



## Deliverable 5.2: A definition of the ENVRIplus Reference Model

### WORK PACKAGE 5 – Reference-model guided RI design

LEADING BENEFICIARY: CARDIFF UNIVERSITY

Author(s):	Beneficiary/Institution
Alex Hardisty	Cardiff University
Abraham Nieva de la Hidalga	Cardiff University
Dan Lear	Marine Biological Association
Barbara Magagna	Austrian Environment Agency
Malcolm Atkinson	University of Edinburgh
Keith G. Jeffery	Natural Environment Research Council
Paul Martin	University of Amsterdam
Zhiming Zhao	University of Amsterdam

Accepted by: Paola Grosso (WP 5 leader)

**Deliverable type:** REPORT and Wiki

**Dissemination level:** PUBLIC

**Deliverable due date:** 31.October.2016/M18

**Actual Date of Submission:** 09.January.2017/M21



## ABSTRACT

The former ENVRI project (FP7 283465) made an analysis of the characteristics and requirements of six environmental research infrastructures (RIs) to identify commonalities and contrasts of function. This led to the design of the ENVRI Reference Model – an integrative reference approach that the environmental research community can use to secure interoperability between infrastructures, to enable reuse of common components, to permit sharing of resources, and to provide a common language of communication between those responsible for the design and construction of Research Infrastructures.

The present report starts from the conclusions of ENVRIplus deliverable D5.1 “A consistent characterisation of existing and planned RIs” (principally, sections 3.10 and 4.2.13) to further address the issues of:

1. Assessing the requirements of the new, larger group of RIs represented in the ENVRIplus project and whether these dictate revisions to the existing 3 viewpoints of the Reference Model;
2. Developing the Reference Model further, especially so that it can be used to formalise critical aspects of the architecture for interoperability between RIs, and for collaborative construction and maintenance involving multiple organisations.

ENVRI Reference Model version 2.1 (9<sup>th</sup> November 2016) is a refreshed and revised version incorporating changes arising from use in practice, as well as new developments since the previous version was published (version 1.1, August 2013). It reflects the needs of the wider base of Research Infrastructures surveyed as part of the requirements analysis and technology review activities in the present project.

This deliverable document is a description for the stakeholders of the changes to the ENVRI Reference Model that have been introduced to both improve the Reference Model generally and to respond to the newly collected and analysed requirements of the Research Infrastructures. As such, it is a snapshot at a stage in the evolution of the Reference Model for reference purposes in the future.

The work reported in the present document has been undertaken as part of Task 5.2 in Work Package (WP) 5, which itself is part of a closely related group of work packages forming Theme 2 “Data for Science”. This theme is concerned with the design, development and implementation of e-Infrastructure, methods, services and tools, that will help RIs more easily manage and fully exploit their data. The present report and the ENVRI RM itself should help Theme 2 integrate and steer its work to meet the priorities of the Research Infrastructures. Thus, beneficiaries and participants of Theme 2 constitute a second audience for this deliverable.



Project internal reviewer(s):

Project internal reviewer(s):	Beneficiary/Institution
Helen Glaves	Natural Environment Research Council
Markus Stocker	University of Bremen

Document history:

Date	Version
03.6.2016	Outline of document structure for comments
11.11.2016	First almost complete draft of text for comments
02.12.2016	Draft for internal review
05.01.2017	Corrected version
06.01.2017	Accepted by Paola Grosso (WP5 leader)

## DOCUMENT AMENDMENT PROCEDURE

Amendments, comments and suggestions should be sent to the editors (Alex Hardisty [HardistyAR@cardiff.ac.uk](mailto:HardistyAR@cardiff.ac.uk), Abraham Nieva [NievadelaHidalgaA@cardiff.ac.uk](mailto:NievadelaHidalgaA@cardiff.ac.uk), or one of the authors listed above.)

## TERMINOLOGY

A complete project glossary is provided online here: [envriplus.manageprojects.com/s/text-documents/LFCMXHHCwS5hh](http://envriplus.manageprojects.com/s/text-documents/LFCMXHHCwS5hh)

## PROJECT SUMMARY

ENVRIplus is a Horizon 2020 project bringing together Environmental and Earth System Research Infrastructures, projects and networks together with technical specialist partners to create a coherent, interdisciplinary and interoperable cluster of Environmental Research Infrastructures across Europe. It is driven by three overarching goals: 1) promoting cross-fertilisation between RIs, 2) implementing innovative concepts and devices across RIs, and 3) facilitating research and innovation in the field of environmental understanding and decision-making for an increasing number of users outside the RIs.

ENVRIplus aligns its activities to a core strategic plan where sharing multi-disciplinary expertise will be most effective. The project aims to improve Earth observation monitoring systems and strategies, including actions to improve harmonisation and innovation, and generate common solutions to many shared information technology and data related challenges. It also seeks to harmonise policies for access and provide strategies for knowledge transfer amongst RIs. ENVRIplus develops guidelines to enhance trans-disciplinary use of data and data-products supported by applied use-cases involving RIs from different domains. The project coordinates actions to improve communication and cooperation, addressing Environmental RIs at all levels, from management to end-users, implementing RI-staff exchange programs, generating material for RI personnel, and proposing common strategic developments and actions for enhancing services to users and evaluating the socio-economic impacts.

ENVRIplus is expected to facilitate structuration and improve quality of services offered both within single RIs and at the inter-RI (European and Global) level. It promotes efficient and multi-disciplinary research offering new opportunities to users, new tools to RI managers and new communication strategies for environmental RI communities. The resulting solutions, services and other project outcomes are made available to all environmental RI initiatives, thus contributing to the development of a coherent European RI ecosystem.



Blank page





## TABLE OF CONTENTS

Executive summary .....	11
1 Introduction .....	12
2 Concordance with the task description .....	12
2.1 Work Package objective .....	12
2.2 Task description .....	12
2.3 Principal concepts used in the present document .....	13
2.4 Abbreviations used in the present document .....	14
3 Connecting the ENVRI RM with other reference models adopted by existing RIs and related projects .....	14
3.1 Reference Models supporting Research Infrastructures .....	15
3.2 Reference models support in relation to the domain .....	16
3.3 Other reference models .....	17
3.4 ENVRI Reference Model for systems engineering of RIs .....	19
4 Validation of the Reference Model .....	19
4.1 The DASSH validation case .....	20
4.2 The EUFAR experience using the ENVRI Reference Model .....	21
4.3 Assessment of validation progress .....	23
5 Developing the Reference Model further .....	23
5.1 General improvements .....	23
5.2 Re-orientating around the data life cycle model .....	24
5.2.1 Explanation .....	24
5.2.2 Changes to RM Overview section .....	26
5.2.3 Changes to Science Viewpoint .....	26
5.2.4 Changes to Information Viewpoint .....	26
5.2.5 Changes to Computational Viewpoint .....	26
5.2.6 Other changes .....	26
5.2.7 Further considerations around the data life cycle .....	27
5.3 Continuous Consistency Assessments of ENVRI RM .....	27
5.3.1 Data Life cycle Reorientation Consistency Assessment .....	27
5.3.2 Viewpoint Correspondence Assessment .....	28
5.3.3 Corrective actions and recommendations for improvements .....	30
5.4 Alignment with outputs from Research Data Alliance (RDA) .....	30
5.4.1 Terms from the Data Foundations and Terminology Working Group (DFT WG) .....	30
5.4.2 Interaction with Data Fabric Interest Group (DFIG) .....	31
5.5 Contribution of Task 12.1, embedding the HES approach .....	31
6 Assessing the requirements of the new, larger community of environmental RIs .....	31
6.1 Approach for analysing RIs requirements .....	32
6.2 Recommendations derived from the analysis .....	32
6.3 Semantic Harmonisation (C.13) .....	34
6.3.1 Analysis and recommendation .....	34
6.3.2 Changes made to implement recommendations .....	35
6.4 Provenance Tracking .....	36
6.4.1 Analysis and recommendation .....	36
6.4.2 Changes made to implement recommendations .....	36
6.5 Data Cataloguing (B.4) .....	37



6.5.1	Analysis and recommendation .....	37
6.5.2	Changes made to implement recommendations .....	37
6.6	Data Identification (B.3).....	38
6.6.1	Analysis and recommendation .....	38
6.6.2	Changes made to implement recommendations .....	38
6.7	Data Citation (C.12) .....	39
6.7.1	Analysis and recommendation .....	39
6.7.2	Changes made to implement recommendations .....	40
6.8	Data Product Generation (B.5) .....	40
6.8.1	Analysis and recommendation .....	40
6.8.2	Changes made to implement recommendations .....	40
6.9	Data Processing Control (D.9).....	41
6.9.1	Analysis and recommendation .....	41
6.9.2	Changes made to implement recommendations .....	41
6.10	Data Use (E) .....	41
6.10.1	Analysis and recommendation .....	41
6.10.2	Changes made to implement recommendations .....	41
6.11	Data Publication (C.11).....	42
6.11.1	Analysis and recommendation .....	42
6.11.2	Changes made to implement recommendations .....	42
6.12	Data Discovery and Access (C.14).....	42
6.12.1	Analysis and recommendation .....	42
6.12.2	Changes made to implement recommendations .....	42
7	Moving towards engineering and technology .....	43
7.1	Engineering Viewpoint design approach .....	43
7.1.1	The Engineering Viewpoint subsystems .....	46
7.1.2	Interaction points .....	46
7.1.3	Identification of core competencies .....	47
7.1.4	Alignment with existing viewpoints.....	48
7.1.5	Recommended architectural style and sector trends.....	49
7.2	Approaching the Technology Viewpoint .....	50
7.2.1	Main objects of the Technology Viewpoint .....	51
7.2.2	Correspondence with the other four viewpoints .....	51
7.2.3	Relation of the TV to architectural and engineering models.....	52
8	Future enhancements to the Reference Model .....	52
8.1	Pending issues from T5.1 requirements analysis .....	52
8.1.1	Provenance Tracking.....	53
8.1.2	Non-Functional Requirements.....	53
8.1.3	Research campaigns .....	53
8.1.4	Adding support for canonical metadata .....	54
8.1.5	User-user interaction .....	54
8.2	Other potential enhancements .....	54
8.2.1	Scientific Workflows .....	54
8.2.2	Data management plan .....	54
8.2.3	Bulk data ingest.....	55
9	Steps into practice – putting the RM to work.....	55
10	Outlook and next steps .....	56
11	References .....	57



Appendix 1: National Marine Biodiversity Data Archive Centre (DASSH) – A validation case.....	60
Appendix 2: European Facility for Airborne Research (EUFAR) – A validation case .....	65
Appendix 3: Alignment of ENVRI RM with RDA terms and definitions .....	67
Appendix 4: Grouping of Analysis of results from D5.1 .....	71
Appendix 5: Research campaigns .....	81
Appendix 6: Ideas to facilitate RIs engagement.....	86
Appendix 7: ENVRI Reference Model version 2.1, November 2016 – Snapshot .....	89



## TABLE OF FIGURES

Figure 1: The concept of actor and its relationship to roles .....	22
Figure 2: Stages in the data life cycle .....	25
Figure 3 Brokered immediate data export from data archive on request from user(s) in a Virtual Laboratory .....	44
Figure 4 Brokered data import for curation in a data archive .....	44
Figure 5 Potential interaction points between EV subsystems .....	47
Figure 6 Identifying Core Competencies of ENVRI plus RIs .....	48
Figure 7 The Digital Object Cloud .....	50
Figure 8 ENVRI RIs within the Digital Object Cloud .....	50
Figure 9 Existing data management process .....	60
Figure 10 Existing dataflows at national level .....	61
Figure 11 Diagrams for the RM modelling of DASSH, Example 1 .....	62
Figure 12 Diagrams for the RM modelling of DASSH, Example 2 .....	62
Figure 13 Diagrams for the RM modelling of DASSH, Example 3 .....	63
Figure 14 Diagrams for the RM modelling of DASSH, Example 4 .....	63
Figure 15 Diagrams for the RM modelling of DASSH, Example 5 .....	63
Figure 16 Diagrams for the RM modelling of DASSH, Example 6 .....	64

