interactions by providing primitive bindings to all required interfaces.



The use of binding objects removes the imperative to decompose complex interactions into sets of pairwise bindings between objects; this suits the level of abstraction at which the Reference Model is targeted, the specific distribution of control between interacting objects being specific to different infrastructure architectures.

- ☞ The names of binding objects are typically italicised in diagrams to better distinguish them from 'normal' computational objects.



3.3.1 Data Acquisition

The basis for environmental research is the observation and measurement of environmental phenomena. The archetypical environmental research infrastructure provides access to data harvested from an extended network of sensors, instruments and other contributors deployed in the field.

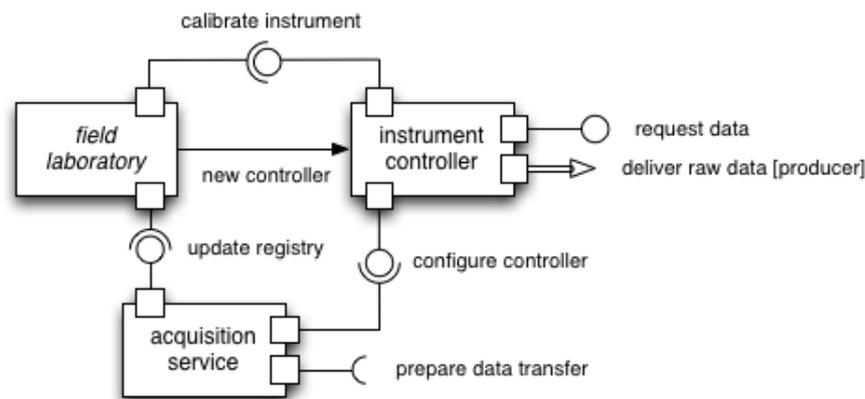


Figure 3.24: Computational Specification of Data Acquisition Subsystem

The **data acquisition** subsystem provides pathways by which data sources integrated into the infrastructure can deliver data to suitable data stores. Data acquisition is computationally described as a set of *instrument controllers* (encapsulating the accessible functionalities of instruments and other raw data sources out in the field), monitored and managed by one or more *acquisition services* (responsible for ensuring that any data is delivered into the infrastructure in accordance with current policies).

Acquisition is manipulated via *field laboratories*, community proxies by which authorised agents can add and remove instruments from the network (by registering and de-registering instrument controllers) as well as calibrate instrument readings where applicable in accordance with current community best-practice.

acquisition service

Oversight service for integrated data acquisition.

An acquisition service object encapsulates the computational functions required to monitor and manage a network of instruments. An acquisition service can translate acquisition requests into sets of individual instrument configuration operations as appropriate.

An acquisition service should provide at least three operational interfaces:

- **update registry (server)** provides functions for registering and deregistering instruments within the data acquisition subsystem.
- **configure controller (client)** is used to configure data collection (and other configurable factors) on individual instruments.
- **prepare data transfer (client)** is used to negotiate data transfers into the data curation subsystem.



ENVRI Common Operations of Environmental Research Infrastructures

field laboratory

Community proxy for interacting with data acquisition instruments.

A sub-class of virtual laboratory binding object encapsulating the functions required to access, calibrate, deploy or undeploy instruments within the data acquisition subsystem.

☞ A field laboratory is created by a science gateway in order to allow researchers in the field to interact with the data acquisition subsystem.

Deployment of an instrument entails the deployment of an instrument controller by which the instrument can be interacted with.

- A field laboratory object can instantiate any number of *instrument controller* objects.
- A field laboratory should provide at least two operational interfaces in addition to those provided by any virtual laboratory:
- **calibrate instrument (client)** is used to calibrate the reading of data by instruments based (in principle) on scientific analysis of data output. This interface can also be used to monitor activity on a given instrument.
- **update registry (client)** is used to register and unregister instruments with the data acquisition subsystem.

☞ The degree of freedom with which a field laboratory interacts with the instruments of a data acquisition subsystem is contingent on the nature of the research infrastructure and policed by a security service object (as defined for all user laboratories).

instrument controller

An integrated raw data source.

An instrument is considered *computationally* to be a source of raw environmental data managed by an acquisition service. An instrument controller object encapsulates the computational functions required to calibrate and acquire data from an instrument.

☞ 'Instrument' is a logical entity, and may to multiple physical entities deployed in the real world should they act in tandem sufficiently closely to justify being treated as one data source. Any instrument represented by an instrument controller should however be considered independently configurable and monitorable from other instruments managed by the same acquisition service.

An instrument controller should provide three operational interfaces:

- **calibrate instrument (server)** provides functions to calibrate the reading of data by an instrument (if possible).
- **configure controller (server)** provides functions to configure how and when an instrument delivers data to a data store.
- **request data (server)** provides functions to directly request data from an instrument.
- An instrument controller should provide at least one stream interface:
- **deliver raw data (producer)** is used to deliver raw data streams to a designated data store.



3.3.2 Data Curation

One of the primary responsibilities of an environmental research infrastructure is the curation of the significant corpus of acquired data and derived results harvested from the data acquisition subsystem, data processing and community contributions. Scientific data must be collected, catalogued and made accessible to all authorised users. The accessibility requirement in particular dictates that infrastructures provide facilities to ensure easy availability of data, generally by replication (for optimised retrieval and failure-tolerance), assignment of persistent identifiers (to aid discovery) and cataloguing (aiding discovery and allowing more sophisticated requests to be made over the entirety of curated data).

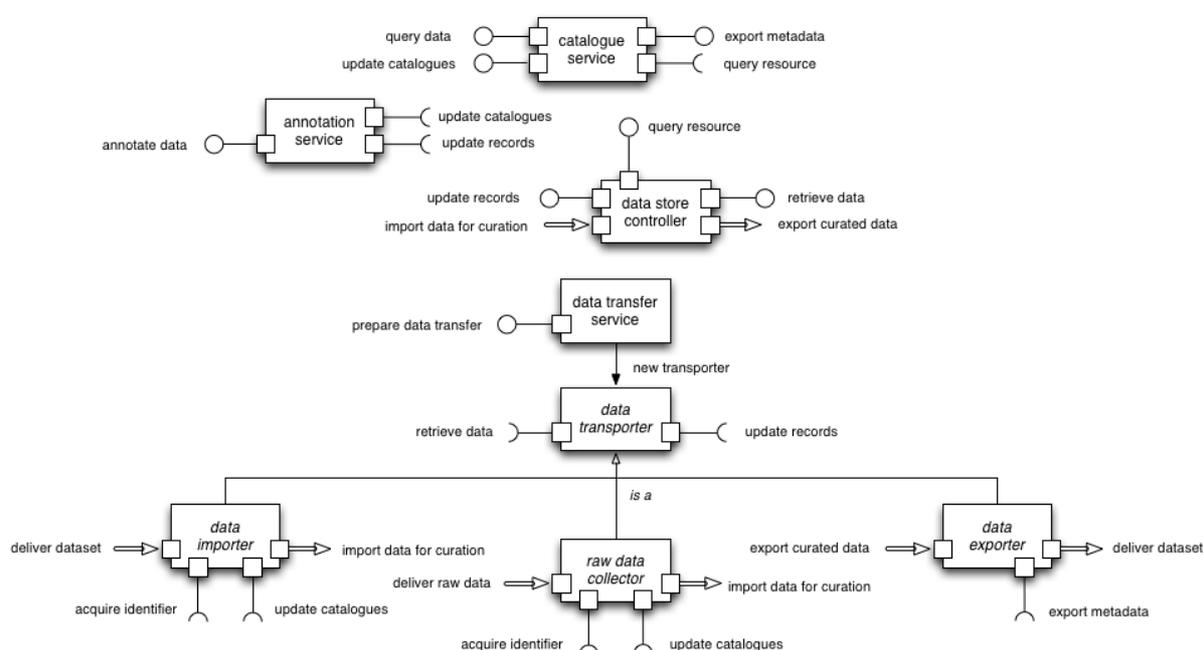


Figure 3.25: Computational Specification of Data Curation Subsystem

The **data curation** subsystem provides the core services required for data preservation and management. Computationally, data curation is handled by a set of *data store controllers* (encapsulating the interfaces required to use data stores within the infrastructure), monitored and managed by a number of *curation services*, including:

- The **annotation service**, concerned with the updating of records (such as datasets) and catalogues in response to user annotation requests.
- The **catalogue service**, concerned with the cataloguing of metadata and other characteristic data associated with datasets stored within the infrastructure.
- The **data transfer service**, concerned with the movement of data into and out of the infrastructure.

Data store controllers provide access to data stores that may have their own internal data management regimes. A data transfer service is able to provide *data transporters* for managing the movement of data from one part of a research infrastructure to another; because a number of additional services may be involved in (for example) repackaging the data for different contexts, it is possible to derive a number of different kinds of data transporter. The Model defines the following:



ENVRI Common Operations of Environmental Research Infrastructures

- **Raw data collectors** manage the movement of data from one or more instruments in the data acquisition subsystem to one or more data stores in the data curation subsystem. A raw data collector can acquire identifiers for new datasets and can immediately update data catalogues in order to register such datasets and their characteristic metadata.
- **Data importers** manage the movement of data from external sources (such as user-originating datasets and derived datasets from data processing) to one or more data stores in the data curation subsystem. A data importer can acquire identifiers for new datasets and can immediately update data catalogues in order to register datasets.
- **Data exporters** manage the movement of data from one or more data stores in the data curation subsystem to an external destination (such as a user machine or downstream service gathering data from the research infrastructure). A data exporter can invoke the catalogue service in order to extract metadata to package alongside exported datasets in order to maintain contextual information including provenance.

annotation service

Oversight service for adding and updating records attached to curated datasets.

An annotation service object collects the functions required to annotate datasets and collect observations that can be associated with the various types of data managed within a research infrastructure.

An annotation service should provide three operational interfaces:

- **annotate data (service)** provides functions for requesting the annotation of existing datasets or the creation of additional records (such as qualitative observations made by researchers).
- **update catalogues (client)** is used to update catalogues or catalogue information managed by a catalogue service.
- **update records (client)** is used to directly annotate existing datasets curated within one or more data stores.

catalogue service

Oversight service for cataloguing curated datasets.

A catalogue service object collects the functions required to manage the construction and maintenance of catalogues of metadata or other characteristic data associated with datasets (including provenance and persistent identifiers) stored within data stores registered with the data curation subsystem.

☞ A data catalogue is itself a dataset, and can therefore be accessed and queried exactly as any other dataset.

A catalogue service should provide four operational interfaces:

- **export metadata (server)** provides functions for gathering metadata to be exported with datasets extracted from a data curation subsystem.
- **query data (server)** provides functions for querying data held by the infrastructure, including the retrieval of datasets associated with a given persistent identifier.
- **update catalogues (server)** provides functions for harvesting (meta)data from datasets in order to derive or update data catalogues.
- **query resource (client)** is used to retrieve data from data stores.



data store controller

A data store within the data curation subsystem.

Data stores record data collected by the infrastructure, providing the infrastructure's primary resources to its community. A data store controller encapsulates the functions required to store and maintain datasets and other data artefacts produced within the infrastructure within a data store, as well as to provide access to authorised agents.

A data store controller should provide three operational interfaces:

- **update records (server)** provides functions for editing data records within a data store as well as preparing a data store to ingest new data through its import stream interface described below.
- **query resource (server)** provides functions for querying the data held in a data store.
- **retrieve data (server)** provides functions to negotiate the export of datasets from a data store.

A data store controller should provide two stream interfaces:

- **import data for curation (consumer)** receives data packaged for curation within the associated data store.
- **export curated data (producer)** is used to deliver data stored within the associated data store to another service or resource.

data export

Binding object for exporting curated datasets.

A sub-class of *data transporter* binding object encapsulating the functions required to move and package curated datasets from the data curation subsystem to an outside destination.

A data exporter should provide at least one operational interface in addition to those provided by any data transporter:

- **export metadata (client)** is used to retrieve any additional metadata to be associated with the data being transferred.

☞ Generally, metadata is exported alongside datasets being exported from the infrastructure where data is repackaged to be more self-describing.

A data exporter must also provide two stream interfaces through which to pass data:

- **export curated data (consumer)** is used to retrieve curated datasets stored within data stores in the data curation subsystem.
- **deliver dataset (producer)** is used to deliver (repackaged) curated data to a designated external data store outside of the data curation subsystem.

data import

Binding object for importing external datasets.

A sub-class of *data transporter* binding object encapsulating the functions required to move and package external datasets from outside sources into the data curation subsystem.

A data importer should provide at least two operational interfaces in addition to those provided by any data transporter:



ENVRI Common Operations of Environmental Research Infrastructures

- **acquire identifier (client)** is used to request a new persistent identifier to be associated with the data being transferred.

☞ Generally, identifiers are requested when importing new data into an infrastructure.

- **update catalogues (client)** is used to update (or initiate the update of) data catalogues used to describe the data held within an infrastructure to account for new datasets.

A data importer must also provide two stream interfaces through which to pass data:

- **deliver dataset (consumer)** is used to retrieve external datasets stored in external data stores outside of the data curation subsystem.
- **import data for curation (producer)** is used to deliver (repackaged) datasets to one or more data stores within the data curation subsystem.

data transfer service

Oversight service for the transfer of data into and out of the data curation subsystem.

A data transfer service object encapsulates the functions required to integrate new data into the data curation subsystem of a research infrastructure and export that integrated data on demand. The data transfer service is responsible for setting up data transfers, including any repackaging of datasets necessary prior to delivery.

- A data transfer object can create any number of new *data transporter* objects.

☞ The actual coordination of data transfers is handled by data transporter objects; the data transfer service is responsible for specifying the behaviour of a given transporter.

A data transfer service should provide one operational interface:

- **prepare data transfer [server]** provides functions for negotiating and scheduling a data transfer either into or out of the data curation subsystem of an infrastructure.

data transporter

Generic binding object for data transfer interactions.

A data transporter binding object encapsulates the coordination logic required to deliver data into and out of the data curation subsystem of a research infrastructure. A data transporter object is created whenever data is to be streamed from one locale to another.

A data transporter is configured based on the data transfer to be performed, but must have at least the following two interfaces:

- **prepare storage (client)** is used to inform downstream resources about impending data transfers.
- **retrieve data (client)** is used to request data from a given data source.

raw data collector

Binding object for raw data collection.

A sub-class of *data transporter* binding object encapsulating the functions required to move and package raw data collected by the data acquisition subsystem.



ENVRI Common Operations of Environmental Research Infrastructures

A raw data collector should provide at least two operational interfaces in addition to those provided by any data transporter:

- **acquire identifier (client)** is used to request a new persistent identifier to be associated with the data being transferred.

☞ Generally, identifiers are requested when importing new data into an infrastructure.

- **update catalogues (client)** is used to update (or initiate the update of) data catalogues used to describe the data held within an infrastructure to account for new datasets.

A raw data collector must also provide two stream interfaces through which to pass data:

- **deliver raw data (consumer)** is used to collect raw data sent by instruments in the data acquisition subsystem.
- **import data for curation (producer)** is used to deliver (repackaged) raw data to data stores in the data curation subsystem.

3.3.3 Data Access

Aside from the curation of scientific data, a research infrastructure must provide means to access that data. Access can be provided in a number of ways, including the export of curated datasets and the querying of data catalogues. Beyond the actual mechanism of access however are the issues of discovery and interpretation. Specific datasets may be found via citation (the publication of persistent identifiers associated with data) or by browsing data catalogues (permitting queries over multiple datasets); the capability should exist to allow an infrastructure to identify the location of specific datasets in data stores as well. It should also be possible to identify the ontologies, taxonomies and other semantic metadata associated with datasets or data requests and provide some form of mapping between representations as necessary.

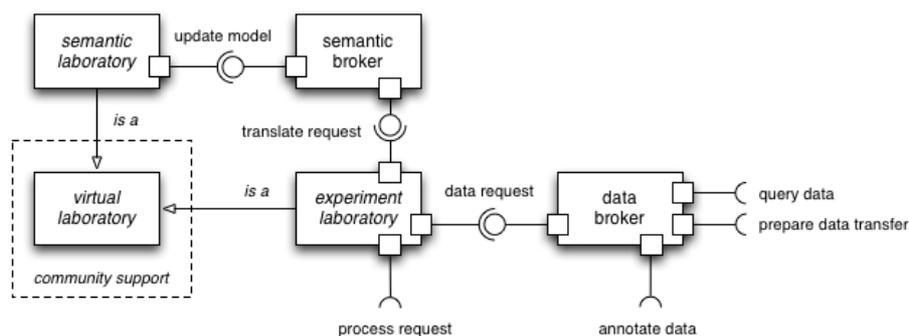


Figure 3.26: Computational Specification of Data Access Subsystem

The **data access** subsystem provides data brokers that act as intermediaries for access to data held within the data curation subsystem, as well as semantic brokers for performing semantic interpretation. These brokers are responsible for verifying the agents making access requests and for validating those requests prior to sending them on to the relevant data curation service. These brokers can be interacted with directly via *virtual laboratories* such as *experiment laboratories* (for general interaction with data and processing services) and *semantic laboratories* (by which the community can update semantic models associated with the research infrastructure).



ENVRI Common Operations of Environmental Research Infrastructures

data broker

Broker for facilitating data access/upload requests.

A data broker object intercedes between the data access subsystem and the data curation subsystem, collecting the computational functions required to negotiate data transfer and query requests directed at data curation services on behalf of some user. It is the responsibility of the data broker to validate all requests and to verify the identity and access privileges of agents making requests. It is not permitted for an outside agency or service to access the data stores within a research infrastructure by any means other than via a data broker.

☞ Data brokers are not responsible for brokering the collection of raw data from the data acquisition subsystem, as this is handled more efficiently by an acquisition service.

A data broker should provide four operational interfaces:

- **data request (server)** provides functions for requesting the import or export of datasets, the querying of data or the annotation of data within a research infrastructure.
- **annotate data (client)** is used to request annotation of data held within the data curation subsystem of a research infrastructure.
- **prepare data transfer (client)** is used to negotiate data transfers with the data curation subsystem of a research infrastructure.
- **query data (client)** is used to forward queries onto the data curation subsystem of a research infrastructure and receive the results.

experiment laboratory

Community proxy for conducting experiments within a research infrastructure.

A sub-class of virtual laboratory binding object encapsulating the functions required to schedule the processing of curated and user-provided data in order to perform some task (analysis, data mining, modelling, simulation, etc.).

☞ An experiment laboratory is created by a science gateway in order to allow researchers to interact with data held by a research infrastructure in order to achieve some scientific output.

An experiment laboratory should provide at least three operational interfaces:

- **data request (client)** is used to make requests of the research infrastructure pertaining to curated datasets.
- **process request (client)** is used to make requests of the research infrastructure pertaining to data processing.
- **translate request (client)** is used to invoke a semantic broker where some mapping between different semantic domains is deemed necessary.

semantic broker

Broker for establishing semantic links between concepts and bridging queries between semantic domains.



ENVRI Common Operations of Environmental Research Infrastructures

A semantic broker intercedes where queries within one semantic domain need to be translated into another to be able to interact with curated data. It also collects the functionality required to update the semantic models used by an infrastructure to describe data held within.

A semantic broker should provide two operational interfaces:

- **translate request (server)** provides functions for translating requests between two semantic domains.
- **update model (server)** provides functions for updating semantic models associated with a research infrastructure.

semantic laboratory

Community proxy for interacting with semantic models.

A sub-class of *virtual laboratory* binding object encapsulating the functions required to update semantic models (such as ontologies) used in the interpretation of curated data (and infrastructure metadata).

☞ A semantic laboratory is created by a science gateway in order to allow researchers to provide input on the interpretation of data gathered by a research infrastructure.

A semantic laboratory should provide at least one operational interface in addition to those provided by any virtual laboratory:

- **update model (client)** is used to update semantic models associated with a research infrastructure.

3.3.4 Data Processing

The processing of data can be tightly integrated into data handling systems, or can be delegated to a separate set of services invoked on demand; generally the more involved the processing, the more likely that separate resources will be required. The provision of dedicated processing services becomes significantly more important when large quantities of data are being curated within a research infrastructure, especially for scientific data, that is often subject to extensive post-processing and analysis in order to extract new results. The data processing subsystem of an infrastructure encapsulates the dedicated processing services made available to that infrastructure, either within the infrastructure itself or delegated to a client infrastructure.

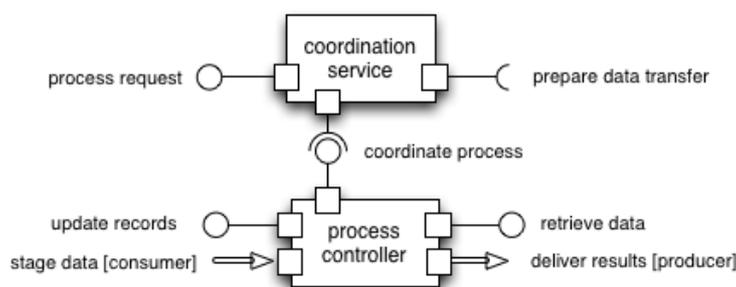


Figure 3.27: Computational Specification of Data Processing Subsystem

The **data processing** subsystem is computationally described as a set of *process controllers* (representing the computational functionality of registered execution resources) monitored



ENVRI Common Operations of Environmental Research Infrastructures

and managed by a *coordination service*. The coordination service delegates all processing tasks sent to the data processing subsystem to particular execution resources, coordinates multi-stage workflows and initiates execution. Data may need to be staged onto individual execution resources and results retrieved for curation; data channels can be established with resources via their process controllers.

coordination service

Oversight service for data processing tasks deployed on infrastructure execution resources.

A coordination service should provide at least three operational interfaces:

- **process request (server)** provides functions for scheduling the execution of data processing tasks; this potentially includes complex workflows involving many (parallel) sub-tasks.
- **coordinate process (client)** is used to coordinate the execution of data processing tasks on execution resources presented by process controllers.
- **prepare data transfer (client)** is used to move data into and out of the data curation subsystem in order to register new results or in preparation for the generation of such results.

process controller

Part of the execution platform provided by the data processing subsystem.