## TABLE OF TABLES

Table 1: Stages in the data life cycle .....	25
Table 2 Initial Assessment of Viewpoints consistency .....	28
Table 3 Top ten mapped functionalities .....	32
Table 4 Summary of recommendations for enhancing the RM .....	33
Table 5: IV activities that support semantic harmonisation and metadata harvesting .....	35
Table 6 Relevant modelling components of the Science Viewpoint .....	65
Table 7 DFT WG Summary of Outputs .....	67
Table 8 IV terms compared to DFT WG terms .....	68
Table 9 DFT WG terms compared to IV Terms .....	69
Table 10 Distribution of requirements grouped according to D5.1 categories .....	71
Table 11 Distribution of requirements aligned to the Data life cycle phases .....	71
Table 12 Unmapped requirements .....	71
Table 13 Acquisition requirements .....	72
Table 14 Requirements which can map to any curation functionality .....	72
Table 15 Curation requirements which are single instances .....	73
Table 16 Curation requirements related to workflow enactment (D.6) .....	73
Table 17 Curation requirements related to data storage and preservation (B.8) .....	73
Table 18 Curation requirements for data identification (B.3) .....	73
Table 19 Curation requirements for data cataloguing (B.4) .....	74
Table 20 Curation requirements for data product generation (B.5) .....	74
Table 21 Publishing requirements with low count .....	75
Table 22 Publishing requirements mapped to data publication (C.11) .....	75
Table 23 Publishing requirements mapped to data citation (C.12) .....	75
Table 24 Publishing requirements mapped to semantic harmonisation (C.13) .....	76
Table 25 Publishing requirements mapped to discovery and access (C.14) .....	77
Table 26 Processing requirements with low count .....	77
Table 27 Processing requirements mapped to Scientific Workflow Enactment (D.6) .....	77
Table 28 Processing requirements mapped to data processing control (D.9) .....	78
Table 29 Use requirement mapped to data visualisation .....	79
Table 30 Use requirements mapped to authentication, authorisation, and accounting (E.1, E.2, E.3) .....	79
Table 31 Use requirements which cannot be mapped to existing functionalities .....	80



Table 32: Potential concepts in the Science Viewpoint .....	83
Table 33: Potential Relationships in the Science Viewpoint .....	83
Table 34: Potential concepts in the Information Viewpoint .....	84
Table 35: Potential Relationships in the Information Viewpoint .....	84
Table 36: Potential concepts in the Computation Viewpoint .....	84
Table 37: Potential Relationships in the Computation Viewpoint .....	85
Table 38: Potential concepts in the Engineering Viewpoint .....	85
Table 39: Potential Relationships in the Engineering Viewpoint .....	85



Blank page



## Executive summary

ENVRI Reference Model (RM) version 2.1, 9<sup>th</sup> November 2016 is a refreshed and revised version incorporating changes arising from use in practice, as well as new developments since the previous version was published (version 1.1, August 2013). Changes have been made both to improve the RM generally and to respond to newly collected and analysed requirements of the Research Infrastructures (RI). As such, version 2.1 is a snapshot at a stage in the evolution of the ENVRI RM for reference purposes by the principal stakeholders, namely the RIs themselves.

Undertaken as part of Task 5.2 in Work Package (WP) 5 of the ENVRIplus project, this work is part of a closely related group of work packages forming Theme 2 “Data for Science”, concerned with the design, development and implementation of e-Infrastructure, methods, services and tools, that will help RIs more easily manage and fully exploit their data. The present report and the ENVRI RM itself helps Theme 2 integrate and steer its work to meet the priorities of the RIs. Thus, project beneficiaries and participants of Theme 2 constitute a second audience.

The main advantage of the ENVRI RM over other kinds of reference model is that it has been developed specifically to address the needs of environmental research infrastructures. Serving as a knowledge base and transfer mechanism for RI professionals, it assists to raise the level of discourse in systems engineering and architectural design of ICTs for RIs. We have shown that the ENVRI RM relates to and complements other reference models, including both explicit models such as those promoted by the Open Geospatial Consortium (OGC) and the Group on Earth Observation System of Systems (GEOSS) and models implicit in the work of the Research Data Alliance (RDA).

Assessing the requirements of the new, larger group of RIs and whether these dictate revisions to the existing 3 viewpoints of the Reference Model is a main topic of the present report. Enhancements have been introduced to the RM to better support RI needs for data identification, cataloguing, product generation, publication, citation, semantic harmonisation, discovery and access, provenance, processing and use. Changes have also been made to re-orient the ENVRI RM towards and align it with a typical life cycle model for research data. Internal self-consistency of the RM has been improved. The presentation style of the RM documentation has been adjusted to make the concepts and vocabulary easier to comprehend.

Developing the Reference Model further, especially so that it can be used to formalise critical engineering aspects of the architecture for interoperability between RIs, and for collaborative construction and maintenance involving multiple organisations, is the next major step of evolution. To that end, the Engineering Viewpoint approach defines ‘interaction points’ for achieving interoperability between parts of an RI or between RIs, at which interfaces can be defined. It recommends a metadata-driven ‘brokered Service Oriented Architecture (SOA)’ as the preferred future architectural style for interoperability. From the technology perspective, the RM will encourage use of and foster seamless integration with shared European e-Infrastructures, including the European Open Science Cloud (EOSC).

There are several possible further enhancements of the RM by which it can gain value, including the provision of support for research campaigns, canonical metadata models, provenance tracking, scientific workflows and data management plans. Of these, the last – data management plans – offers the most immediate addition of value that can be beneficially exploited by the RIs.



# 1 Introduction

The ENVRI Reference Model (ENVRI RM) is an enabling tool – a framework, a ‘context for participation’ – that helps the Environmental Research Infrastructures (RI) to achieve their shared goals of cross-fertilisation, harmonisation and innovation between RIs; generating common solutions to many shared information and communications technology (ICT) and data related challenges. The ENVRI RM achieves this by offering a defined vocabulary for communication between participants, and a set of conceptual relationships between those terms as the basis for creating a common understanding and language for discourse.

In its details, the ENVRI RM has been elaborated based on reverse engineering 6 RIs (EISCAT-3D, EMSO, EPOS, EuroARGO, ICOS, LifeWatch) to determine the functionalities these RIs possess in common. These functionalities have been modelled according to three principal viewpoints (Scientific, Informational, Computational) informed by a general standardised model for “Open Distributed Processing” (ODP) [ISO/IEC 10746]. A minimal model of common functions of a research infrastructure within the context of a common life cycle for research data has been documented (<http://envri.eu/rm><sup>1</sup>) to serve as a guide for systems architects and others.

The ENVRI RM has been one of the successful outcomes from the original ENVRI project (EC grant agreement 283465) but with limited exploitation and use thus far. Its potential is therefore latent.

The ENVRI RM has been further enhanced in the ENVRIplus project to address a wider range of RI needs. Specifically, it has been enhanced to better support RI needs for data identification, cataloguing, product generation, publication, citation, semantic harmonisation, discovery and access, provenance, processing and use. Changes have also been made to re-orient the ENVRI RM towards and align it with a typical life cycle model for research data. The present report records and explains those developments up to the date of its publication. For the record, it includes a snapshot of the corresponding version 2.1 of the Reference Model. See Appendix 7<sup>1</sup> (page 86 of the present document).

## 2 Concordance with the task description

### 2.1 Work Package objective

The work reported in the present document is described as Task 5.2 in the Description of Activities of the ENVRIplus project. The relevant objective of the Work Package states that it aims to: “... promote interoperability among RIs by providing a novel ENVRIplus Reference Model which should be developed not only based on the existing ENVRI RM but should also include the latest development insights from other successful RIs.”

The decision has been taken at the commencement of the work to keep the name “ENVRI Reference Model” as a matter of maintaining the existing market branding, already receiving recognition in the areas to which it is targeted. This is in accordance with wider project decisions on promoting “ENVRI Community” as an overall brand for the community of the Environmental research infrastructures, projects and networks as well as other diverse stakeholders interested in the environmental research infrastructure matters.

### 2.2 Task description

The task description for Work Package 5, Task 5.2 states:

*“The ENVRIplus RM will be developed to be able to describe architectures at both the conceptual level (using Semantic Web/LOD) and at the data performance level with the*

---

<sup>1</sup> At any time, the current and most up-to-date published version of the ENVRI Reference Model can be found at the following url: <http://envri.eu/rm>.





former being generated from the latter (to preserve integrity) as demonstrated in the ENGAGE project. The current ENVRI RM focuses on the design of small set of RIs and was produced at a time when most of them were in their preparatory phase. During the ENVRI project, RIs made significant progress; partly exceeding the expressiveness of the existing ENVRI RM. The ENVRIplus Reference Model takes input not only from the existing ENVRI RM but also from the characterisation of RIs derived from Task 5.1, as well as from the contribution by Task 12.1. embedding the HES approach and from developments in EGI, Helix Nebula and EUDAT.

This task will take actions to:

- a) Develop the ENVRIplus RM based on the existing ENVRI RM, and results from Task 5.1.
- b) Connect the ENVRIplus RM with other successful reference models adopted by existing RIs or related projects;
- c) The Marine Biological Association coordinates the DASSH Data Archive Centre, which is a national facility for the archival of species and habitat data. It needs to be integrated with other European marine biological data (e.g. data curated by EMSO, SeadataNet, JERICO and EMBRC) as a joint contribution to EMODNET Biology, the COPERNICUS provider. This is a typical test case and will help improve the Reference Model.”

The present report is principally concerned with the results of action (a). This is reported in sections 5 (page 23), 6 (page 31), 7 (page 43), and 8 (page 52) below. It also addresses some aspects related to actions (b) (section 3 below, page 14) and (c) (section 4 below, page 19). Sections 9 (page 55) and 10 (page 56) are, respectively, concerned with activities associated with putting the RM into practice and an outlook of next steps in the evolution of the ENVRI RM.

Appendices 1 – 6 provide supplementary information. Appendix 7 (page 89 onwards) contains a full snapshot of version 2.1 of the ENVRI RM, taken on 9<sup>th</sup> November 2016.

## 2.3 Principal concepts used in the present document

For an introduction to the ENVRI Reference Model, the reader is referred to Appendix 7 beginning on page 89. The sections on “Getting Started”, “Introduction” and “Model Overview” provide a comprehensive introduction to the model and its concepts. [Zhao 2015] is an explanatory article.

Of specific importance, the following principal concepts, derived from terms defined in the [ISO/IEC 10746] series of International Standards for Open Distributed Processing (ODP) are used throughout the present document:

<b>Viewpoint (on a system)</b>	A form of abstraction achieved using a selected set of architectural concepts and structuring rules, focusing on specific concerns within a system (cf. the notion of different viewpoints as used in engineering and architectural blueprints).
	In the context of the present document, the system is the ICT system needed by an RI to support its operations on data. The environment within which that system operates is the specific RI it serves.
<b>Science Viewpoint (SV)<sup>2</sup></b>	A viewpoint on a system and its environment that focuses on the purpose (business), scope and policies for that system.
<b>Information Viewpoint (IV)</b>	A viewpoint on a system and its environment that focuses on the semantics of the information to be handled and the information processing to be performed by that system.

<sup>2</sup> In ODP, this viewpoint is referred to as the Enterprise Viewpoint but in ENVRI it has been named as the Science Viewpoint to more closely reflect that the business is science.



<b>Computational Viewpoint (CV)</b>	A viewpoint on a system and its environment that focuses on distribution through functional decomposition of the system into objects that interact at interfaces.
<b>Engineering Viewpoint (EV)</b>	A viewpoint on a system and its environment that focuses on the mechanisms and functions required to support distributed interaction between objects in the system.
<b>Technology Viewpoint (TV)</b>	A viewpoint on a system and its environment that focuses on the choice of technology in that system.

Viewpoints can be applied at an appropriate level of abstraction to specify a complete RI in the ENVRI context, in which case the environment defines the context in which the RI operates. The ENVRI RM is a minimal model representing the minimum set of most likely components to be found in an archetypical RI. In use, the RM can be extended as far as needed for any specific RI.

Viewpoints can also be applied to individual components of RI, in which case the component's environment will include some abstraction of both the RI's environment and other system components.

## 2.4 Abbreviations used in the present document

CV	Computational Viewpoint
EV <sup>3</sup>	Engineering Viewpoint
IV	Information Viewpoint
RI	Research Infrastructure (plural: RIs)
RM	Reference Model
SV <sup>3</sup>	Science Viewpoint
TV	Technology Viewpoint

## 3 Connecting the ENVRI RM with other reference models adopted by existing RIs and related projects

The Software Engineering Institute at Carnegie Mellon University gives definitions for Reference Model and Reference Architecture [SEI] as follows:

**Reference architecture:** A reference model mapped onto software elements that implements the functionality defined in the reference model.

**Reference model:** A division of functionality into elements together with the data flow among those elements.

The SEI definitions show that a reference model is concerned with providing a foundation to understand the functionalities of information systems at a high level. Meanwhile, reference architectures are concerned with mapping the functionalities to a set of software components.

There are many reference models in use; at least one for every software architecture model used for developing information systems. Wikipedia, for example lists 15 examples [Wikipediaorg 2016]. Reference models concentrate on describing the issues that are most important in relation to the architectural model and the types of applications to be built using that architecture.

For example, the Open Systems Interconnection Basic Reference Model [ISO/IEC 7498 1994] is an example of a general-purpose reference model designed to support the development of distributed systems using different standard communication protocols. The Open Geospatial

---

<sup>3</sup> Note that formally in ODP, the abbreviations NV and EV are used to denote the Engineering Viewpoint and Enterprise Viewpoint respectively. However, in adaptation to the ENVRI context we use EV to denote the Engineering Viewpoint and SV to denote the Science Viewpoint (see previous footnote also).

Consortium Reference Model [OGC ORM 2011] is an example of a domain specific reference model designed to support specifying and implementing systems for geospatial services, data, and applications. A reference model describes the common functionalities for the types of systems being described. Specific requirements that are important for a domain are not included in general-purpose reference models, while domain-specific reference models address the concerns of a specific domain, offering a further mapping of requirements to existing solution strategies.

### 3.1 Reference Models supporting Research Infrastructures

Environmental Research Infrastructures are supported by ICT systems that should be developed following state-of-the-art software engineering methods and architectures. These presently include the application of Layered, Service Oriented Architecture (SOA) and Cloud Architectural models<sup>4</sup>. Such systems are under constant pressure to connect with other systems to share data, manage increasing quantities of data, supporting growing research communities, and respond rapidly to new problems.

In the environmental RI domain, a multi-tiered or layered architecture model is often used in structuring applications that include large quantities of data backed up in database systems. Layered (tiered) architecture is based on the distribution of concerns. The number of tiers can vary but will at least separate aspects of presentation, processes, data access and physical resources clearly. The reference model associated with this architecture is mainly focused on the allocation of computing functions/units to the identified layers and the relationships between them. In this reference model, each layer (except for the top layer) provides services to the layer above it. The layered architecture model aims to clearly separate components that provide specific functionalities, such as persistence of data, implementation of process rules, or presentation details. Components in each tier collaborate directly with each other and can ask for services from lower tier components.

Some advanced infrastructures<sup>5</sup> provide access to data (sets) using a Service Oriented Architecture (SOA) framework. SOA helps to organise and utilise distributed capabilities that can be provided by systems that are remotely located [OASIS 2006]. The concepts of service, service provider and service consumer are important for this reference model and are terms that are often used in the context of conversations about ENV RIs and e-Infrastructures. SOA helps the integration of remote sourced components in the form of services. The services can be provided by entities (service providers) that can be external or internal with respect to the entities using those services (service consumers). European e-Infrastructures provide services based on the SOA model; for example, the B2xxxx services offered by EUDAT.

Cloud computing is one current model for enabling access to a shared pool of computing resources. Units in the pool (e.g., networks, servers, storage, applications, services) can be rapidly deployed and provisioned with minimal management effort or service provider interaction. The definition of cloud computing provided by the National Institute of Science and Technology (NIST) [NIST 2011] can be seen itself as a basic reference model for cloud computing, composed of five essential characteristics, three service models, and four deployment models. The essential characteristics (on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured service) highlight the main differences between cloud computing and traditional computing models. The three service models group the services in a layered structure with three levels Software, Platform and Infrastructure (as a Service), with the corresponding acronyms SaaS, PaaS and IaaS. The cloud deployment models describe who owns and operates the cloud resources (public, private, community, and hybrid). In the environmental RI domain RIs are not expected to develop their own cloud resources. They are encouraged, where appropriate to take advantage of

<sup>4</sup> The RI systems studied during the requirements analysis phase of ENVRIplus exhibit traits from these three architectural styles [Atkinson 2016].

<sup>5</sup> Although not an ENV RI in the context of the ENVRIplus project, the Global Biodiversity Information Facility (GBIF), <http://www.gbif.org/> is an example of an SOA-based infrastructure.



cloud computing resources provided by e-Infrastructures such as those provided by EGI.eu and EUDAT.

## 3.2 Reference models support in relation to the domain

In the environmental RI domain, ICT systems tend to follow the RM principles of their main adopted architectural models. During the development process, the systems are designed to respond to specific requirements from the institutions that sponsor them. The Layered, SOA, and Cloud Architecture reference models help to explain and support important engineering and technical solutions. Often, the explanation of specific functional requirements for the environmental research domain cannot be documented or explained directly without them.

The Layered architecture reference model helps to explain the need for designing system components at different levels depending on the functionalities they support. This RM provides the best strategies to integrate systems that bring together different technologies.

The SOA reference model describes the components of service oriented systems and how they are integrated. The focus of this RM is usually the distribution of functionalities across institutional boundaries using networks.

The Cloud Architecture reference model describes the fundamental principles for extending the use of distributed computing assets at infrastructure, platform and software levels. The focus of this RM is on the technical requirements for extending systems that are deployed on distributed computing resources.

One focus of RIs is the sharing of research data with a wide user base. In this domain, layered architectures can support the building of backend systems used for the administration of RIs and for cataloguing and publishing of data via Web portals. SOA can help RIs to provide access to data sources using service Application Programming Interfaces (API). Cloud computing can help in provisioning and scaling multiple services such as data transfer, storage and processing. However, designing an ICT system looking at these three models in isolation from one another makes it difficult to see that system as a single coherent entity.

In contrast, the ENVRI RM presents a unified view of the RI ICT system as a coherent whole, supporting the aims of the research community. The ODP model on which ENVRI RM is based already considers bringing together different architectural models (styles) to construct the complex distributed system. The ENVRI RM addresses the major types of ICT and data-oriented operations that a given RI likely needs to support. It exposes and explains how concentrating on certain types of operations on data in each of the phases of a typical data life cycle model will help to determine what RIs need provide with their ICTs. Medium-term, we expect this new approach to reduce the cost of building, maintaining and connecting RIs by letting each RI concentrate on its core competencies.

In-house developed environmental RI systems are maintained by small ICT staff teams and their most common model is layered architecture. In addition to these systems, solutions provided by e-Infrastructures service providers such as EUDAT and EGI can extend and complement in-house RI systems offering services designed following a mix of layered, SOA and Cloud architectures.

One of the greatest problems for RIs is that they are developed as one-of-a-kind systems. Apart from data sharing, RIs are not traditionally built for sharing and reuse of their engineered assets. This development mode is in contrast with current trends and proposals for shared infrastructures at National, European, and World levels [Hodson 2016; EOSC 2016; RDA 2016a]. The current efforts by e-Infrastructures offering shared solutions to common problems are an example of these trends. However, RIs tendency to reinvent rather than reuse existing assets leaves them unprepared for taking advantage of those solutions. There is a consequential loss of opportunities for shared operational support and maintenance, thereby reducing overall life cycle costs.



These considerations and those of section 3.1 above lead towards the Engineering Viewpoint proposals in section 7.1.5, page 49 below.

### 3.3 Other reference models

As mentioned before, there are many reference models. Moreover, the emergence of customised models to address the needs of specific sectors gives rise to the appearance of domain specific reference models. This is why the discussion of the previous sub-section focused on the models most well-known for the development of environmental RIs. However, other models could also be used to provide the support that the ENVRI RM provides, such as TOGAF [TOGAF 2011], OGC ORM [OGC ORM 2011], DL RM [Candela 2011], RM OAIS [CCSDS], EuroGEOSS Model [Santoro 2016] and the emerging RDA Data Fabric model [RDA 2016b]. TOGAF is a general-purpose reference model while the others are focussed towards specific priorities, such as representation of geospatial systems, digital libraries, systems of systems, and research data sharing frameworks.

The Open Group Architecture Framework (TOGAF) [TOGAF 2011] is a reference architecture designed to facilitate the construction of enterprise systems. TOGAF has two associated RMs: *Foundation Architecture: Technical Reference Model* (FA TRM) and the *Integrated Information Infrastructure Reference Model* (III RM). FA TRM describes the generic services and functions that provide a foundation on which more specific architectures and architectural components can be built. The III RM is a subset of the FA TRM which expands the business applications and infrastructure applications parts to provide help in addressing the need to design an integrated information infrastructure to enable unbounded information flow. TOGAF is the like RM ODP (and by extension, ENVRI RM) because it also organises the description of enterprise systems in viewpoints, which it calls architectures. As such, TOGAF Business Architecture can be mapped to the Enterprise Viewpoint (Science Viewpoint in ENVRI RM), the TOGAF Information Systems Architecture can be mapped to the Information Viewpoint, the Information System Data Architecture is equivalent to the Computational Viewpoint, the Information System Application Architecture is equivalent to the Engineering Viewpoint, and the Technology Architecture is equivalent to the Technology Viewpoint. From this perspective TOGAF could potentially be adapted to produce architecture and a reference model like those developed by ENVRI.

OGC ORM [OGC ORM 2011] is designed to specify and implement interoperable solutions for geospatial services, data, and applications, including sensor and sensor networks. The relevance to RIs' ICT systems for environmental research is obvious. This OGC ORM also presents the different perspectives for the description of systems. The Geospatial Information section can be mapped to the Information Viewpoint. The Geospatial Services section is like the Computational Viewpoint. The Reusable Patterns section can be mapped to the Engineering Viewpoint, and the OGC standards section can be mapped to the Technology Viewpoint. The Enterprise (Science Viewpoint) is the least developed of the perspectives in OGC ORM. In this sense the OGC ORM aims to keep generality in terms of applicability. This makes sense since Geospatial services are important for multiple domains, which may need to use them in different areas including industrial research, academic research, government, and commerce.

Digital Libraries Reference Model (DL RM) [Candela 2011] is designed to facilitate construction and validation of digital libraries. Digital libraries are highly related to RIs because both are based on the sharing of digital data objects. DL RM fosters the sharing of digital objects that include several forms of data including text, images, videos and sound, in addition to databases and metadata. All these types of objects are important to RIs. The DL RM also overlaps with the ENVRI RM in terms of the functionalities for preserving, curating, publishing and using data (digital objects). The acquisition and processing of data are the areas which DL RM does not support completely. The presentation of the DL RM is organised as domains. Each domain models a set of aspects of digital libraries. These domains are not comparable to viewpoints one to one. For instance, the DL RM user, policy and functionality domains have elements that can be mapped to the Science Viewpoint in the ENVRI RM, while the architecture domain can be mapped to elements of the



computational, engineering and technology domains. Many elements from all the DL RM domains can also be mapped to the Information Viewpoint of the ENVRI RM. The DL RM is domain specific in the sense of focusing on the preservation, curation and publishing of data (digital objects), however, these activities are important in multiple areas such as industrial research, academic research, government, and commerce.

The Reference Model for an Open Archival Information System (RM OAIS [CCSDS]) describes an archive that accepts the responsibilities to preserve information and make it available for a designated community. The RM OAIS describes the responsibilities which distinguish an OAIS Archive using a functional model consisting of six functional entities: Ingest, Archival Storage, Data Management, Administration, Preservation Planning, and Access. The Ingest, Archival Storage, Data Management, and Access functional entities participate directly in data processing. The Administration functional entity provides functionalities for the operation of the archive. The Preservation Planning functional entity provides functions and services to monitor the OAIS. RM OAIS describes the evolution of data in the archive as three types of information packages: Submission Information Package (SIP), Archival Information Package (AIP), and Dissemination Information Package (DIP). An Information Package is a logical container composed of optional Content Information and optional associated Preservation Description Information. Associated with this Information Package is Packaging Information used to delimit and identify the Content Information and Package Description information used to facilitate searches for the Content Information. The main focus points of the RM OAIS are the curation, preservation, publishing and use of data, these focus points are closely related to the same activities in the ENVRI RM. However, ENVRI also covers data acquisition. Regarding the depth of abstraction coverage, the issues addressed by OAIS are equivalent to the ones covered by the SV and IV, while the RM also provides the computational viewpoint which supports a concrete IT architecture model. The long-term preservation planning functional entity in OAIS which separates steering/management from administration, is an important issue to consider for the further development of the ENVRI RM which will be needed by all RIs and any subset that chooses to interoperate (as discussed in section 3.4).

The GEO (Group on Earth Observation) in its GEOSS (Global Earth Observation System of Systems) program has proposed a model to extend the SOA architectural model using advanced brokering to facilitate the integration of research repositories into executable workflows [Santoro 2016]. The approach proposes a high-level architecture that can be used to integrate different data sources into a dedicated pipeline, starting from the definition of a high level abstract business process (ABP), the building of a workflow from the ABP, the execution of the workflow and the publishing of results from execution. In the context of environmental research infrastructures, the GEOSS proposal could be used in three different areas: (1) defining an acquisition-curation-publishing pipeline, (2) building processing tools supported by workflows and (3) facilitating the integration of RIs as data sources.

The definition of an acquisition-curation-publishing pipeline (1) scenario envisions the RI describing the acquisition, curation, and publishing tasks as an ABP. The ABP could then be converted into a workflow and published in a web portal for repeated execution. In the building workflow-supported processing tools scenario (2), an RI could define its data processing needs as an ABP, which in turn could be converted into executable workflows and made available to users via a web interface. Finally, the facilitating the integration of RIs scenario (3), could propose a case in which a group of RIs define the integration of their repositories into different sets of ABPs which in turn are built into workflows. These workflows could then be exposed in a common virtual research environment. The GEOSS proposal is complementary to the ENVRI RM proposal. The possibility of integration of RIs proposed by GEOSS seems to be the most relevant scenario. The approach relies strongly on the existence of data repositories, but does not go into detail about how those repositories are built and sustained. Additionally, the GEOSS brokering approach is presently (late 2016) under consideration by the RDA data fabric interest group (DFIG) [RDA 2016b] as a guide for the development of an architectural model.