A process controller object encapsulates the functions required for using an execution resource (generically, any computing platform that can host some process) as part of any infrastructure workflow.

A process controller should provide at least three operational interfaces:

- **coordinate process (server)** provides functions for controlling the execution resource associated with a given process controller.
- **retrieve data (server)** provides functions for retrieving data from an execution resource.
- **update records (server)** provides functions for modifying data on an execution resource, including preparing the resource for the ingestion of bulk data delivered through its *stage data* stream interface.

A process controller should provide at least two operational interfaces:

- **stage data (consumer)** is used to acquire data sent from the data curation subsystem of a research infrastructure needed as part of some process.
- **deliver results (producer)** is used to deliver any new data produced for integration into the data curation subsystem of a research infrastructure.

3.3.5 Community Support

A research infrastructure cannot be considered an isolated entity; the modes of interaction between a research infrastructure and the broader scientific community that it serves must be accounted for in the design of the infrastructure. In the Reference Model it is assumed that the principal method by which prospective users interact with an infrastructure is via a *scientific gateway* or *virtual research environment* - in essence a community portal, usually web-based, by which a number of services are exposed both for human users and for remote procedure invocation. These services may range from fundamental (data discovery and retrieval) to more interactive (user contribution and dataset annotation) to more 'social' (concerning user profiling, reputation mechanisms and workflow sharing).



ENVRI Common Operations of Environmental Research Infrastructures

The community support subsystem also encapsulates services generally provided by outside agencies that are nevertheless invoked by services within the infrastructure proper, such as a service for providing globally-readable persistent identifiers to datasets.

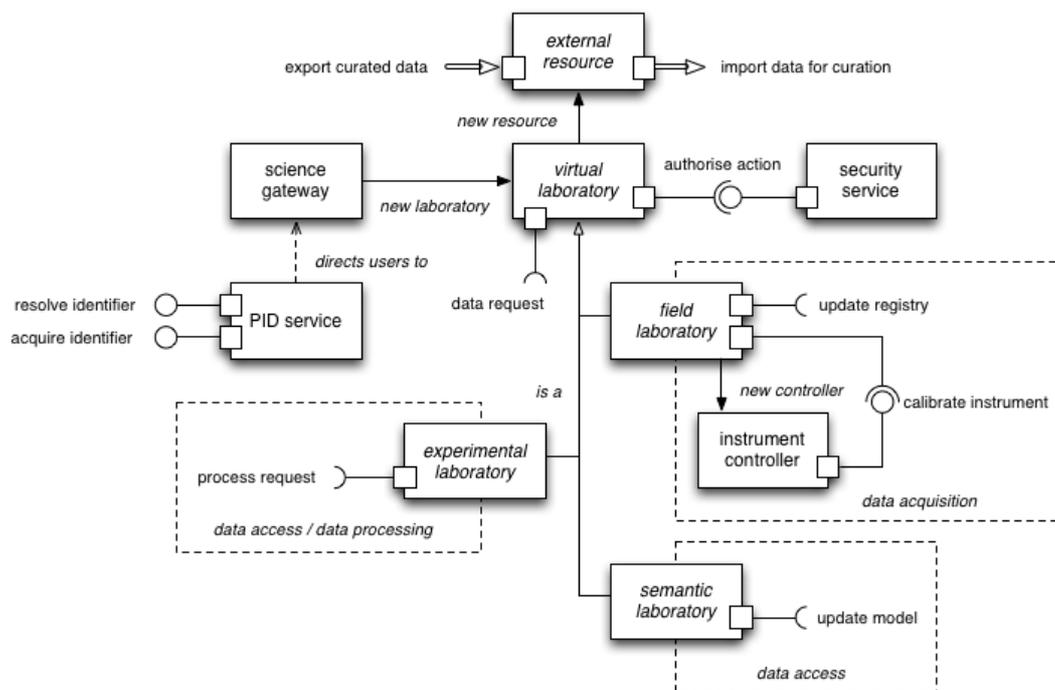


Figure 3.28: Computational Specification of Community Support Subsystem

In the Reference Model, more complex interactions between the community and other infrastructure subsystems are mediated by *virtual laboratories*; these objects are deployed by *science gateways* in order to provide a persistent context for such interactions between certain groups of users and particular infrastructure subsystems. The Reference Model recognises the following specific subclasses of laboratory:

- *Field laboratories* (so-named because they interact with raw data sources 'in the field') are used to interact with the *data acquisition* subsystem, allowing researchers to deploy, calibrate and un-deploy instruments as part of the integrated data acquisition network used by an infrastructure to collect its primary 'raw' data. Field laboratories have the ability to instantiate new *instrument controllers* within the data acquisition subsystem.
- *experimental laboratories* are used to interact both with curated data and data processing facilities, allowing researchers to deploy datasets for processing and acquire results from computational experimentation.
- *semantic laboratories* are used to interact with the semantic models used by a research infrastructure to interpret datasets and characteristic (meta)data.

Regardless of provenance, all laboratories must interact with a *security service* in order to authorise requests and authenticate users of the laboratory before they can proceed with any privileged activities.



ENVRI Common Operations of Environmental Research Infrastructures

A *PID service* provides globally-readable persistent identifiers to infrastructure entities (principally datasets, but also permissibly processes, services and data sources) that may be cited by the community. This service is assumed in the Model to be provided by an external party, and is expected to direct agents attempting to read citations to one of the infrastructure's science gateways.

PID service

External service for persistent identifier assignment and resolution.

Persistent identifiers are generated by a global service generally provided by an outside entity supported by the research community. A PID (persistent identifier) service object encapsulates this service and is responsible for providing identifiers for all entities that require them.

☞ Different versions of artefacts, where maintained separately, are assumed to have different identifiers, but those identifiers can share a common root such that the family of versions of a given artefact can be retrieved in one transaction, or only the most recent (or otherwise dominant) version is returned.

A PID service should provide at least two operational interfaces:

- **acquire identifier (server)** provides a persistent identifier for a given entity.
- **resolve identifier (server)** resolves identifiers, referring agents to the identified entity (in practice a science gateway providing access to the entity).

service gateway

Community portal for interacting with an infrastructure.

A science gateway object encapsulates the functions required to interact with a research infrastructure from outside the subsystems of data acquisition, data curation, data brokering and data processing. A science gateway should be able to provide virtual 'laboratories' for authorised agents to interact with and possibly configure many of the science functions of a research infrastructure.

- A science gateway object can instantiate any number of *virtual laboratory* objects.

security service

Oversight service for authentication and authorisation of user requests to the infrastructure.

A security service object encapsulates the functions required to authenticate agents and authorise any requests they make to services within a research infrastructure. Generally, any interaction occurring via a science gateway object or a virtual laboratory object will only proceed after a suitable transaction with a security service object has been made.

A security service should provide at least one operational interface:

- **authorise action (server)** provides functions to verify and validate proposed actions, providing authorisation tokens (for example) where required.

virtual laboratory

Community proxy for interacting with infrastructure subsystems.

A virtual laboratory binding object encapsulates interaction between a user or group of users and a subset of the science functions provided by a research infrastructure. Its role is to bind a security service with (potentially) any number of other infrastructure objects.

A virtual laboratory object must provide at least one interface:

- **authorise action** [client] is used to retrieve authorisation for any restricted interactions with the data acquisition subsystem.

Specific sub-classes of virtual laboratory should be defined to interact with the infrastructure in different ways. The Reference Model defines the *field laboratory* object for interaction with the *data acquisition* subsystem.

3.3.6 Brokered Data Export

Exporting data out of a research infrastructure entails retrieving data from the data curation subsystem and delivering it to an external resource. This process must be brokered by the community support and data access subsystems.

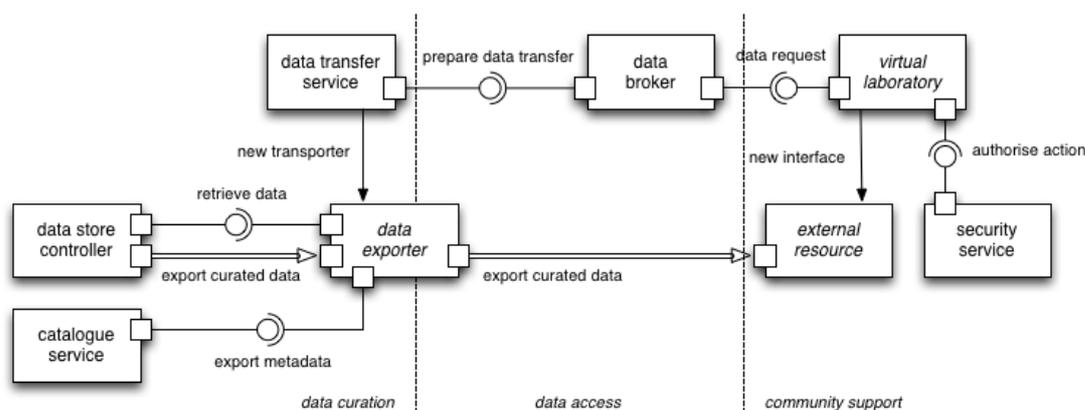


Figure 3.29: Brokered Data Export

Generally requests for data to be exported to an external resource originate from a *virtual laboratory*. All requests are validated by the *security service* via its *authorise action* interface. The laboratory provides an interface to an *external resource* (this might take the form of a URI and a preferred data transfer protocol) and submits a request to a *data broker* in the data access subsystem via its *data request* interface. The data broker will translate any valid requests into actions; in this scenario, a data transfer request is sent to the *data transfer service* within the data curation subsystem.

The data transfer service will configure and deploy a *data exporter*; this exporter will *retrieve data* from all necessary data stores, opening a data-flow from data store to external resource. The exporter is also responsible for the repackaging of exported datasets where necessary – this includes the integration of any additional metadata or provenance information stored separately within the infrastructure that needs to be packaged with a dataset if it is to be used independently of the infrastructure. As such, the exporter can invoke the *catalogue service* to retrieve additional meta-information via its *export metadata* interface.

3.3.7 Brokered Data Import

Importing data from sources other than the acquisition network requires that the import be brokered by the data access subsystem before data can be delivered into the data curation subsystem.

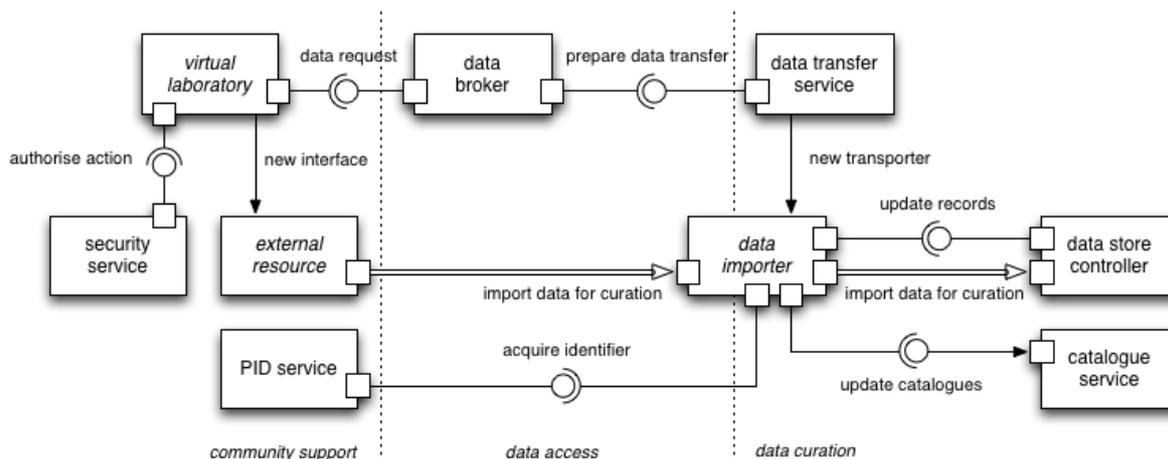


Figure 3.30: Brokered Data Import

A *virtual laboratory* can be used by researchers to upload new data into a research infrastructure. All requests are validated by the *security service* via its *authorise action* interface. The laboratory provides an interface to an *external resource* (this might take the form of a URI and a preferred data transfer protocol) and submits a request to a *data broker* in the data access subsystem via its *data request* interface. The data broker will translate any valid requests into actions; in this scenario, a data transfer request is sent to the *data transfer service* within the data curation subsystem.

The data transfer service will configure and deploy a *data importer*; this importer will open a data-flow from an external resource to one or more suitable data stores within the infrastructure and *update records* within those stores as appropriate. The importer is responsible for the annotation and registration of imported datasets – this generally entails obtaining a global persistent identifier for any new datasets and updating the catalogues used by the research infrastructure to identify and sort its data inventory. As such, the importer can invoke the *catalogue service* to *update catalogues* and invoke any community-used *PID service* to *acquire identifiers*.

3.3.8 Brokered Data Query

Querying curated data resources requires that the request be brokered by the data access subsystem before any results will be retrieved from the data curation subsystem and delivered to the client from which the source came.

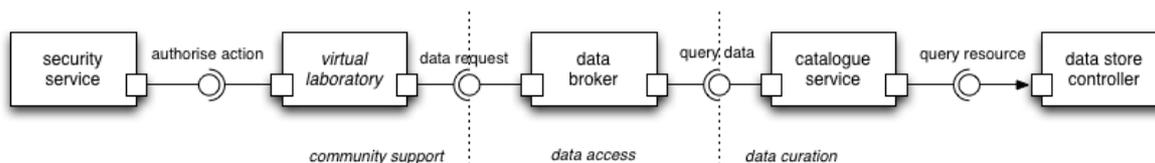


Figure 3.31: Brokered Data Query

Any kind of *virtual laboratory* is able to query the data held within a research infrastructure subject to access privileges governed by the *security service* (invoked via its *authorise action* interface). Data requests are forwarded to a *data broker* within the data access subsystem, which will interpret the request and contact any internal services needed to fulfil it. In this case, the data broker will invoke

the *catalogue service* via its *query data* interface; the catalogue service will locate the datasets needed to answer any given query and then proceed to *query resources* within infrastructure *data stores*.

3.3.9 Citation

The citation of datasets involves reference to persistent identifiers assigned to objects within a research infrastructure. Such citations are resolved by referring back to the infrastructure, which can then return a report describing the data cited.



Figure 3.32: Citation

A user or external service tries to *resolve an identifier* (found in a citation) with the global *PID service* used by the research infrastructure. By dereferencing the given identifier, that user or service is directed to a *science gateway* used to interact with the infrastructure. From there, the desired provenance information about the citation can be immediately retrieved, or a *virtual laboratory* can be deployed for more complex interactions with the research infrastructure.

3.3.10 Internal Data Staging

The internal staging of data within an infrastructure for processing requires coordination between the data processing subsystem (which handles the actual processing workflow) and the data curation subsystem (which holds all scientific datasets within the infrastructure).

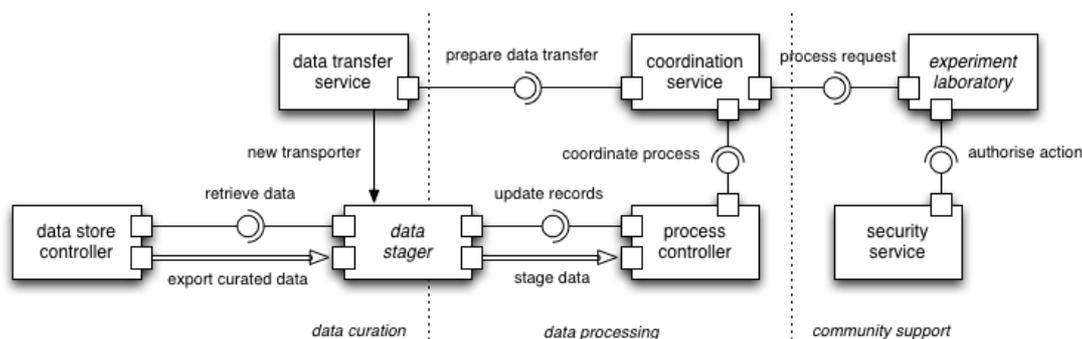


Figure 3.33: Internal Data Staging

Data processing requests generally originate from *experiment laboratories* where authorised by a *security service*. A *process request* is sent to a *coordination service*, which interprets the request and coordinates any processing workflow. In the case where data needs to be staged onto an execution platform from a curated data store, the coordination service will request that a *data transfer service* within the data curation subsystem *prepare a data transfer*. The data transfer service will then configure and deploy a *data stager*.

The data stager coordinates the transfer of data between curation and processing subsystems. A data-flow is established between all required *data stores* and all *process controllers* (representing in this case the execution platform on which data processing will be executed) by requesting data via the data



stores' *retrieve data* interfaces and *updating records* on the process controllers' side (to ensure that a destination for the data has been prepared and to ensure that all data is tracked correctly).

3.3.11 Processed Data Import

The formal ingestion of data derived during internal processing of existing datasets within an infrastructure requires coordination between the data processing subsystem (which produces the new results) and the data curation subsystem (which is responsible for all scientific datasets within the infrastructure).

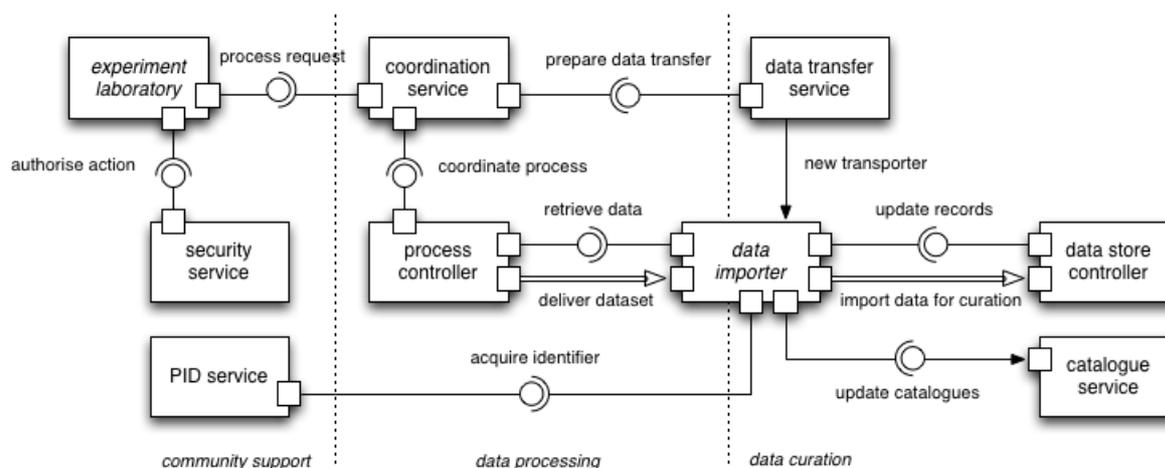


Figure 3.34: Processed Data Import

Requests for data processing typically originate from an *experiment laboratory*, mediated by a *security service* (via its *authorise action* interface). Process requests are sent to a *coordination service* that will interpret requests and coordinate any internal resources needed accordingly. In the case of derived data being created that should be reintegrated into the research infrastructure, the coordination service will invoke a *data transfer service* via its *prepare data transfer* interface. That data transfer service will configure and deploy a *data importer* to coordinate the transfer of results back into the data curation subsystem.

The data importer *retrieves data* from any *process controllers* producing new data and establishes a data-flow from the execution platform underlying those controllers to suitable *data stores* needed to host the data. The *update records* interface is used to prepare data store controllers and ensure that the ingested data is properly recorded. The data importer is responsible for ensuring that ingested data is packaged correctly; it may also be necessary for a new persistent identifier to be obtained and for research infrastructure data catalogues to be updated. These responsibilities are handled by the *PID service* (via its *acquire identifier* interface) and the *catalogue service* (via its *update catalogues* interface) respectively.

3.3.12 Raw Data Collection

The collection of raw scientific data requires coordination between the *data acquisition* subsystem (which extracts the raw data from instruments) and the *data curation* subsystem (which packages and stores the data).

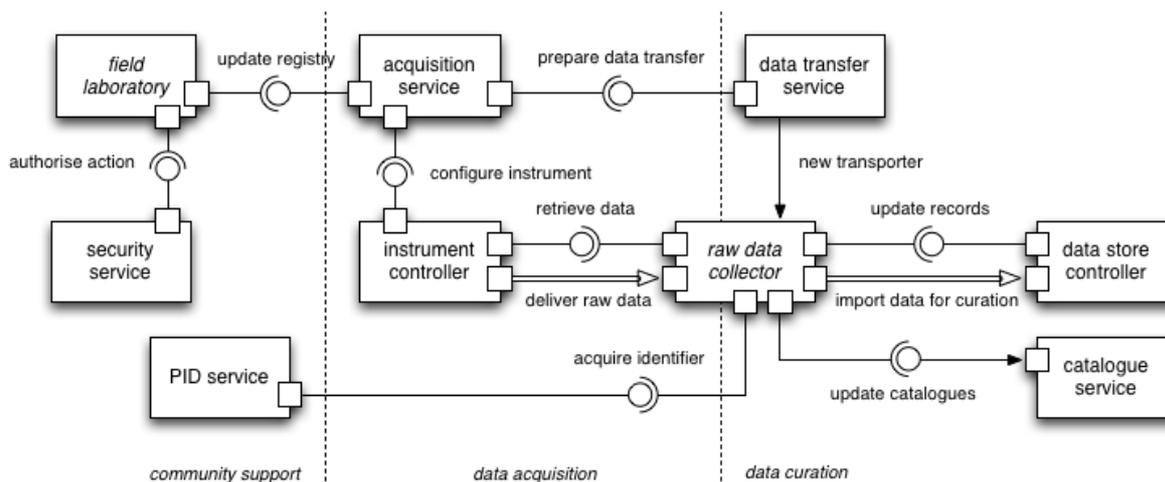


Figure 3.35: Raw Data Collection

The delivery of raw data into a research infrastructure is driven by collaboration between an *acquisition service* and a *data transfer service*. This process can be configured using a *field laboratory* subject to a *security service* (actions are validated via the security service's *authorise action* interface). Regardless, the acquisition service identifies the instruments that act as data sources and provides information on their output behaviour, whilst the data transfer service provides a *data transporter* that can establish (multiple, persistent) data channels between instruments and data stores. The data transporter (a *raw data collector*) can initiate data transfer by requesting data from one or more *instrument controllers* and preparing one or more *data store controllers* to receive the data.