The Research Data Alliance (RDA) [RDA 2016a] is an international organisation focused on the development of recommendations, guidelines and community activities aimed at reducing barriers to research data sharing and exchange; and at promoting acceleration of data driven innovation worldwide. The RDA addresses many of the same issues as the ENVRI RM. Many of the interest and working groups within the RDA focus on topics that directly impact those addressed by the ENVRI RM. However, so far RDA does not provide a reference model for integrating the technical recommendations and standards in the development of systems. Within the RDA, the aim of the data fabric interest group (DFIG) [RDA 2016b] is to build a set of recommendations for integrating the RDA outputs and facilitating the building of systems to support data intensive research – the same goal as the ENVRI RM. The goal of DFIG is to identify common components and define characteristics and services that can be used across boundaries to solve a variety of data scenarios such as replicating data in federations, and developing virtual research environments. The work of this interest group is at an early exploratory stage. However, it is positioned to influence the way in which RI products (data collections and identification systems) are integrated at a high level. Consequently, customising the outputs through the ENVRI RM as they are developed can prepare the environmental RI domain.

### 3.4 ENVRI Reference Model for systems engineering of RIs

Connecting the ENVRI RM to the reference models adopted by existing RIs or related projects has been realised during the basic definition of the ENVRI RM. It has exposed clearly the areas where the ENVRI RM overlaps with other models, the areas where they diverge, and the perceived advantages or disadvantages of each model. The main advantage of the ENVRI RM over other RMs is that it was developed specifically to address the requirements of environmental research infrastructures. The adoption and use of reference models in the environmental research domain for the development of research infrastructures has been limited thus far and not clearly oriented towards the specific systems engineering level challenges of the domain. The ENVRI RM presents an approach aimed at addressing typical environmental research requirements that is technology-neutral. The ENVRI RM serves as a knowledge base and transfer mechanism derived from domain knowledge that can be mapped to technical standards and implementations used to define architectures for environmental research infrastructures.

We note that in domains where RMs are well embedded, their respective communities recognise the importance of RMs as mechanisms steer engineering. In the domain of environmental RIs, towards which the ENVRI RM is targeted, RIs presently depend on internal and informal mechanisms to steer their e-infrastructure design and development. As they grow in complexity, and as the inter-RI issues are addressed this requires the establishment of formal processes adjudicating the mappings of RMs and other long-term decisions to shape their support for data-driven science. In the context of ENVRI RM, the formal process must be performed through a long-term governance mechanism for refining and updating the ENVRI RM. In this scenario, the ENVRI community is expected to eventually take up the governance responsibilities.

## 4 Validation of the Reference Model

Validation of the RM by attempts to use / exploit it in different systems design contexts is an essential part of understanding whether the RM has been specified correctly and of demonstrating its value to stakeholders.

As part of the task work, two implementation cases, DASSH and EUFAR are presently being studied jointly by project experts and RIs with the aim of describing aspects of two RIs using the ENVRI Reference Model (RM). This will help to detect the RM's limits and thus also support its general improvement. The following subsections (4.1 and 4.2) report experiences to date for work that is presently still in progress. Appendices 1 and 2 contain further background information on the two cases.



## 4.1 The DASSH validation case

*“Within a week I could understand Reference Model concepts and the relation between viewpoints, and I can use the model to frame the problem case. It has become evident that the focus of the problem is on the sharing of the published data from DASSH into other kinds of systems”*

*Dan Lear – Head of Data, Information and Technology. Marine Biological Association (UK).*

The DASSH Data Archive Centre is hosted at the Marine Biological Association (MBA) and is part of the UK Marine Environmental Data and Information Network (MEDIN). Data are routinely supplied to the UK National Biodiversity Network, EMODNET Biology network and GBIF. It is planned that the e-Infrastructure will form part of the MBA's contribution to the European Marine Biological Resource Centre (EMBRC). The infrastructure has grown organically since 2005, in response to technological changes, policy drivers and the requirements of funders. Staff involved in the DASSH Data Centre are typically marine biologists with some technical aptitude. They are not ICT specialists.

The ENVRI Reference Model presents an ideal opportunity to review the current systems and working processes/practices, describe them in a standard way, facilitate integration with other e-Infrastructures and assist in the identification of bottlenecks or areas for improvement.

Given the previous lack of experience with Reference Models in any form, the initial challenge faced has been to understand the overall concept of what a Reference Model represents, and the benefits of describing the DASSH infrastructure in this manner. With the support of the team at Cardiff University, MBA received a comprehensive introduction to the RM, and how it can be applied.

With this basic understanding established the first task was to pick a Viewpoint and begin to map existing data flow processes and systems to terms from the RM. The Science Viewpoint was selected as the most accessible point of entry, and the one that most closely resembled existing data flow and process diagrams developed by DASSH.

However, choosing the starting point of describing the infrastructure in the terms of the Science Viewpoint proved one of the hardest tasks. The available guidance does not help in the prioritisation or refinement of selecting a path through the viewpoints, and can leave the naïve user unsure or unclear of the optimal mechanism for transcribing an existing set of processes and behaviours into those terms available in the Reference Model. Whilst this lack of a prescribed methodology represents a strength and inherent flexibility for more confident users of the RM, it can leave new users with an intimidatingly “blank” canvas. Work on how best to render the DASSH infrastructure in terms of the RM was also needed. The lack of mature or easily accessible tools necessitated an initial paper-based approach, with this subsequently being interpreted to an electronic form. It is unclear how to best represent the correspondences between viewpoints.

A further challenge, especially for a first-time user of the Reference Model is assessing the most appropriate level of detail at which to work. This should be informed by the challenge or tasks that the RM is being used to address. However, clearer guidance on both where to start and how much detail to include would be beneficial.

MBA opted to focus on the data publication elements of the DASSH data life cycle as this represents the interface with other e-Infrastructures. It is also the area where MBA have identified an existing bottle-neck in the data flow process. Once a starting point was selected, and with help from the RM team at Cardiff University, MBA split and transcribed the existing data flows into the corresponding RM viewpoints, moving from the Science Viewpoint to the Information and Computational Viewpoints.





The provision of a suggested tool (UMlet<sup>6</sup>) and guidance on the recommended notation is also a great help for those users with no previous experience in UML or ODP.

The experience of the Data Team at the MBA in learning and using the Reference Model can be summarised in the following points:

- Clearer, simpler documentation using non-specialist language would greatly benefit the inexperienced user, including how to illustrate correspondences;
- More and varied examples would help define a starting point and the level of detail;
- Develop templates using the UMlet tool (or similar) linked to examples; and,
- Build a community of users who can share experience and support uptake.

## 4.2 The EUFAR experience using the ENVRI Reference Model

EUFAR works to coordinate the operation of instrumented aircraft and remote sensing instruments, exploiting the skills of experts in airborne measurements in the fields of environmental and geosciences, to provide researchers with the infrastructure best suited to their needs. EUFAR's networking activity "Standards and Protocols" aims at contributing to the development of standards and recommendations to harmonise the structure and methods of operating research infrastructures. This will allow it to interact with EU integrating activities for environmental research (e.g., ACTRIS, IAGOS, EUROFLEET, etc.). In this context, ENVRIplus is a major driver, and EUFAR has taken the opportunity to apply to serve as an implementation/use case for the ENVRI Reference Model. This predefined professional framework allows EUFAR to clearly define roles and processes in their Research Infrastructure operations and helps to describe the current situation as well as to find missing and/or duplicated actions. This activity will be of major benefit to EUFAR in preparing a sustainable legal structure

During the 3<sup>rd</sup> EUFAR GA meeting in Prague, April 2016, the kick-off of the cooperation between EUFAR and ENVRIplus (Barbara Magagna, EEA) took place. A special session on ENVRIplus has been organized to introduce the purpose of the intended cooperation and to explain the RM in more detail. It has been decided, that the EUFAR leaders of "Standards and Protocols" (Stefanie Holzwarth, DLR) and "Database" (Wendy Garland, STFC) work packages will be the persons in charge of implementing the RM for EUFAR. Both are experienced in using models to map complex processes.

To start the implementation case, a requirements collection questionnaire (Version 5.2: 16 November 2015) provided by ENVRIplus had to be filled out concerning the seven Theme 2 vertical pillar topics of ENVRIplus (i.e., data identification and citation, cataloguing, etc.). The completion of the questionnaire helped EUFAR experts to understand the wording of the RM and identify the different viewpoints. Several iteration processes were necessary to finalise the input. It became clear, that some of the topics are out of scope for EUFAR, e.g., processing which is done by each data provider individually using the local infrastructure and tools.

Assisted by EEA (B. Magagna), the key EUFAR people started to collect all relevant modelled components of the Science Viewpoint into a Microsoft® Excel spreadsheet. These are Communities, Roles and Behaviours (illustrated in Table 6 in Appendix 2, page 65). Sometimes it was not possible to do a one-to-one mapping of the EUFAR roles to the given Science Viewpoint Roles. Further work is in progress on this aspect, including perhaps enhancing the 'standard RM' with roles specific to EUFAR.

This output has been used to create instances within the OIL\_E ontology [Martin 2015] to make the EUFAR description comparable with other RI descriptions such as ICOS. In principle (but as yet

---

<sup>6</sup> <http://www.umlet.com/>, an open-source Java-based UML tool designed for teaching the Unified Modelling Language (UML) and for quickly creating UML diagrams.

untested) if one can use the same classes to define instances one can potentially compare different descriptions.

Data collection has been the initial focus of modelling. Since this is also the most relevant part of EUFAR, and therefore also the most familiar one, the translation seemed to be easy. Looking at the result it has become clear that some process steps must be described in more detail, since more actors are involved. In EUFAR many different actors with different responsibilities are involved in relevant community behaviours. Thus, it was decided to analyse the concept of actor and its relationship to roles, according to the implemented OIL\_E ontology. This is illustrated in Figure 1.

## Science viewpoint relationships

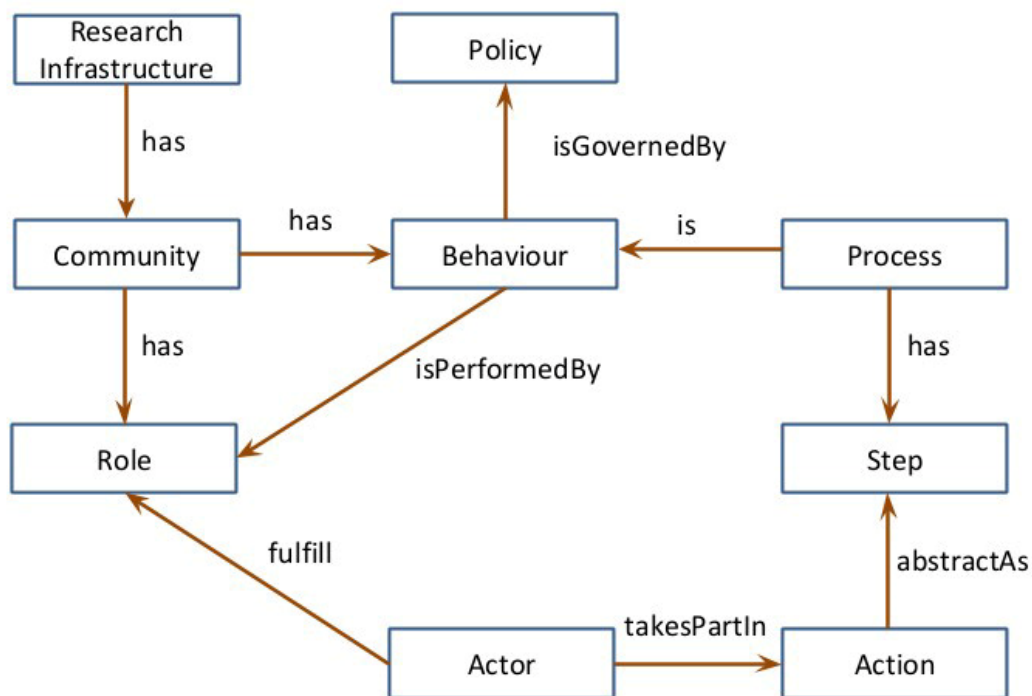


FIGURE 1: THE CONCEPT OF ACTOR AND ITS RELATIONSHIP TO ROLES

In addition, UMLet diagrams are being produced to document the different elements. To be able to make transparent the different steps within the processes describing the behaviours, activity diagrams are being produced. These diagrams indicate easily whether or not the interfaces between different actors are defined sufficiently. The use of the diagrams eases the iterative process of behaviour description.

The first behaviours analysed in these details are:

- Data Collection
- Data Curation
- Data Discovery and Access
- User Behaviour Tracking

The next focus of work will be the Information Viewpoint with the modelling of the data life cycle of EUFAR datasets.

There are many relevant activities across EUFAR consortium members. Modelling these as the Science Viewpoint has proved to be crucial to gaining a better understanding of how these are distributed, and in supporting internal decision-making.

### 4.3 Assessment of validation progress

The two validation cases described above have been in progress since early 2016. A few general conclusions can be drawn, that supplement what is already known from earlier ENVRI work.

Approximately 3-4 weeks of effort has been invested by each team to initially learn the RM, familiarise with concepts and outline some of their first thoughts on design. In both validation cases, payback has been achieved quite early in the process. In the DASSH case, this has been to achieve thinking on the separation of different concerns. In the EUFAR case, it has been the realisation of greater complexities than first imagined when working across multiple stakeholders.

Adopting the RM approach is a gradual process that shifts thinking and mindset towards the systems-oriented approach that is necessary when dealing with complex ICT design and interoperability. This is known to some but not appreciated by many that are active in the RI arena. Once an initial understanding has been acquired, early adopter users of the RM are finding it helpful to consider their RI from different viewpoints, and that there are answers provided by the RM to most of the questions they have about their design.

The usefulness and benefit of the RM is only as good as the available tools and the community of experts around it. Growth will ease other people's future experiences, especially with the availability of more guidance materials based on practical examples and visual representations. This would help not only the early adopters but also the wider community.

While there is only a small user community at present, without much contact and cross-fertilisation between segments of it, the ENVRIplus project has planned a number of activities to address this issue. Materials aspects are being partially addressed in Work Package 15 of the ENVRIplus project by preparation of e-Training materials. Section 9 below considers the community building aspect in more detail.

One recognized shortcoming of the validation cases is that there is almost no resource effort to pursue them to their logical conclusions. The participants are already finding they could spend a lot more time on them. The perspective of validation must therefore be more long-term, allowing for a process of gradual enlightenment and adoption. Bringing about such change is slow, which makes the process of RM validation slow. This is related to the lack of significant numbers of persons working in the RI arena that have previous experience of 'the commercial/industrial level' of applying ICT architecture and conceptual modelling processes and tools to large-scale RI ICT design problems. Such persons are valuable and scarce but where they exist they facilitate matters in terms of marketing, adopting and gaining benefit of the RM. The RM is by nature complex, as is the process of Research Infrastructure ICT design. Appropriately skilled individuals are essential to the successful outcomes of RI design. Such individuals help to embed the work in the target organisations, which increases its value to those and increases survivability when project funding ends. Section 9, page 55 below makes some suggestions for some further practical steps into practice; for putting the RM to work in the ENVRI community.

## 5 Developing the Reference Model further

### 5.1 General improvements

The opportunity has been taken to incorporate changes aimed at improving readability and comprehension, and to simplify the overall presentation of the RM. This has been in response to feedback received about the difficulty of becoming familiar with and getting something useful from the RM. These changes have included re-drawing of diagrams, re-structuring of the page hierarchy in Information and Computational Viewpoints and numerous typographical improvements.



Nevertheless, the Reference Model remains an extensive technical document that requires some investment of time on the part of the reader to become familiar with it.

## 5.2 Re-orientating around the data life cycle model

### 5.2.1 Explanation

Early versions of the RM and its 3 viewpoints (science, information, computational) had presented the model in a sub-systems oriented view of the world. With hindsight, we learnt that this emphasis on subsystems early in the description of the RM makes the model appear more technical than it needs to be. Sub-systems more properly have their place in elaborating the Computational and Engineering Viewpoints (see section 7.1 below) whereas what is needed in the introductory parts of the RM and in its Science and Information Viewpoints is a more business oriented view. Additionally, it is now thought useful to be able to show how the RM contributes to supporting typical activities in the life cycle of research data, implemented by research and e-infrastructures.

Thus, version 2.0 of the ENVRI RM has been updated to re-orientate and align it with a “typical” data life cycle model. The adopted life cycle model, illustrated in Figure 2 and Table 1 below has been designed in accordance with the main state changes to data being processed by RIs<sup>7</sup> and initially documented in deliverable D5.1 [Atkinson 2016]. It has been checked against other published variants of a life cycle for research data that are often cited (e.g., DataONE [Michener 2012], Digital Curation Centre [Higgins 2008], Data Documentation Initiative [DDI 2015]) and found to be compatible.

---

<sup>7</sup> As originally documented in the Information Viewpoint of version 1.1 of the RM, and based on analysis of the original 6 RIs.

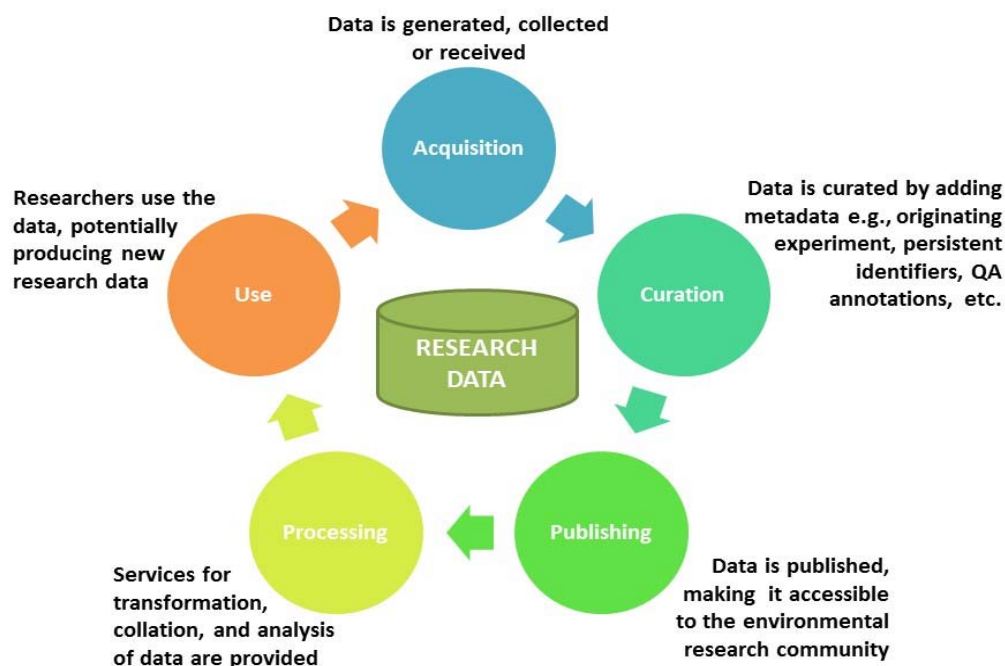


FIGURE 2: STAGES IN THE DATA LIFE CYCLE

TABLE 1: STAGES IN THE DATA LIFE CYCLE

Phase	Data state	Definition	Supporting activities
<b>Acquisition</b>	Acquired [Registered]	Data generated by experimental processes, manual observation or automatic recording of events or phenomena, and stored in digital form	Activities related to setting up monitoring devices or networks of such devices. Providing recording tools for individuals. Transmission and storage of data in digital form (digitisation), regardless of the lifespan assigned to collected data. Note that some elements of curation (such as annotating data with metadata at the point of acquisition in the field) can considerably improve the overall quality of data.
<b>Curation</b>	Curated [Annotated, QA assessed, reviewed, mapped]	Additional data created to facilitate identification and retrieval	Activities designed to preserve, link, and identify data; such as: quality assessment, annotation, digital identification (DOI) and cataloguing.
<b>Publishing</b>	Published	Additional data created to facilitate access	Activities designed to make data accessible to other parties
<b>Processing</b>	Processed	Data which has been further processed (visualised, summarised, further annotated)	Activities designed to support advanced processing of data (querying, summarising, visualising, modelling, among others)
<b>Use</b>	Processed [used, cited, referenced]	Data used for creating new data which can undergo further processing and storage	Activities designed to derive new data products, including information and knowledge

The first set of changes required a review of the RM and determining whether pre-existing references to “subsystems” should all be changed to life cycle phases. This included introducing the data life cycle model as part of the RM overview, and editing each viewpoint to change the

focus from the definition of subsystems to supporting the research data life cycle. This led to the publication of a re-vamped version 2.0 of the RM in July 2016.

The following sub-sections (5.2.2 - 5.2.6) summarise the principal changes made in version 2.0.

### 5.2.2 Changes to RM Overview section

The RM Overview has been edited to introduce the data life cycle. The changes include adding a diagram to explain the life cycle model, and editing the description of how the RM supports systems that process research data following the life cycle model. In the previous version (1.x) the overview started with the presentation of subsystems. In this version (2.x), subsystems descriptions are replaced by the description of the phases of the data life cycle.

The RM has value because of its harmonising role. This is not to say that it harmonises the solutions adopted by RIs nor does it harmonise the requirements that RIs express. What it does do is to harmonise the context in which we talk about these things so that we can understand one another and so that we can identify points at which interactions and/or integrations between an RI and either another RI or a digital infrastructure for research (DI4R) can take place.

The relationships between the life cycle phases are introduced as “integration points”. Integration points are identified as the main transition boundaries between different phases of the data life cycle, expressed as the possibility to achieve some integration between one RI and another or between one RI and a digital infrastructure for research (DI4R). At these points, data from one phase is passed to another phase for further processing.

### 5.2.3 Changes to Science Viewpoint

Aligning with the data life cycle touched the three main parts of the Science Viewpoint (SV) – Communities, Roles and Behaviours. However, changes have been minimal compared to the changes in the other two viewpoints.

### 5.2.4 Changes to Information Viewpoint

The changes to the Information Viewpoint (IV) derived from aligning with the data life cycle have been significant. Re-structuring and naming of the sections have been changed to reflect purpose. For instance, the “Dynamic Schemata” section has been changed to “Data Evolution”, and the “Static Schemata” section has been renamed as “Data Management Constraints”. Additionally, sections have been rearranged to provide a more compact presentation of the IV concepts.

In addition to this, many diagrams have been replaced with versions that can be created and shared using open source tools for UML modelling.

### 5.2.5 Changes to Computational Viewpoint

The Computational Viewpoint (CV) is the first part of the model where an approach to description based around the subsystems idea seems to become more appropriate. The existing description has been edited to emphasise that the purpose of subsystems is to support the different phases of the data life cycle and to prepare for engineering. In addition to this, the sections of the CV have been reorganised to differentiate between component configurations that support the data life cycle directly and those configurations that support infrastructure integration.

### 5.2.6 Other changes

In addition to the changes outlined above, this review of the RM also led to corrections and updates of information. For instance, sections with duplicated material have been removed; and dead or erroneous links have been corrected/updated as needed.



### 5.2.7 Further considerations around the data life cycle

Figure 2 and Table 1 summarise the phases of a typical life cycle for research data. By necessity, these are simplified representations of what happens in many real-world scenarios. Here we present some additional information to be considered further and accounted for in the RM.

During the curation phase, cataloguing is vital to make data 'FAIR' – Findable, Accessible, Interoperable, Reusable [FAIR 2016]. Catalogues must provide access not only to the raw data but to frequently computed derivatives so that scientists re-use those, rather than requesting re-computation. This is significant in terms of saving energy and costs, and helps with standardisation of interpretations. However, scientists must be able to access the raw data when they need to.

In data publishing, there can be policies governing the timing and nature of publication e.g., in respect of embargo periods and the accessibility of data. Data discovery and retrieval processes must recognise these and permit, for example authorised retrieval based on valid credentials of the researcher making the request; or limitations on direct retrieval of data based on, for example its sensitivity and/or quantity.

Much processing and use of data can involve interaction with other curated data and/or operational data from wider international (i.e., global) activities and infrastructure agreements. Consideration must be given as to how such interactions at the process level are accommodated.

Many of the above and similar considerations are active topics of discussion in the research infrastructures community and in the Research Data Alliance [RDA 2016a]. It is a complex area with many variations and possibilities that still must be further refined towards consensus agreements on best practice.

## 5.3 Continuous Consistency Assessments of ENVRI RM

The analysis of the consistency of the ENVRI RM has been carried out as a continuous progression guided by two main objectives. The first objective of the consistency review is to guarantee that the ENVRI RM provides coverage for all the functionalities included in the minimal set defined in the RM. The second objective is to verify that the models between the different viewpoints are consistent with each other. This section describes the gaps found while trying to verify the fulfilment of these two objectives by inspection and cross-checking and explains the solutions advanced to bridge those gaps.

The consistency assessment has covered two phases, the assessment carried out during the reorientation with the data life cycle model (version 2.0) and the assessment during the alignment with new requirements (version 2.1). These phases are presented as follows.

### 5.3.1 Data Life cycle Reorientation Consistency Assessment

The minimal set of functionality which is aligned with the common functions within the data life cycle should be consistent with the concepts defined in different viewpoints. An important guiding principle for identifying key correspondences came from the 2<sup>nd</sup> ENVRIweek meeting<sup>8</sup> during which it was established that:

“RIs are autonomous entities interacting with each other to service scientists' needs; this suggests that identifying the correspondences that need to be made visible in the pursuit of interoperability between RIs is of great relevance.”

In line with this definition, the assessment of the ENVRI RM V 1.1 exposed six specific consistency issues which needed attention. These issues and their resolution are detailed in Table 2.