The raw data collector is considered responsible for packaging any raw data obtained into a format suitable for curation - this may entail chunking data streams, assigning persistent identifiers and associating metadata to the resulting datasets. To assist in this, a raw data collector may acquire identifiers from a *PID service*. It may also want to register the presence of new data and any immediately apparent data characteristics in infrastructure data catalogues - this is done by invoking an update operation on the *catalogue service*.

3.3.13 Instrument Integration

Data acquisition relies on an integrated network of data sources (referred to generically as 'instruments') that provide raw measurements and observations continuously or on demand. This network is not necessarily static; new instruments can be deployed and existing instruments can be taken off-line or re-calibrated throughout the lifespan of a research infrastructure. In the Reference Model, modifications to the acquisition network should be performed via a 'virtual laboratory' that permits authorised agents to oversee acquisition and calibrate instruments based on current community practice or environmental conditions.

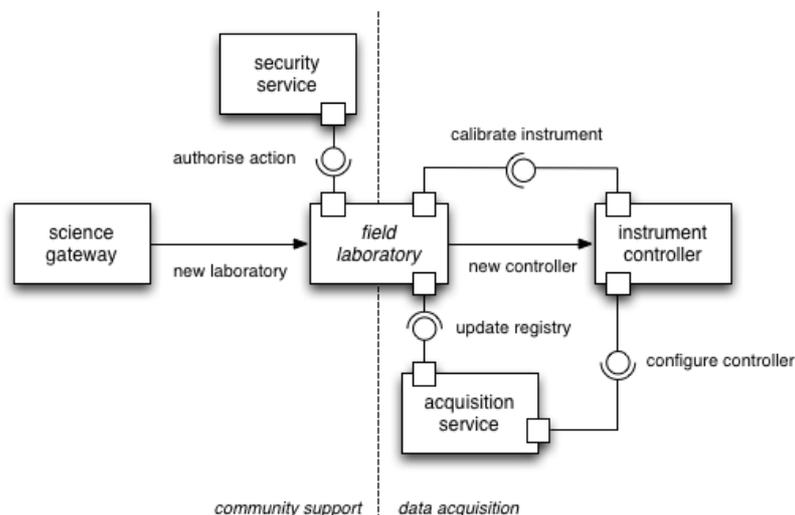


Figure 3.36: Instrument Integration

Instruments can be added to and removed from a data acquisition network by a *field laboratory* provided by a *science gateway*. The field laboratory must be able to provide an *instrument controller* for any new instrument added in order to allow the data acquisition subsystem to interact with the instrument. Deployment, un-deployment or re-calibration of instruments requires authorisation - this can only be provided a valid *security service* (via its *authorise action* interface). Any changes to the data acquisition network must be registered with an *acquisition service* (via its *update registry* interface).

The behaviour of an instrument controller can be configured by the acquisition service by invoking functions on the controller via its *configure controller* interface.

A field laboratory also provides the means to calibrate instruments based on scientific best practice where applicable - this is done via the instrument controller's *calibrate instrument* interface.



4 CONCLUSION AND FUTURE WORK

The ENVRI Reference Model is a work in progress. Currently, attention is focused on three of the five ODP viewpoints: enterprise, information and computational. The remaining viewpoints of engineering and technology have been deferred to a later date.

Much work remains. Stronger correspondence between the three primary viewpoints is necessary to ensure that the three sub-models are synchronised in concept and execution. Further refactoring of individual components and further development of individual elements is to be expected as well. Further development of the presentation of the model is also essential, in order to both improve clarity to readers not expert in ODP and in order to promote a coherent position. In the immediate next step, the following tasks are planned:

Validation

The reference model will be validated from several aspects.

1. Usability. The users from different RIs will be invited to use the reference model to describe the research infrastructures in the ENVRI. The feedback will be collected and analysed to improve the definition of the reference model.
2. Interoperability. The descriptions of different RIs will be compared and check the commonality of the operations, and validate the effectiveness of the reference model in realizing the interoperability between RIs. The development of the use case in the work package 4 will also be used as the scenario to test the reference model.
3. Application. The linking model and the reference model will be tested in the application planning systems to check the data, resource and infrastructure interoperability

Semantic linking model

The reference model will be used as an important input for task 3.4, namely the development of semantic linking model among the reference model, data and infrastructure. The linking model provides an information framework to glue different information models of resources and data. It couples the semantic descriptions of the data with the infrastructures and provides semantic interoperability between data and resources. It needs to address fault tolerance, optimization and scheduling of linked resources, while making a trade-off between fuzzy logic and full information. The linking model is part of the development effort of the reference model.

The linking model will take different aspects into considerations:

- The application (such as workflow) aspect captures the main characteristics of the application supported by the research infrastructure, including issues such as main flow patterns, quality of services, security and policies in user communities, and linking them to the descriptions of the data and infrastructures.
- The computing and data aspect focuses on operations and different data and meta data standards at different phase of data evolution (raw data, transfer, calibration, fusion etc.) and model them with linking of the data storing, accessing, delivery and etc. on (virtualized) e-Infrastructure.
- The Infrastructure aspect links the semantic model of the different layers of components in the physical infrastructure such as network elements and topologies, and also the monitoring information of the runtime status of the infrastructure. This part will enable the constraint solving of quality constraints to reserve and allocating resources for high level applications (processes).



5 REFERENCES

- [1] W. Los, "Introduction to ENVRI and the workshop objectives," in *ENVRI Frascati Meeting 5-7 Feb 2013*, Presentation. Frascati, Italy, 2013.
- [2] "Global Change: Towards global research infrastructures," *European Commission, Directorate-General For Research and Innovation*, 2012.
- [3] S. Sorvari. "Environmental research in harmony," *International Innovation - Disseminating, science research and technology*. Dec. 2012 Page 28, 2012. Available: <http://www.research-europe.com/magazine/ENVIRONMENT/2012-15/index.html>
- [4] ISO/IEC, "ISO/IEC 10746-1: Information technology--Open Distributed Processing--Reference model: Overview," *ISO/IEC Standard*, 1998.
- [5] ISO/IEC, "ISO/IEC 10746-2: Information technology--Open Distributed Processing--Reference model: Foundations," *ISO/IEC Standard*, 2009.
- [6] ISO/IEC, "ISO/IEC 10746-3: Information technology--Open Distributed Processing--Reference model: Architecture," *ISO/IEC Standard*, 2009.
- [7] ISO/IEC, "ISO/IEC 10746-4: Information technology--Open Distributed Processing--Reference model: Architecture Semantics," *ISO/IEC Standard*, 1998.
- [8] OASIS, "Reference Model for Service Oriented Architecture 1.0," *OASIS Standard*, 2006.
- [9] L. Candela, and A. Nardi, "The Digital Library Reference Model," *DL.org*, 2010.
- [10] ISO/IEC, "Open System Interconnection (OSI), ISO/IEC 7498-1," *ISO/IEC Standard*, 1994.
- [11] CCSDS, "Reference Model for an Open Archival Information System (OAIS)," *CCSDS Standard*, 2012.
- [12] C. Atkinson, M. Gutheil, and K. Kiko, "On the Relationship of Ontologies and Models," *Lecture Notes in Informatics, Gesellschaft für Informatik, Bonn*, INI Proceedings, 1996.
- [13] D. C. Schmidt, "Model-Driven Engineering," *IEEE Computer* vol. 39, 2006.
- [14] N. F. Noy, and D. L. McGuinness, "Ontology Development 101: A Guide to Creating Your First Ontology," *Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880*, 2001.
- [15] P. Tetlow, J. Z. Pan, D. Oberle, E. Wallace, M. Uschold, and E. Kendall, "Ontology Driven Architectures and Potential Uses of the Semantic Web in Systems and Software Engineering," *W3C Standard*, 2006.
- [16] SEKE, "International Conference on Software Engineering (SEKE 2005)".
- [17] VORTE, "International Workshop on Vocabularies, Ontologies and Rules for The Enterprise (VORTE 2005-2013)".
- [18] MDSW, "The Model-Driven Semantic Web Workshop (MDSW 2004)".
- [19] SWESE, "Workshop on Semantic Web Enabled Software Engineering (SWESE 2005-2007)".
- [20] ONTOSE, "Workshop on Ontology, Conceptualizations and Epistemology of Software and Systems Engineering (ONTOSE 2005-2009)".
- [21] WoMM, "Workshop on Meta-Modeling and Corresponding Tools (WoMM 2005)".



ENVRI Common Operations of Environmental Research Infrastructures

- [22] Kwaaitaal, M. Hoogeveen, and T. V. D. Weide, "A Reference Model for the Impact of Standardisation on Multimedia Database Management Systems," *Computer Standards & Interfaces*, vol. 16, pp. 45-54, 1994.
- [23] OGC, "OGC Reference Model," *Open Geospatial Consortium*, OGC Standard, 2011.
- [24] T. Uslander, "Reference Model for the ORCHESTRA Architecture (RM-OA) V2," *Open Geospatial Consortium*, OGC Standard, 2007.
- [25] V. Hernandez-Ernst, *et al.*, "LIFEWATCH. Deliverable 5.1.3: Data & Modelling Tool Structures -- Reference Model," *the EU LifeWatch consortium*, 2010.
- [26] D. Hollingsworth, "The Workflow Reference Model," *the Workflow Management Coalition*, 1995.
- [27] Mayk, and W. C. Regli, "Agent Systems Reference Model Release Version 1.0a," *US Army Communications and Electronics Command Research Development and Engineering Center (CERDEC)*, 2006.
- [28] E. H. Chi, and J. T. Riedl, "An Operator Interaction Framework for Visualization Systems," *Symposium on Information Visualization (InfoVis '98)*, 1998.
- [29] E. H. Chi, "A Taxonomy of Visualisation Techniques using the Data State Reference Model," *Proceedings of the IEEE Symposium on Information Visualization 2000 (InfoVis'00)*, 2000.
- [30] N. Koch, and M. Wirsing, "The Munich Reference Model for Adaptive Hypermedia Applications," in *2nd International Conference on Adaptive Hypermedia and Adaptive Web-based Systems, Proceedings. P. De Bra, P. Brusilovsky, and R. Conejo (eds.) LNCS 2347*, ©Springer Verlag, pp. 213-222, 2002.
- [31] OMG, "Data Distribution Service for Real-time Systems Version 1.2," *OMG Standard*, 2007.
- [32] OASIS, "Content Management Interoperability Services (CMIS) Version 1.0," *OASIS Standard*, 2011.
- [33] Hardisty, "WP3 Progress and Issues for Full Plenary with SAB, Wednesday 6th Feb 2013," in *ENVRI Frascati Meeting, 5-7 Feb 2013*, ed. Frascati, Italy, 2013
- [34] R. Kahn and R. Wilensky, "A framework for distributed digital object services", *International Journal on Digital Libraries (2006) 6(2):115-123*, 2006.
- [35] Barros, M. Dumas and P. Oaks, "A Critical Overview of the Web Services Choreography Description Language (WS_CDL)", *BPTrends*, Mar. 2005.
- [36] L. Candela. "Data Use - Virtual Research Environments". In K. Ashley, C. Bizer, L. Candela, D. Fergusson, A. Gionis, M. Heikkurinen, E. Laure, D. Lopez, C. Meghini, P. Pagano, M. Parsons, S. Viglas, D. Vitlacil, and G. Weikum, (ed.) *Technological & Organisational Aspects of a Global Research Data Infrastructure - A view from experts, GRDI2020, 91-98*, 2012.
- [37] P. F. Linington, Z. Milosevic, A. Tanaka, and A. Vallecillo, Ed., *Building Enterprise Systems with ODP*. CRC Press, 2012.