---

<sup>8</sup> Zandvoort, Netherlands, May 2016



TABLE 2 INITIAL ASSESSMENT OF VIEWPOINTS CONSISTENCY

Issue	Advance towards resolution
1. The SV does not present a process view. Processes would tie together the definition of behaviours, roles and communities, without a process view it is hard to identify the relationship between roles and behaviours.	Work in progress: The diagrams that present the interaction of actions in the evolution of information objects were redesigned to better illustrate the alignment with the data life cycle. The correspondences between IV activities and SV activities is being used to derive similar diagrams which include swim lines for assigning behaviours to roles.
2. The IV of the ENVRI RM does not present a clear view of the different information objects and relationships between them	Solution: The IV models were redesigned to illustrate the relationships between information objects types.
3. Some object types not sufficiently clearly defined e.g., Concept, Conceptual model	Solution: The IV information object definitions were revised and organised to clarify terms.
4. Specification of investigation may contain specifications of measurements which in turn can be used to evaluate measurement results; but this is not well illustrated; leaving the reader to infer it.	Solution: The definitions were left open to allow the RIs to define their own constraints with regards of limits, validation and QA.
5. The definitions of data, information, concept, knowledge are not clearly specified in the RM. The RM uses these terms interchangeably, impeding the understanding of what is meant.	Solution: The definition for data, information, concept, knowledge should be aligned with the needs of different RIs. To limit ambiguity, all references to information or data have been edited as information objects.
6. The definition of architecture should start in the computational viewpoint and then be refined further in the engineering and technology viewpoints. The CV lacks a description of the architecture for the distribution of the computation objects.	Solution: The presentation of CV objects was reorganised according to architectural layers to support the design of different configurations for all phases of the data life cycle

### 5.3.2 Viewpoint Correspondence Assessment

The three viewpoints of the ENVRI RM help to model the systems that support the operation of environmental research infrastructures from three complementary perspectives.

The high-level models of the Science Viewpoint are representations of the descriptions of the scientific community that will work with the RI, and of their business. The data models of the Information Viewpoint represent the translation of science viewpoint processes into actions designed to encode those processes and their results in machine readable formats. The model of the Computational Viewpoint assigns the information actions to different computational units that will effectively carry out the behaviours and processes defined in the Science Viewpoint.

This means that the subsystems, processes, actors, and related objects that constitute an RI are modelled from three different angles or perspectives, as in other areas of engineering (e.g., civil, mechanical). To guarantee that the different models from each viewpoint are consistent and correspond to one another, ODP supports the notion of 'viewpoint correspondences'. Correspondences are relationships between an object in one viewpoint and objects in another viewpoint(s). Correspondences can be one-to-one, one-to-many, or many-to-many.



Consistency across viewpoints is verified by predefined correspondences between the objects of those viewpoints. To facilitate verification, the correspondences can be grouped according to the phases of the research data life cycle. This minimises the number of objects that must be considered at any one time. A judicious selection of correspondences must be made to signify the main relations whilst avoiding complexity through proliferation. Even so, automation of consistency verification is essential to avoid human error.

The following sections 5.3.2.1 - 5.3.2.3 describe the key correspondences between the three viewpoints: SV-IV, IV-CV, and SV-CV, including information about gaps between viewpoints.

#### **5.3.2.1 SV-IV correspondences**

There are two types of correspondence between the Science Viewpoint and the Information Viewpoint: (1) SV roles correspond to IV information objects and (2) SV behaviours to IV actions.

The first type of SV-IV correspondences oversees the encoding of SV roles into IV information objects. In its current version, the ENVRI RM does not include a detailed structure for defining entities such as users, institutions, or research groups. They are considered basic roles that any research infrastructure will have codified somehow in their administrative systems<sup>9</sup>. The main information objects modelled thus far are data centric and correspond to the encoding of the products of research in digital form. These include: the specification of experiment, the details of the measurements, the persistent data objects, and the catalogues of metadata associated with those objects. Nevertheless, in the future consideration must be given to adding this support (see 8.1.4).

The second type of SV-IV correspondences maps SV behaviours to IV actions. IV actions are applied to IV objects, changing their state and creating new IV objects. The correspondences between the SV behaviours and the IV actions during acquisition, curation and publishing are the most consistent ones. However, the mappings between SV behaviours to IV actions in the processing and use phases present some gaps. For instance, the SV behaviours for the data use community cannot presently be mapped to any specific action of the IV. Additionally, the IV actions such as query data, query metadata, do data mining and process data cannot be mapped to existing SV behaviours.

#### **5.3.2.2 IV-CV correspondences**

There are two types of correspondences between the Information Viewpoint and the Computational Viewpoint: (1) IV information objects correspond to CV objects and (2) IV actions to CV objects functionalities.

Information Viewpoint information objects correspondences to Computational Viewpoint objects are expressed as IV objects manipulated or stored by specific CV objects. These correspondences are implicit and do not require further analysis. The consistency of these correspondences is verified as all the IV objects can be mapped to at least one CV component that will either store or manipulate that object.

Information Viewpoint actions correspondences to Computational Viewpoint objects functionalities are expressed as IV actions performed by specific CV objects. The functionalities of CV objects are expressed through their Interfaces. In this case, verifying correspondence looking at interfaces only is difficult, because the interface definitions are mostly oriented towards facilitating the integration of components rather than describing the different ways in which those components can be used. In this case, the description of the functionalities of the CV objects should describe explicitly which functionalities are supported by the components. For instance, the IV action do data mining could be supported using a combination of CV objects that could include Virtual Laboratory, Data Broker, Catalogue Service, Annotation Service, and Data Store

---

<sup>9</sup> For example, CERIF based 'current research information systems' (CRIS) such as Converis from Thomson Reuters or the open source DSpace tool.



Controller. However, none of the interfaces to those components indicate that they could be used for data mining.

### 5.3.2.3 SV-CV correspondences

The correspondences between the Science Viewpoint and the Computational Viewpoint are defined as transitive SV to CV correspondences. Transitive SV to CV correspondences are defined through SV-IV correspondences and IV-CV correspondences. In this case, if for instance, SV object A corresponds to IV object B and if IV object B corresponds to CV object C, then it can be said that SV object A corresponds to CV object C. In this case the problems gaps detected in the SV-IV and IV-CV correspondences will affect directly the SV-CV correspondences. In addition to this, some correspondences seem to bypass the IV. For instance, the CV security service can be mapped to the SV behaviours for the data use community, which the IV does not address (user behaviour tracking, user profile management, user working profile management, user working work relationship management and user group work support). This is partly due to missing IV elements, as mentioned in 5.3.2.1 above. It is important for security and privacy that users (especially in role experts) are considered part of the RM. Similarly, methods underpinning the definitions of behaviours in the SV should map directly as interfaces in the CV.

### 5.3.3 Corrective actions and recommendations for improvements

Up to this point, we assume that the gaps and mismatches between the different viewpoints are not yet a serious problem for the users of the ENVRI RM. Users of the ENVRI RM can fill the gaps by adding specific objects required when modelling specific needs in their domain – as they can add additional objects for their domain. However, the analysis above will be used in the future to try to bridge the detected gaps:

- Science Viewpoint: Create or redefine behaviours that support the correspondences to the IV actions: query data, query metadata, do data mining and process data;
- Information Viewpoint: Create information objects and actions that support the data use community behaviours i.e., existing and planned working practices;
- Computational Viewpoint: Redefine objects so that their functionalities can be mapped to IV actions and SV behaviours rather than relying solely on the definitions of their interfaces. Review the naming and definition of objects so that correspondences are easier to identify;
- For security and privacy purposes, consider what additional elements may be needed in the IV.

## 5.4 Alignment with outputs from Research Data Alliance (RDA)

### 5.4.1 Terms from the Data Foundations and Terminology Working Group (DFT WG)

Based on data models presented from different disciplines, including from the former ENVRI project and on interactions with different scientists and scientific departments, the Data Foundations and Terminology (DFT) working group of the Research Data Alliance (RDA) has produced several inter-related reports (see bibliography, page **Error! Bookmark not defined.** in Appendix 3). RDA defined several simple definitions for digital data in a registered domain, based on an agreed conceptualisation. The essential output of the DFT group – definitions and model – is condensed in the document "DFT Core Terms and Model-v1-6" [DFT WG – RDA 2015] where 14 terms and their relationships are defined.

The terms defined in the ENVRI RM Information Viewpoint (IV) are the closest match to those described by the DFT WG. Appendix 3 (page 67) summarises and makes a comparison (Table 8 on



page 68 and Table 9 on page 69) of the terms in the IV and the corresponding term in the DFT Core Term Definitions (DFTWG-CTD).

The compatibility of the RDA DFT terminology and the ENVRI RM set of IV information objects will facilitate aligning the ENVRI RM with the work of other RDA groups. Moreover, the alignment can support the development of collaboration/cross-fertilisation activities between the ENVRI community and the RDA.

#### 5.4.2 Interaction with Data Fabric Interest Group (DFIG)

The RDA's Data Fabric Interest Group (DFIG) is concerned with identifying a set of Common Components and defining their characteristics and services that can be used across boundaries in such a way that they can be combined to solve a variety of data scenarios such as replicating data in federations, developing virtual research environments, supporting the exchange of data and so forth<sup>10</sup>.

In late October 2016, contact has been initiated between ENVRIplus project and DFIG with the aim of ensuring alignment and compatibility of the work emerging from the two efforts. Specifically, there is agreement to identify how and whether the ENVRI RM can contribute to addressing the problem areas being studied in DFIG.

### 5.5 Contribution of Task 12.1, embedding the HES approach

The HES (Human-Environment Systems) approach is an integrated approach towards studying the interface and reciprocal interactions that link human sub-systems (e.g., economic, social, agricultural) to natural sub-systems (e.g., hydrologic, atmospheric, biological) of the planet<sup>11</sup>.

The societal grand challenges identified and studied by WP12 could be modelled within the Community concept of the Science Viewpoint, where one can already find a differentiation according to its objective. The application area could be related to the five life cycle phases (data acquisition, data curation, data publishing, data processing, data use) or at least this approach could be used as a starting point for further modelling in this respect. Whether it should be provided as an integral part of the RM or as a supplement to it has still to be decided. On the other hand, the grand challenges are not yet defined properly and still under discussion. It has been decided to provide a detailed model for this part within the RM only after a clear acceptance of the concept of grand challenges between the RIs. This should be possible during 2017.

## 6 Assessing the requirements of the new, larger community of environmental RIs

The ENVRI RM has been developed originally based on the common functionalities identified across 6 RIs [Chen 2013a, Chen 2013b].

Assessing the requirements of the new, larger community of RIs represented in the ENVRIplus project, the Reference Model has been improved [ENVRI RM version 2.1 2016]. This has involved mapping the outcomes of the requirements collection and analysis task, Task 5.1 [Atkinson 2016] into changes that must be made to the RM. The approach used and the main recommendations made are explained in sections 6.1 and 6.2 respectively. The detailed analysis, recommendations and modifications made are explained for each of the main topic areas in sections 6.3 –6.12 following below.

<sup>10</sup> <https://www.rd-alliance.org/group/data-fabric-ig/wiki/recommendations.html>

<sup>11</sup> e.g., [https://en.wikipedia.org/wiki/Coupled\\_human%E2%80%93environment\\_system](https://en.wikipedia.org/wiki/Coupled_human%E2%80%93environment_system)



## 6.1 Approach for analysing RIs requirements

The requirements report, deliverable D5.1 [Atkinson 2016] has been reviewed and analysed as follows:

1. First, requirements have been collected from the document into a spreadsheet.
  - a. Each sentence expressing a requirement was converted to a spreadsheet row;
  - b. Each requirement was mapped to a phase of the data life cycle and to one of the functionalities described within the ENVRI RM by adding a tag indicating phase and a tag indicating functionality;
  - c. Each requirement that was not a straightforward mapping received an interpretation and/or a keyword which supported mapping; and,
  - d. A comment about the mapping or what to do with the requirement was added.
2. Requirements have been grouped and summarised (see Appendix 4, page 71); and,
3. A recommendation on how to proceed after the analysis has been provided (see section 6.2).

## 6.2 Recommendations derived from the analysis

A total of 178 requirements have been derived from deliverable D5.1. These requirements have been mapped to the set of functionalities used for building the ENVRI RM. The count of requirements mapped to functionalities was used to prioritise which requirements indicated areas of concern. This mapping produced a list of the 10 most important functionalities to target during the update of the ENVRI RM. The full list of identified requirements and the mapping to functionalities for selection is described in Appendix 4. The top ten functionalities that map to requirements arising from D5.1 are listed in Table 3. The first three columns contain the identifier (ID), Name and Definition<sup>12</sup> of the functionalities. The fourth column contains the count of requirements derived from D5.1 mapped to that functionality. These top ten functionalities account for 137 of the 178 requirements.

TABLE 3 TOP TEN MAPPED FUNCTIONALITIES

ID	Functionality	Definition	Mapping count
B.3	Data Identification	A functionality that assigns (global) unique identifiers to data contents.	10
B.4	Data Cataloguing	A functionality that associates a data object with one or more metadata objects that contain data descriptions.	12
B.5	Data Product Generation	A functionality that processes data against requirement specifications and standardised formats and descriptions. (optional)	8
C.11	Data Publication	A functionality that provides clean, well-annotated, anonymity-preserving datasets in a suitable format, and following specified data-publication and sharing policies to either make the datasets publicly accessible or to those who agree to certain conditions of use, and to individuals who meet certain professional criteria.	11
C.12	Data Citation	A functionality that assigns an accurate, consistent and standardised reference to a data object, which can be cited in scientific publications.	9
C.13	Semantic Harmonisation	A functionality that unifies similar data (knowledge) models based on the consensus of collaborative domain experts and derived published data to achieve better data (knowledge) reuse and semantic interoperability.	25
C.14	Data Discovery and Access	A functionality that retrieves requested data from a data resource by using suitable search technology.	5

<sup>12</sup> <https://wiki.envri.eu/display/EC/Appendix+A+Common+Requirements+of+Environmental+Research+Infrastructures>

ID	Functionality	Definition	Mapping count
D.6	(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.	20
D.9	Data Processing Control	A functionality that initiates the calculation and manages the outputs to be returned to the client.	19
E	Data Use	Data use in general - Usability, user support, configurability, service provision, portals	19

The shortlisted functionalities have been reviewed to determine whether they cover all aspects of the mapped requirements. If they cover all aspects, then no further action was needed. Otherwise, there are three possibilities: i) modify the definition; ii) refine the functionality, or iii) create a new functionality. The first option requires the least changes. The second option extends the existing functionality by adding refinements that account for the requirements needs. The third option creates a new functionality in the RM (maybe even at a different level than the one mapped initially); this would point to new requirements not recognised previously (i.e., from the initial analysis of 6 RIs carried out by the forerunner ENVRI project, [Chen 2013a]).

Table 4 lists the recommendations for modifying the RM according to the analysis. The following sections (6.3– 6.12) explain in detail how these recommendations have been derived for each of the top ten functionalities.

**TABLE 4 SUMMARY OF RECOMMENDATIONS FOR ENHANCING THE RM**

1. Data Identification (B.3)
<ul style="list-style-type: none"> <li>Correct the model at science and computational viewpoints. <ul style="list-style-type: none"> <li>Science Viewpoint should accommodate identification within the curation community both in roles and in behaviours.</li> <li>Computational Viewpoint: place this functionality at the use phase. Move to curation phase.</li> </ul> </li> <li>Specific types of identification should be included as alternatives in the technology viewpoint specification. This should include use of handles (DOI, PID, or other similar) and strategies for using file names and other types of identifiers (e.g. URI).</li> </ul>
2. Data Cataloguing (B.4)
<ul style="list-style-type: none"> <li>Catalogue manager could be a new role in the SV. The role could be a passive role (i.e. software agent implements it); this would correspond directly to the Catalogue service defined in the CV.</li> <li>Specific types of catalogues could be included as alternatives in the technology viewpoint specification.</li> </ul>
3. Data Product Generation (B.5)
<ul style="list-style-type: none"> <li>Create the data process manager. The role could be a passive role (i.e. software agent implements it); this would correspond directly to the process controller service defined in the CV.</li> <li>New data products are created at all stages of the data life cycle. For instance, backup, metadata, provenance, and other such data objects can be seen as data products. In a sense the creation of data products is already accounted for in the RM. A discussion about the pertinence of being explicit at naming data products would be required but not urgent.</li> <li>Highly relevant for Provenance.</li> </ul>
4. Data Publication (C.11)
<ul style="list-style-type: none"> <li>Requirements lack enough support or relevance to merit modification recommendations</li> </ul>
5. Data Citation (C.12)
<ul style="list-style-type: none"> <li>Follow the model proposed in the computational viewpoint and place this functionality at the use phase. The use phase is where citations are produced (to indicate that data has been used).</li> <li>Specific services for handling citations may be needed, especially for linking citations to provenance and resources.</li> </ul>
6. Semantic Harmonisation (C.13)
<ul style="list-style-type: none"> <li>Expand the minimal model to include semantic harmonisation</li> <li>Add a role in the SV to deal with harmonisation explicitly</li> <li>Make explicit the support of harmonisation in the IV, indicate the support in the definition of mapping rule, setup mapping rule and perform mapping.</li> <li>Make explicit the support of harmonisation in the CV, may need a specific component to facilitate mapping.</li> </ul>

- |     |  |   |
|-----|--|---|
| 7.  | Data Discovery and Access (C.14)   | <ul style="list-style-type: none"> <li>Requirements lack enough support or relevance to merit modification recommendations</li> </ul>   |
| 8.  | (Scientific) Workflow Enactment (D.6) – Rebranded as Provenance tracking | <ul style="list-style-type: none"> <li>These requirements were wrongly aligned to scientific workflow processing functionalities given that it is the only one explicitly supporting provenance. Provenance is wider as demonstrated in the IV.</li> <li>Provenance tracking a data use requirement. Users of data are informed about the processes which generated the data by provenance information. Provenance should be linked to the use phase</li> <li>Generation of provenance data should be linked to the use phase of the data life cycle</li> <li>The SV should define roles and behaviours supporting provenance tracking</li> <li>The CV should define specific objects and interfaces which support provenance tracking</li> <li>Provenance Tracking is not part of the minimal model. Should be included in the minimal model.</li> </ul> |
| 9.  | Data Processing Control (D.9)  | <ul style="list-style-type: none"> <li>Requirements lack enough support or relevance to merit modification recommendations</li> </ul>   |
| 10. | Data Use (E)   | <ul style="list-style-type: none"> <li>Requirements lack enough support or relevance to merit modification recommendations</li> </ul>   |

Resulting from this analysis, the priorities to address the changes to the RM have been determined to be the following:

1. Semantic Harmonisation (C.13) – Not part of the minimal model
2. Provenance Tracking – Closely related to processing control and workflow enactment (D.6)
3. Data Cataloguing (B.4) – Revise status
4. Data Identification (B.3) – Revise status
5. Data Citation (C.12) – Revise status
6. Data Product Generation (B.5) – Revise status
7. Data Processing Control (D.9) – Revise status
8. Data Use (E) – A new set of functionalities can be derived from these requirements
9. Data Publication (C.11) – Revise status
10. Data Discovery and Access (C.14) – Revise status

The following sections (6.3 - 6.12) explain the analysis of the specific sets of requirements and the changes made to the RM to address them in detail.

## 6.3 Semantic Harmonisation (C.13)

### 6.3.1 Analysis and recommendation

The requirements mapped to harmonisation point out needs to assure interoperability, data sharing, linking RIs, homogenisation, collaboration between RIs, discovery and reuse, enhanced interaction, and synchronisation. Of the different RIs, twelve indicated these needs in the generic information section of D5.1: ACTRIS, AnaEE, EISCAT-3D, ELIXIR, EMBRC, EMSO, FixO3, IAGOS, INTERACT, IS-ENES2, LTER, SeaDataNet (SeaDataNet includes three other RIs: EUROFLEETS2, JERICO and ESONET).

The ENVRI RM version 2.0 describes semantic harmonisation as “Functionality that unifies similar data (knowledge) models based on the consensus of collaborative domain experts to achieve better data (knowledge) reuse and semantic interoperability”. However, semantic harmonisation is not part of the minimal model.

**Science Viewpoint:** Semantic Harmonisation is defined as one of the behaviours of the data publication community. No role is linked expressly to semantic harmonisation, in theory any of the actors could then perform the behaviour.



**Information Viewpoint:** No information object or activity is expressly marked as supporting semantic harmonisation. However, the definition of the mapping rule object, together with the setup mapping rule and perform mapping action indicate support for harmonisation.

**Computational Viewpoint:** The Experiment laboratory implements a translate request interface which is used when mappings between semantic domains are required. Experiment laboratory is used in processed data import and internal data staging.

The following are the recommendations for enhancement:

- Expand the minimal model to include semantic harmonisation;
- Add a role in the SV to deal with harmonisation explicitly;
- Make explicit the support of harmonisation in the IV, indicate the support in the definition of mapping rule, setup mapping rule and perform mapping;
- Make explicit the support of harmonisation in the CV, may need a specific component to facilitate mapping.

### 6.3.2 Changes made to implement recommendations

The following changes have been made to implement the above recommendations.

Semantic Harmonisation is defined as the functionality that unifies similar data (knowledge) models based on the consensus of collaborative domain experts to achieve better data (knowledge) reuse and semantic interoperability.

Metadata Harvesting is defined as the functionality that (regularly) collects metadata in agreed formats from different sources.

In the Science Viewpoint one role and two behaviours have been added to the SV to explicitly support semantic harmonisation and metadata harvesting:

**Semantic Curator:** An active role, which is a person who designs and maintains local and global conceptual models and uses those models to annotate the data and metadata.

**Select or Build Conceptual Model:** A behaviour performed by a Semantic Curator which supports the annotation of data and metadata.

**Data Annotation:** A behaviour performed by a Semantic Curator which supports the linking of data and metadata with conceptual models (global and local).

The activities in the Information Viewpoint (IV) already support semantic harmonisation and metadata harvesting.

TABLE 5: IV ACTIVITIES THAT SUPPORT SEMANTIC HARMONISATION AND METADATA HARVESTING

IV Activity	Semantic Harmonisation	Metadata Harvesting
annotate metadata	Metadata is linked to local and global conceptual models	--
annotate data	Data is linked to local and global conceptual models	--
annotate action	Create links to local and global conceptual models	--
resolve annotation	Dereference from local or global conceptual models to data objects	--
build conceptual models	Manually build the local and global conceptual models	Harvesting can support collecting and generating the models automatically
setup mapping rules	Manually create rules to link data objects to global and local conceptual models	Harvesting and pattern matching can be used to create rules automatically



IV Activity	Semantic Harmonisation	Metadata Harvesting
perform mapping		Mapping can be automatically performed

In the Computational Viewpoint, the data curation and data publishing subsystems have been redefined to include configurations where the semantic laboratory is used to annotate data and metadata and to align those annotations with local and global models. The semantic broker has been redefined to support harvesting and harmonisation.

## 6.4 Provenance Tracking

### 6.4.1 Analysis and recommendation

In the previous version of the ENVRI RM, provenance was only mentioned in the context of scientific workflow enactment. However, the requirements for provenance go beyond supporting scientific workflow enactment. In this case, the definition could be passed back to workflow enactment (D.6) or refined as provenance tracking.

The ENVRI RM version 2.0 describes scientific workflow enactment as “specialisation of Workflow Enactment supporting the composition and execution of computational or data manipulation steps in a scientific application. Important processing results should be recorded for provenance purposes”.

**Science Viewpoint:** There are no roles linked directly to support workflow enactment or provenance tracking. There are no specific behaviours linked to support workflow enactment or provenance tracking.

**Information Viewpoint:** Provenance is mentioned several times. The IV defines both an object and an action that implement provenance tracking.

**Computational Viewpoint:** The CV defines coordination and process controller services which in turn implement interfaces which support workflow execution and monitoring. Provenance is only mentioned marginally no component designed to support provenance tracking.

The following are the recommendations for enhancement:

- These requirements were wrongly aligned to scientific workflow processing functionalities given that it is the only one explicitly supporting provenance. Provenance is wider as demonstrated in the IV. By wider we mean that it should be included in all phases of the data life cycle from acquisition to use.
- Provenance tracking is a data use requirement. Human users of data are informed about the processes that generated the data by provenance information. Software components may also be informed about provenance and act upon it. Provenance should be linked to the use phase.
- Generation of provenance data should be linked to the use phase of the data life cycle.
- The SV should define roles and behaviours supporting provenance tracking.
- The CV should define specific objects and interfaces which support provenance tracking.
- Provenance Tracking is not part of the minimal model; it should be included in the minimal model.

### 6.4.2 Changes made to implement recommendations

The following changes have been made to implement the above recommendations.





The IV presents provenance as an activity that parallels the activities in the data life cycle, registering the outcomes of each phase as important provenance elements.

The diagrams of the IV have been redesigned to make clear the points where provenance data can be collected.

The corresponding roles and behaviours in the SV still require definition, as do the corresponding objects in the CV.

A new functionality needs to be included if provenance tracking is added as a parallel activity to all processes during the data life cycle.

For provenance reasons, all derived data products must be identified (see sections 6.6 and 0).

## 6.5 Data Cataloguing (B.4)

### 6.5.1 Analysis and recommendation

Requirements include the description of different types of functionalities for cataloguing including several types of: catalogues of observation systems and laboratory equipment; catalogues of physical samples; catalogues of data products and/or results; catalogues of publications; reference catalogues; federated catalogues; and processing catalogues. In addition to this, several characteristics of the catalogues are presented including: standardised approach (OGC/SWE, SSN, ISO/IEC 191XX), links for provenance between publications and datasets, automatically supplying the entire set of metadata characterising the task (e.g., through the provenance framework), optimise the management of provenance data streams, preserving the link between data and their provenance metadata.

The ENVRI RM version 2.0 describes data cataloguing as “A functionality that associates a data object with one or more metadata objects which contain data descriptions”.

**Science Viewpoint:** In the data curation community, the Data Curator role is described as being responsible for catalogues. No behaviour is proposed to support data cataloguing specifically.