APPENDIXES

A. Terminology and Glossary

A.1 Acronyms and Abbreviations

CCSDS	Consultative Committee for Space Data Systems
CMIS	Content Management Interoperability Services
CERIF	Common European Research Information Format
DDS	Data Distribution Service for Real-Time Systems
ENVRI	Environmental Research Infrastructure
ENVRI_RM	ENVRI Reference Model
ESFRI	European Strategy Forum on Research Infrastructures
ESFRI-ENV RI	ESFRI Environmental Research Infrastructure
GIS	Geographic Information System
IEC	International Electrotechnical Commission
ISO	International Organisation for Standardization
OAIS	Open Archival Information System
OASIS	Advancing Open standards for the Information Society
ODP	Open Distributed Processing
OGC	Open Geospatial Consortium
OMG	Object Management Group
ORCHESTRA	Open Architecture and Spatial Data Infrastructure for Risk Management
ORM	OGC Reference Model
OSI	Open Systems Interconnection
OWL	Web Ontology language
SOA	Service Oriented Architecture
SOA-RM	Reference Model for Service Oriented Architecture
RDF	Resource Description Framework
RM-OA	Reference Model for the ORCHESTRA Architecture
RM-ODP	Reference Model of Open Distributed Processing
UML	Unified Modelling Language
W3C	World Wide Web Consortium
UML4ODP	Unified Modelling Language For Open Distributed Processing



ENVRI Common Operations of Environmental Research Infrastructures

A.2 Terminology

Access Control: A functionality that approves or disapproves of access requests based on specified access policies.

Acquisition Service: Oversight service for integrated data acquisition.

Active role: A active role is typically associated with a human actor.

Add Metadata: Add additional information according to a predefined schema (metadata schema). This partially overlaps with data annotations.

Annotate Data: Annotate data with meaning (concepts of predefined local or global conceptual models).

Annotate Metadata: Link metadata with meaning (concepts of predefined local or global conceptual models). This can be done by adding a pointer to concepts within a conceptual model to the data. If e.g. concepts are terms in and SKOS/RDF thesaurus, published as linked data then this would mean entering the URL of the term describing the meaning of the data.

Annotation Service: Oversight service for adding and updating records attached to curated datasets.

Assign Unique Identifier: Obtain a unique identifier and associate it to the data.

Authentication: A functionality that verifies a credential of a user.

Authentication Service: Security service responsible for the authentication of external agents making requests of infrastructure services.

Authorisation: A functionality that specifies access rights to resources.

Authorisation Service: Security service responsible for the authorisation of all requests made of infrastructure services by external agents.

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

Behaviour : A behaviour of a community is a composition of actions performed by roles normally addressing separate business requirements.

Build Conceptual Models: Establish a local or global model of interrelated concepts.

Capacity Manager: An active role, which is a person who manage and ensure that the IT capacity meets current and future business requirements in a cost-effective manner.

Carry out Backup: Replicate data to an additional data storage so it may be used to restore the original after a data loss event. A special type of backup is a long term preservation.

Catalogue Service: Oversight service for cataloguing curated datasets.

Check Quality: Actions to verify the quality of data.



ENVRI Common Operations of Environmental Research Infrastructures

Citation: Citation in the sense of IT is a pointer from published data to:

- the data source(s)
- and / or 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

Community: A collaboration which consists of a set of *roles* agreeing their objective to achieve a stated business purpose.

Community Support Subsystem: A subsystem that provides functionalities to manage, control, and track users' activities and supports users to conduct their roles in the community.

Concept: 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. Concepts within ontologies and their relations can be seen as machine readable sentences. Those sentences can be used to establish a self-description. It is, however, practice today, to have both, the human readable description and the machine readable description. In this sense a conceptual model can also be seen as a collection of human and machine readable sentences. Conceptual models can reside within the persistence layer of a data provider or a community or outside. Conceptual models can be fused with the data (e.g. within a network of triple stores) or kept separately.

Coordination Service: An oversight service for data processing tasks deployed on infrastructure execution resources.

Data Access Subsystem: A subsystem that enables discovery and retrieval of data housed in data resources.

Data Acquisition Community. A community, which collects raw data and bring (streams of) measures into a system.

Data Acquisition Subsystem: A subsystem that collects raw data and brings the measures or data streams into a computational system.

Data Analysis: A functionality that inspects, cleans, transforms data, and provides data models with the goal of highlighting useful information, suggesting conclusions, and supporting decision making.

Data Assimilation: A functionality that combines observational data with output from a numerical model to produce an optimal estimate of the evolving state of the system.

Data Broker: Broker for facilitating data access/upload requests.

Data Cataloguing: A functionality that associates a data object with one or more metadata objects which contain data descriptions.

Data Citation: A functionality that assigns an accurate, consistent and standardised reference to a data object, which can be cited in scientific publications.

Data Collection: A functionality that obtains digital values from a sensor instrument, associating consistent timestamps and necessary metadata.

Data Collector: An active role, which is a person who prepares and collects data. The purpose of data collection is to obtain information to keep on record, to make decisions about important issues, or to pass information on to others.

Data Consumer: Either an active or a passive role, which is an entity who receives and use the data.

Data Curation Community: A community, which curates the scientific data, maintains and archives them, and produces various data products with metadata.

Data Curation Subsystem: A subsystem that facilitates quality control and preservation of scientific data.

Data Curator: An active role, which is a person who verifies the quality of the data, preserve and maintain the data as a resource, and prepares various required data products.

Data Discovery & Access: A functionality that retrieves requested data from a data resource by using suitable search technology.

Data Exporter: Binding object for exporting curated datasets.

Data Extraction: A functionality that retrieves data out of (unstructured) data sources, including web pages, emails, documents, PDFs, scanned text, mainframe reports, and spool files.

Data Identification: A functionality that assigns (global) unique identifiers to data contents.

Data Importer: An Oversight service for the import of new data into the data curation subsystem.

Data Mining: A functionality that supports the discovery of patterns in large data sets.

Data Originator: Either an active or a passive role, which provide the digital material to be made available for public access.

Data Processing Control: A functionality that initiates the calculation and manages the outputs to be returned to the client.

Data Processing Subsystem: A subsystem that aggregates the data from various resources and provides computational capabilities and capacities for conducting data analysis and scientific experiments.

Data Product Generation: A functionality that processes data against requirement specifications and standardised formats and descriptions.

Data Provenance: Information that traces the origins of data and records all state changes of data during their lifecycle and their movements between storages.

Data Provider: Either an active or a passive role, which is an entity providing the data to be used.

Data Publication: A functionality that provides clean, well-annotated, anonymity-preserving datasets in a suitable format, and by following specified data-publication and sharing policies to make the datasets publically accessible or to those who agree to certain conditions of use, and to individuals who meet certain professional criteria.

Data Publication Community: A community that assists the data publication, discovery and access.

(Data Publication) Repository: A passive role, which is a facility for the deposition of published data.

Data Quality Checking: A functionality that detects and corrects (or removes) corrupt, inconsistent or inaccurate records from data sets.

Data Service Provision Community: A community that provides various services, applications and software/tools to link, and recombine data and information in order to derive knowledge.

Data State: Term used as defined in ISO/IEC 10746-2. At a given instant in time, data state is the condition of an object that determines the set of all sequences of actions (or traces) in which the object can participate.

Data Storage & Preservation: A functionality that deposits (over long-term) the data and metadata or other supplementary data and methods according to specified policies, and makes them accessible on request.

Data Store Controller: A data store within the data curation subsystem.

Data Transfer Service: Oversight service for the transfer of data into and out of the data curation subsystem.

Data Transmission: A functionality that transfers data over communication channel using specified network protocols.

Data Transporter: Generic binding object for data transfer interactions.

Data Usage Community: A community who makes use of the data and service products, and transfers the knowledge into understanding.

Describe Service: Describe the accessibility of a service or process, which is available for reuse, the interfaces, the description of behaviour and/or implemented algorithms.

Design of Measurement Model: A behaviour that designs the measurement or monitoring model based on scientific requirements.

Do Data Mining: Execute a sequence of metadata / data request --> interpret result --> do a new request

Education or Trainee: An active role, a person, who makes use of the data and application services for education and training purposes.



ENVRI Common Operations of Environmental Research Infrastructures

Environmental Scientist: An active role, which is a person who conduct research or perform investigation for the purpose of identifying, abating, or eliminating sources of pollutants or hazards that affect either the environment or the health of the population. Using knowledge of various scientific disciplines, may collect, synthesize, study, report, and recommend action based on data derived from measurements or observations of air, food, soil, water, and other sources.

ENVRI Reference Model: A common ontological framework and standards for the description and characterisation of computational and storage systems of ESFRI environmental research infrastructures.

Experiment Laboratory: Community proxy for conducting experiments within a research infrastructure.

Field Laboratory: Community proxy for interacting with data acquisition instruments.

Final review: Review the data to be published, which will not likely be changed again.

General Public, Media or Citizen (Scientist): An active role, a person, who is interested in understanding the knowledge delivered by an environmental science research infrastructure, or discovering and exploring the knowledgebase enabled by the research infrastructure.

Instrument Controller: An integrated raw data source.

Mapping Rule: Configuration directives used for model-to-model transformation.

(Measurement Model) Designer: An active role, which is a person who design the measurements and monitoring models based on the requirements of environmental scientists.

Measurement Result: Quantitative determinations of magnitude, dimension and uncertainty to the outputs of observation instruments, sensors (including human observers) and sensor networks.

Measurer: An active role, which is a person who determines the ratio of a physical quantity, such as a length, time, temperature etc., to a unit of measurement, such as the meter, second or degree Celsius.

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

Metadata Catalogue: A collection of metadata, usually established to make the metadata available to a community. A metadata catalogue has an access service.

Metadata Harvesting: A functionality that (regularly) collects metadata (in agreed formats) from different sources.

Metadata State

- raw: are established metadata, which are not yet registered. In general, they are not shareable in this status
- registered: are metadata which are inserted into a metadata catalogue.
- published: are metadata made available to the public, the outside world. Within some metadata catalogues registered.

Passive Role: A passive role is typically associated with a non-human actor.

Perform Mapping: Execute transformation rules for values (mapping from one unit to another unit) or translation rules for concepts (translating the meaning from one conceptual model to another conceptual model, e.g. translating code lists).

Persistent Data: Term (data) used as defined in ISO/IEC 10746-2. Data is the representations of information dealt by information systems and users thereof. Data which are persistent (stored).

Perform Measurement or Observation: Measure parameter(s) or observe an event. The performance of a measurement or observation produces measurement results.

PID Generator: A passive role, a system which assigns persistent global unique identifiers to a (set of) digital object.

PID Registry: A passive role, which is an information system for registering PIDs.

PID Service: External service for persistent identifier assignment and resolution.

Policy or Decision Maker: An active role, a person, who makes decisions based on the data evidences.

Private Sector (Industry investor or consultant): An active role, a person, who makes use of the data and application service for predicting market so as to make business decision on producing related commercial products.

Process Control: A functionality that receives input status, applies a set of logic statements or control algorithms, and generates a set of analogue / digital outputs to change the logic states of devices.

Process Controller: Part of the execution platform provided by the data processing subsystem.

Process Data: Process data for the purposes of:

- converting and generating data products
- calculations: e.g., statistical processes, simulation models
- visualisation: e.g., alpha-numerically, graphically, geographically

Data processes should be recorded as provenance.

Provenance: The pathway of data generation from raw data to the actual state of data.

Publish Data: Make data public accessible.

Publish Metadata: Make the registered metadata available to the public.

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.

Quality Assessment (QA): Assessment of details of the data generation, including the check of the plausibility of the data. Usually the quality assessment is done by predefined checks on data and their generation process.



ENVRI Common Operations of Environmental Research Infrastructures

Query Data: Send a request to a data store to retrieve required data.

Query Metadata: Send a request to metadata resources to retrieve metadata of interests.

Observer: An active role, which is a person who receives knowledge of the outside world through the senses, or records data using scientific instruments.

Raw Data Collector: Binding object for raw data collection.

Reference Mode: A reference mode is an abstract framework for understanding significant relationships among the entities of some environment.

Register Metadata: Enter the metadata into a metadata catalogue.

Resource Registration: A functionality that creates an entry in a resource registry and inserts resource object or a reference to a resource object in specified representations and semantics.

Role : A role in a community is a prescribing behaviour that can be performed any number of times concurrently or successively.

Science Gateway: Community portal for interacting with an infrastructure.

Scientific Modelling and Simulation: A functionality that supports the generation of abstract, conceptual, graphical or mathematical models, and to run an instance of the model.

Scientist or Researcher: An active role, a person, who makes use of the data and application services to conduct scientific research.

(Scientific) Workflow Enactment: A specialisation of Workflow Enactment, which support of composition and execution a series of computational or data manipulation steps, or a workflow, in a scientific application. Important processes should be recorded for provenance purposes.

Security Service: Oversight service for authentication and authorisation of user requests to the infrastructure.

Semantic Broker: Broker for establishing semantic links between concepts and bridging queries between semantic domains.

Semantic Laboratory: Community proxy for interacting with semantic models.

Semantic Annotation: link from a thing (single datum, data set, data container) to a concept within a conceptual model, enabling the discovery of the meaning of the thing by human and machines.

Semantic Mediator: A passive role, which is a system or middleware facilitating semantic mapping discovery and integration of heterogeneous data.

Sensor: A passive role, which is a converter that measures a physical quantity and converts it into a signal which can be read by an observer or by an (electronic) instrument.



ENVRI Common Operations of Environmental Research Infrastructures

Sensor Network: A passive role, which is a network consists of distributed autonomous sensors to monitor physical or environmental conditions.

Service: Service or process, available for reuse.

Service Consumer: Either an active or a passive role, which is an entity using the services provided.

Service Description: Services and processes, which are available for reuse, be it within an enterprise architecture, within a research infrastructure or within an open network like the Internet, shall be described to help avoid wrong usage. 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)

Service Provider: Either an active or a passive role, which is an entity providing the services to be used.

Service Registry: A passive role, which is an information system for registering services.

Setup Mapping Rules: Specify the mapping rules of data and/or concepts.

Specification of Investigation Design: This is the background information 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 information for the allowed evaluations of data. (E.g. the question whether a sampling design was done randomly or by strategy determines which statistical methods that can be applied or not).

Specification of Measurements or Observations: The description of the scientific measurement model which specifies:

- what is measured;
- how it is measured;
- by whom it is measured; and
- what the temporal design is (single /multiple measurements / interval of measurement etc.)

Specify Investigation Design: specify design of investigation, including sampling design:

- geographical position of measurement or observation (site) -- the selections of observations and measurement sites, e.g., can be statistical or stratified by domain knowledge;
- characteristics of site;
- preconditions of measurements.

Specify Measurement or Observation: Specify the details of the method of observations/measurements.

Storage: A passive role, which is memory, components, devices and media that retain digital computer data used for computing for some interval of time.



ENVRI Common Operations of Environmental Research Infrastructures

Storage Administrator: An active role, which is a person who has the responsibilities to the design of data storage, tune queries, perform backup and recovery operations, raid mirrored arrays, making sure drive space is available for the network.

Store Data: Archive or preserve data in persistent manner to ensure continuing accessible and usable.

Subsystem: A subsystem is a set of capabilities that collectively are defined by a set of interfaces with corresponding operations that can be invoked by other subsystems. Subsystems are disjoint from each other.

Technician: An active role, which is a person who develop and deploy the sensor instruments, establishing and testing the sensor network, operating, maintaining, monitoring and repairing the observatory hardware.

Technologist or Engineer: An active role, a person, who develop and maintains the research infrastructure.

Track Provenance: Add information about the actions and the data state changes as data provenances.

Unique Identifier (UID): With reference to a given (possibly implicit) set of objects, a unique identifier (UID) is any identifier which is guaranteed to be unique among all identifiers used for those objects and for a specific purpose.

User Behaviour Tracking: A behaviour enabled by a *Community Support System* that to track the *Users*. If the research infrastructure has identity management, authorisation mechanisms, accounting mechanisms, for example, a Data Access Subsystem is provided, then the Community Support System either include these or work well with them.

User Group Work Supporting: A behaviour enabled by a *Community Support System* that to support controlled sharing, collaborative work and publication of results, with persistent and externally citable PIDs.

User Profile Management: A behaviour enabled by a *Community Support System* that to support persistent and mobile profiles, where profiles will include preferred interaction settings, preferred computational resource settings, and so on.

User Working Space Management: A behaviour enabled by a *Community Support System* that to support work spaces that allow data, document and code continuity between connection sessions and accessible from multiple sites or mobile smart devices.

User Working Relationships Management: A behaviour enabled by a *Community Support System* that to support a record of working relationships, (virtual) group memberships and friends.

Virtual Laboratory: Community proxy for interacting with infrastructure subsystems.



B. Common Requirements of ENVRI Research Infrastructures

From a pre-study of the ENVRI Research Infrastructures, a set of functionalities commonly provided by those research infrastructure have been identified, which is listed as follows highlighted with the minimal set of core functionalities.

A Data Acquisition Subsystem		
No	Functions	Definitions
A.1	Instrument Integration	A functionality that creates, edits and deletes a sensor.
A.2	Instrument Configuration	A functionality that sets-up a sensor or a sensor network.
A.3	Instrument Calibration	A functionality that controls and records the process of aligning or testing a sensor against dependable standards or specified verification processes.
A.4	Instrument Access	A functionality that reads and/or updates the state of a sensor.
A.5	Configuration Logging	A functionality that collects configuration information or (run-time) messages from a sensor (or a sensor network) and outputs into log files or specified media which can be used by routine troubleshooting and in incident handling.
A.6	Instrument Monitoring	A functionality that checks the state of a sensor or a sensor network which can be done periodically or when triggered by events.
A.7	(Parameter) Visualisation	A functionality that outputs the values of parameters and measured variables to a display device.
A.8	<i>(Real-Time) (Parameter/Data) Visualisation</i>	A specialisation of (Parameter) Visualisation which is subject to a real-time constraint.
A.9	Process Control	An interface that provides operations to receive input status, apply a set of logic statements or control algorithms, and generate a set of analog and digital outputs to change the logic states of devices.
A.10	Data Collection	An interface that provides operations to obtain digital values from a sensor instrument, associating consistent timestamps and necessary metadata.
A.11	<i>(Real-Time) Data Collection</i>	A specialisation of Data Collection which is subject to a real-time constraint.
A.12	Data Sampling	An interface that provides operations to select a subset of individuals from within a statistical population to estimate characteristics of the whole population.
A.13	Noise Reduction	An interface that provides operations to remove noise from scientific data.
A.14	Data Transmission	An interface that provides operations to transfer data over communication channels using specified network protocols.
A.15	<i>(Real-Time) Data Transmission</i>	A specialisation of Data Transmission which handles data streams using specified real-time transport protocols.



ENVRI Common Operations of Environmental Research Infrastructures

A.16	Data Transmission Monitoring	An interface that provides operations to check and report the status of data transferring process against specified performance criteria.
B Data Curation Subsystem		
No	Functions	Definitions
B.1	Data Quality Checking	An interface that provides operations to detect and correct (or remove) corrupt, inconsistent or inaccurate records from data sets.
B.2	Data Quality Verification	An interface that provides operations to support manual quality checking.
B.3	Data Identification	An interface that provides operations to assign (global) unique identifiers to data contents.
B.4	Data Cataloguing	An interface that provides operations to associate a data object with one or more metadata objects which contain data descriptions.
B.5	Data Product Generation	An interface that provides operations to process data against requirement specifications and standardised formats and descriptions.
B.6	Data Versioning	A interface that provides operations to assign a new version to each state change of data, allow to add and update some metadata descriptions for each version, and allow to select, access or delete a version of data.
B.7	Workflow Enactment	An interface that provide operations or services to interprets predefined process descriptions and control the instantiation of processes and sequencing of activities, adding work items to the work lists and invoking application tools as necessary.
B.8	Data Storage & Preservation	An interface that provides operations to deposit (over long-term) the data and metadata or other supplementary data and methods according to specified policies, and make them accessible on request.
B.9	Data Replication	An interface that provides operation to create, delete and maintain the consistency of copies of a data set on multiple storage devices.
B.10	Replica Synchronisation	An interface that provides operations to export a packet of data from on replica, transport it to one or more other replicas and to import and apply the changes in the packet to an existing replica.
C Data Access Subsystem		
No	Functions	Definitions
C.1	Access Control	An interface that provides operations to approve or disapprove of access requests based on specified access policies.
C.2	Resources Annotation	An interface that provides operations to create, change or delete a note that reading any form of text, and to associate them with a computational object.



ENVRI Common Operations of Environmental Research Infrastructures

C.3	<i>(Data) Annotation</i>	A specialisation of Resource Annotation which allows to associate an annotation to a data object.
C.4	Metadata Harvesting	An interface that provides operations to (regularly) collect metadata (in agreed formats) from different sources.
C.5	Resource Registration	An interface that provides operations to create an entry in a resource registry and insert resource object or a reference to a resource object in specified representations and semantics.
C.6	<i>(Metadata) Registration</i>	A specialisation of Resource Registration, which registers a metadata object in a metadata registry.
C.7	<i>(Identifier) Registration</i>	A specialisation of Resource Registration, which registers an identifier object in an identifier registry.
C.8	<i>(Sensor) Registration</i>	A specialisation of Resource Registration which registers a sensor object to a sensor registry.
C.9	Data Conversion	An interface that provides operations to convert data from one format to another format.
C.10	Data Compression	An interface that provides operations to encode information using reduced bits by identifying and eliminating statistical redundancy.
C.11	Data Publication	An interface that provides operations to provide clean, well-annotated, anonymity-preserving datasets in a suitable format, and by following specified data-publication and sharing policies to make the datasets publicly accessible or to those who agree to certain conditions of use, and to individuals who meet certain professional criteria.
C.12	Data Citation	An interface that provides operations to assign an accurate, consistent and standardised reference to a data object, which can be cited in scientific publications.
C.13	Semantic Harmonisation	An interface that provides operations to unify similar data (knowledge) models based on the consensus of collaborative domain experts to achieve better data (knowledge) reuse and semantic interoperability.
C.14	Data Discovery and Access	An interface that provides operations to retrieve requested data from a data resource by using suitable search technology.
C.15	Data Visualisation	An interface that provides operations to display visual representations of data.
D	Data Processing Subsystem	
No	Functions	Definitions
D.1	Data Assimilation	An interface that provides operations to combine observational data with output from a numerical model to produce an optimal estimate of the evolving state of the system.
D.2	Data Analysis	An interface that provides operations to inspect, clean, transform data, and to provide data models with the goal of highlighting useful information, suggesting conclusions, and supporting decision making.

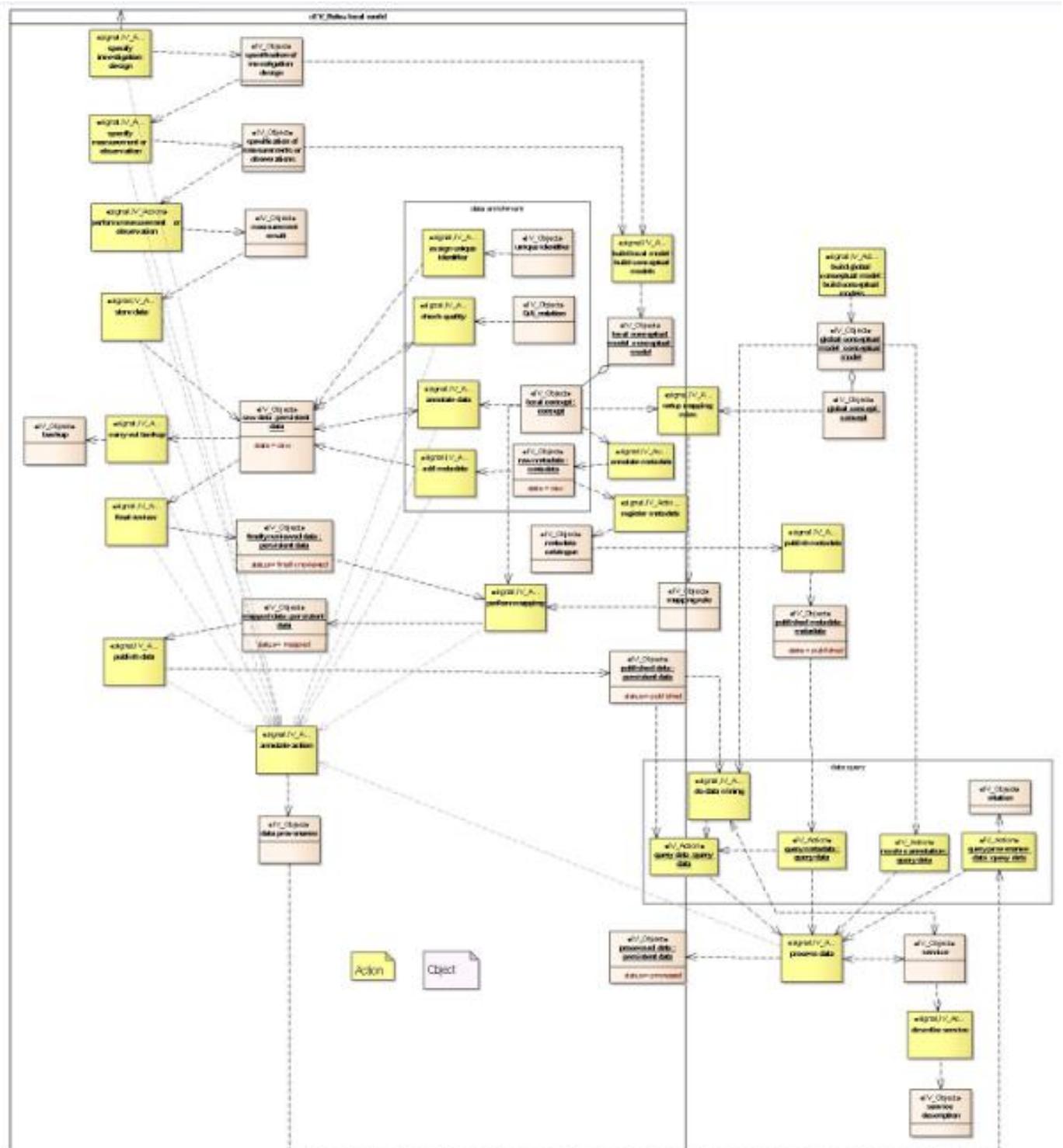


ENVRI Common Operations of Environmental Research Infrastructures

D.3	Data Mining	An interface that provides operations to support the discovery of patterns in large data sets.
D.4	Data Extraction	A interface that provides operations to retrieve data out of (unstructured) data sources, including web pages ,emails, documents, PDFs, scanned text, mainframe reports, and spool files.
D.5	Scientific Modelling and Simulation	An interface that provides operations to support of the generation of abstract, conceptual, graphical or mathematical models, and to run an instance of the model.
D.6	<i>(Scientific) Workflow Enactment</i>	A specialisation of Workflow Enactment, which support of composition and execution a series of computational or data manipulation steps, or a workflow, in a scientific application. Important processes should be recorded for provenance purposes.
D.7	<i>(Scientific) Visualisation</i>	An interface that provides operations to graphically illustrate scientific data to enable scientists to understand, illustrate and gain insight from their data.
D.8	Service Naming	An interface that provides operations to encapsulate the implemented name policy for service instances in a service network.
D.9	Data Processing	An interface that provides operations to initiate the calculation and manage the outputs to be returned to the client.
D.10	Data Processing Monitoring	An interface that provides operations to check the states of a running service instance.
E	Community Support Subsystem	
No	Functions	Definitions
E.1	Authentication	An interface that provides operations to verify a credential of a user.
E.2	Authorisation	An interface that provides operations to specify access rights to resources.
E.3	Accounting	An interface that provides operation to measure the resources a user consumes during access for the purpose of capacity and trend analysis, and cost allocation.
E.4	<i>(User) Registration</i>	A specialisation of Resource Registration which registers a user to a user registry.
E.5	Instant Messaging	An interface that provides operation for quick transmission of text-based messages from sender to receiver.
E.6	<i>(Interactive) Visualisation</i>	An interface that provides operations to enable users to control of some aspect of the visual representations of information.
E.7	Event Notification	An interface that provide operations to deliver message triggered by predefined events.



C. Dynamic Schemata in Details



Before a measurement or observation can be started the design (or setup) must be defined, including the working hypothesis and scientific question, method of the selection of sites (stratified / random), necessary precision of the observation or measurement, boundary conditions, etc. For correctly using the resulting data, detailed information about that process and its parameters have to be available for people processing the data. (e.g. if a stratified selection of sites according to parameter A is done, the resulting value of parameter A cannot be evaluated in the same way as other results)

After defining the overall design of measurements or observations, the measurement method, complying with the design, including devices which should be used, standards / protocols which should be followed, and other details have to be specified. Information of that process and the parameters resulting of the process have to be stored in order to guarantee correct interpretation of the resulting data. (e.g. when you want to model a dependency of parameter B of a parallel measured wind velocity, the limit of detection of the used anemometer influences the range of values of possible assertions).

When the measurement or observation method is defined, it can be carried out, producing measurement results. The handling of those results, all the actions done, to store the data are pulled together in the action "store data". (This action can be very simple when using a measurement device, which periodically sends the data to the data management system, but this can also be a sophisticated harvesting process or e.g. in case of biodiversity observations a process done by humans). The storage process is the first step in the life cycle of data that makes data accessible in digital form and are persisted.

As soon as data are available for IT purposes a backup can be made, independently of the state of the persisted data. This can be done locally or remote, done by the data owners or by dedicated data curation centers. At any status of the data can be processed for QA-assessments, for readjustment of the measurement or observation design and a lot of other reasons. Evaluations, which lead to answers of the scientific question, however, are usually done on data with a certain status - the status "finally reviewed".

Raw data can get enriched, which means that additional information can be added to them:

They can get a unique identifier, necessary for unambiguous identification of data (allowing to identify copies), and resolution within the data provenance.

They can undergo a QA process, checking their plausibility, correcting noise and several other processes, adding the information about that process to the data.

Metadata might be linked to the data, either by application of a standard metadata schema or by following a proprietaries metadata description.

semantic annotation can be added linking the data with their meaning. this semantic annotation can reach from annotation of units over annotation about used taxonomic lists to pointers to concepts in ontologies, describing the background of the measurement or observation in a machine and human readable form.



ENVRI Common Operations of Environmental Research Infrastructures

Making data accessible for users outside the Environment of the data owner at least needs two steps: 1) Mapping the data to the "global" semantics, the semantics the data owner shares with the data user. 2) Publish the data. Mapping data to global semantics may include simple conversions like conversions of units but also need more sophisticated transformations like transformations of code lists and other descriptions like the setup descriptions, measurement descriptions, and data provenance. "global" and "local" are usual, but a little bit confusing terms. There is nothing like a conceptual model for the whole world, which might be expected when we talk about a global conceptual model. A conceptual model (even if it is just a list of used units) is always just valid for a certain community. Whenever the community using some data, is widened, this widened community may have its new conceptual model. The case that two communities have the same model is a very rare luck. The smaller community has the so called "local" conceptual model and the larger community the so called "global".

It is important to know about published data, whether those data have a status: "finally reviewed" and what such a status means. It can mean, that those data will never change again, the optimum for the outside user. But it might also mean, that only under certain circumstances those data will be changed. In this case it is important to know what "certain circumstances" means. And additionally it is important to know, how and where the used semantics are described. A resolvable pointer to them, of course is the solution which can be handled most easily.

All the steps within the life cycle of data can be stored as data provenance, containing at least information about the used objects, the produced objects and the applied action. There are two important use cases for data provenance: 1.) citation of data and all the actors involved in the production of the data. 2.) correct interpretation of the data.

Data can be made directly accessible or indirectly in two or more steps. Direct one step access means, that you send a data request to a data server (query data) and get the data or an error message as answer. Indirect access or two step access means, that you first access metadata (query metadata) , search for a probably fitting data set and then query the data. Those two steps can be extended to more than two steps, when intermediate steps are involved. The two or more step approach is often used for data, which are not open, making metadata open but data not open. For questions touching several datasets and/or filtering the data (like e.g. give me all NO_x air measurement where O₃ exceeds a level of Y ppb) this two-step approach can be seen as a high barrier.