**Information Viewpoint:** metadata catalogue is defined as an information object. Register metadata is defined as an information action which enters metadata into a catalogue. The “registered” state is a metadata state which indicates that metadata has been added to a metadata catalogue.

**Computational Viewpoint:** Catalogue service is defined as a curation service to support the curation of datasets. This service should provide four interfaces: export metadata, query data, update catalogue and query resource. The Catalogue service is used in brokered data export, brokered data import, brokered data query, processed data import, and raw data connection.

The following are the recommendations for enhancement:

- Catalogue manager could be a new role in the SV. The role could be a passive role (i.e. software agent implements it); this would correspond directly to the Catalogue service defined in the CV.
- Specific types of catalogues could be included as alternatives in the technology viewpoint specification.

### 6.5.2 Changes made to implement recommendations

The following changes have been made to implement the above recommendations.

A passive role has been added to the SV curation community to support cataloguing.



**Catalogue System:** A passive role, a catalogue system is a special type of storage system designed to support building of logical structures for classifying data and metadata to aid discovery, access and use.

## 6.6 Data Identification (B.3)

### 6.6.1 Analysis and recommendation

Requirements refer to the need for identification management functionalities that include: DOI management; standard (homogenous) approach to Identification; identification of dynamic data series; identification of results from data queries (e.g., data services); data identification automation; identification of data objects stored as files (using file names as identifiers or suitable alternatives); identifier systems used are based on handles (DOIs from DataCite, ePIC PIDs); persistent and unique identifiers for both data and metadata objects, and ensure availability of identification services.

The ENVRI RM version 2.0 describes data identification as “A functionality that assigns (global) unique identifiers to data contents”.

**Science Viewpoint:** PID Generator is part of the Data Publication community, while Data Identification is linked to Data Curation (possible conflict). No behaviour is proposed to support data identification.

**Information Viewpoint:** Unique Identifier (UID) is defined as an information object. Assign unique identifier is defined as an information action. There is no state linked to identification of data or metadata. The definition may need a state to indicate whether the identifier has been assigned. An object may have more than one identifier associated with it.

**Computational Viewpoint:** PID service is defined as an external service to provide identifiers for data objects and to resolve objects. This service should provide two interfaces: acquire identifier and resolve identifier. The PID service is linked to the data use phase. The PID service is used in brokered data import, processed data import, citation, and raw data connection.

The following are the recommendations for enhancement:

- Correct the model at science and computational viewpoints.
  - Science Viewpoint should accommodate identification within the curation community both in roles and in behaviours.
  - Computational Viewpoint place this functionality at the use phase. Move it to the curation phase.
- Specific types of identification should be included as alternatives in the technology viewpoint specification. This should include use of handles (DOI, PID, or other similar) and strategies for using file names and other types of identifiers (e.g. URI).

### 6.6.2 Changes made to implement recommendations

The following changes have been made to implement the above recommendations.

A new behaviour and a new role have been added to the SV curation community. The PID Generator and PID Registry roles have been removed from the SV publishing community, and combined into a single role of PID Generator in the SV curation community.

**PID Manager:** A passive role, a system or service that assigns persistent global unique identifiers to data and metadata products. The Manager invokes an external entity, the PID Service, to obtain PIDs. The Manager maintains a local catalogue of PIDs that are being used to reference data and



metadata. If the data or metadata in the RI change location or are removed, the PID Manager updates this information locally and informs the PID Service.

**PID Generator:** A passive role, a public system or service which generates and assigns persistent global unique identifiers (PIDs) to sets of digital objects. The PID Generator also maintains a public registry of PIDs for digital objects.

**Data Identification:** A behaviour performed by a PID manager which assigns a PID for data and metadata being curated.

For provenance reasons (6.4 above), all derived data products must be identified. This includes both finalised products and intermediate ones. However, not all intermediate products are suitable for publishing. It is necessary to keep track of the data products but only those meant to be published require a public persistent identifier. Other data products (i.e., intermediate ones) may need a different type of identifier.

## 6.7 Data Citation (C.12)

### 6.7.1 Analysis and recommendation

Requirements refer to the need for data citation functionalities including: citation management; standard (homogenous) approach to citation; data citation automation; guarantee unambiguous resolution of citations; ensure credit to curators and generators of derived data products; facilitate collection of usage statistics; facilitate citation of data subsets (coupling identification with query provenance).

The ENVRI RM version 2.0 describes data citation as “A functionality that assigns an accurate, consistent and standardised reference to a data object, which can be cited in scientific publications and/or from other data collections”.

**Science Viewpoint:** There is no role defined in the SV to handle citation. The Data Publishing community defines Data citation as a behaviour performed by a PID Generator, which is in charge of maintaining a reference between data object and identifier (possible conflict or inaccuracy here).

**Information Viewpoint:** Citation is defined as an information object. There is no information action linked to citation. There is no state linked to citation of data or metadata. The definition may need a state to indicate if the identifier has been used. In the modified version of the RM a Cite Data action is defined consuming provenance data.

**Computational Viewpoint:** Citation is mentioned as one of the ways for finding data sets in CV data publishing. PID Service is defined as an external service to provide identifiers for data objects and to resolve objects. This service should provide two interfaces: acquire identifier and resolve identifier. The PID Service is linked to the data use phase. The PID Service is used in brokered data import, processed data import, citation, and raw data connection. This object is strongly related to citation resolution.

The following are the recommendations for enhancement:

- Follow the model proposed in the computational viewpoint and place this functionality at the use phase. The use phase is where citations are produced (to indicate that data has been used).
- Specific services for handling citations may be needed, especially for linking citations to provenance and resources.

### 6.7.2 Changes made to implement recommendations

The following changes have been made to implement the above recommendations.

The changes for identification described in the section 6.6 are designed to better support data citation. Data citation involves producing a reference for a data source that can be resolved externally and link to the data within the RI. The PID Manager is responsible for ensuring that the links to resources are kept consistent and informing of changes to the PID Generator. The RI can provide the template for citation of data; apart from this the citation activity is a responsibility of the data user.

## 6.8 Data Product Generation (B.5)

### 6.8.1 Analysis and recommendation

Requirements include the description of different functionalities for data product generation mostly related to standardisation of processes for data products.

The ENVRI RM version 2.0 describes data product generation as “A functionality that processes data against requirement specifications and standardised formats and descriptions”. This would support the processing of datasets enabling the automated generation of most data products.

**Science Viewpoint:** The generation of data products is part of the definition of the data curation community. The Data Curator role is described as responsible the generation of data products. Data product generation is a specific behaviour to be performed by the data curator.

**Information Viewpoint:** No information object is linked to data product generation explicitly. Process data is defined as an information action which has data product generation as one of its purposes. There is no state linked specifically to data product generation. In the modified version of the RM a Convert Data action is defined to allow generation of data products (not part of the original RM, but introduced to indicate the possibility of deriving new data products outside of the RI).

**Computational Viewpoint:** Process Controller is defined as a processing service to support the generation of data products. This service should provide one interface specifically for delivering data products: deliver dataset. The process controller is used to process data import and data staging.

The following are the recommendations for enhancement:

- Data process manager could be a new role in the SV. The role could be a passive role (i.e. software agent implements it); this would correspond directly to the process controller service defined in the CV.
- New data products are created at all stages of the data life cycle. For instance, backup, metadata, provenance, and other such data objects are data products. In a sense the creation of data products is already accounted for in the RM. A discussion about the pertinence of defining an explicit list of expected data product types would be helpful but not urgent.

### 6.8.2 Changes made to implement recommendations

The following changes have been made to implement the above recommendations.

Data product generation is described as the result of processing and use activities. In the IV, the data use diagram identifies the production of new data products. Product generation is closely



related to provenance (see 6.4) and revises all stages at which products are expected to be generated and has repercussion to all viewpoints.

## 6.9 Data Processing Control (D.9)

### 6.9.1 Analysis and recommendation

The high count of data processing comes mostly from the optimisation section of D5.1. However, the requirements are not uniform and vary on reach and specificity. The specific requirements from RIs are also broadly stated.

The ENVRI RM version 2.0 describes data processing control as “Functionality that initiates the calculation and manages the outputs to be returned to the client”.

The three viewpoints offer several components that support processing at various levels.

The following are the recommendations for enhancement:

- Distinguish processing control clearly to highlight that the ENVRI RM describes the mechanisms that support producing new data products.

### 6.9.2 Changes made to implement recommendations

The following changes have been made to implement the above recommendations.

The data processing phase in the CV has been redefined to describe how processing control and coordination interact to produce new data products. The diagrams already existed in the original model but have been re-described as integration points.

## 6.10 Data Use (E)

### 6.10.1 Analysis and recommendation

Data use requirements in general refer to concerns about usability, user support, configurability, service provision, portals and virtual environments.

Functionalities in the ENVRI RM version 2.0 at use level are mostly linked to authentication, authorisation, account management and identification. Nevertheless, the three viewpoints offer several components that support data use at various levels.

The following are the recommendations for enhancement:

- Highlight the importance of data use in supporting the interaction of different types of users with the RI during different phases of the data life cycles.

### 6.10.2 Changes made to implement recommendations

The following changes have been made to implement the above recommendations.

The CV presentation of the data use phase and the associated subsystems has been redesigned to highlight the importance of human interaction interfaces provided by presentation objects and program interfaces provided by service objects. The three viewpoints have been updated to support data use.



## 6.11 Data Publication (C.11)

### 6.11.1 Analysis and recommendation

The requirements for data publication cover a wide range of subjects that include publication of data, standardisation of processes, standardisation of formats, specialisation for specific data products, retrieval of data, and identification.

The ENVRI RM version 2.0 describes data publication as “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 publicly accessible or to those who agree to certain conditions of use, and to individuals who meet certain professional criteria”. This would support the publishing of datasets in wide terms.

The data life cycle, the Science Viewpoint, the Information Viewpoint, and the Computational Viewpoint all support the publishing of data

The following are the recommendations for enhancement:

- No new publication requirements detected so no modifications are recommended from this set of requirements.

### 6.11.2 Changes made to implement recommendations

Not applicable.

## 6.12 Data Discovery and Access (C.14)

### 6.12.1 Analysis and recommendation

The requirements mapped to data discovery and access point out to improvements on discovery related to better search and access.

The ENVRI RM version 2.0 describes data discovery and access as “Functionality that retrieves requested data from a data resource by using suitable search technology”.

**Science Viewpoint:** The data publication community is defined to support discovery and access to data. The viewpoint defines roles and behaviours in line with this definition.

**Information Viewpoint:** The Information Viewpoint defines objects and actions that support discovery and access.

**Computational Viewpoint:** The Information Viewpoint defines objects and the interfaces that support discovery and access to data.

The following are the recommendations for enhancement:

- No new discovery and access requirements identified so no modifications are recommended from this set of requirements.

### 6.12.2 Changes made to implement recommendations

Not applicable.



## 7 Moving towards engineering and technology

Adding Engineering (EV) and Technology Viewpoints (TV) to the present three RM viewpoints (Scientific, Information, Computational) provides the ENVRI community with a reference architecture specified in engineering and technological terms. Together, the five viewpoints span all that is needed from science-oriented to technology-oriented views of what must be built. They provide how different RIs can find a common basis for interacting together.

We separate between engineering and technology because different parts of a single RI (and different RIs) can each be built using different technologies, and because technologies change over time. Taking an engineering view provides us with a technology independent understanding of how the overall RI system is constructed (architected). Separating the two aspects assists also with maintenance and transition planning and when necessary, conformance testing.

A complete reference architecture provides the basis against which existing RI designs can be evaluated. It (or parts of it) can be adopted as a starting point and adapted by any RI as that RI's concrete architecture – either as a completely new design basis or as a target towards which longer term convergence can aim. It helps to comprehend the existing situation in higher level terms and to identify the points at which interoperability between RIs can be achieved.

Care must be taken not to confuse engineering decisions with technology choices<sup>13</sup>.

### 7.1 Engineering Viewpoint design approach

The Engineering Viewpoint is concerned with transparently distributing computational objects (from CV) over nodes, either physical and/or virtual. It is concerned with the structure of the nodes, including structures needed to support data objects (from IV), and for the specification of the communicating channels between nodes. It is concerned also with supporting non-functional requirements relating to performance, reliability, load-balancing, etc.<sup>14</sup>

In specifying the Engineering Viewpoint (EV) in the ENVRI RM, the main aim is to control and constrain the architectural style to be adopted at interaction points between RIs to improve the interoperability between RIs; without the need of investment in expensive bilateral transformation / interworking / gateway functionalities. The aim is NOT to control the internal engineering of individual RIs, and thus less emphasis will be given to these aspects when specifying the EV.

The Engineering Viewpoint can be presented as a set of templates (patterns) that assist with deploying the objects and methods specified in the Information and Computational Viewpoints onto the technology platforms<sup>15</sup> selected/recommended by the Technology Viewpoint.

We mentioned in section 3 above the use case to integrate the DASSH Archive of Marine Species and Habitat Data with the EMBRC RI. DASSH data needs to be made accessible through EMBRC.

The constraints for such a typical integration include that:

- Metadata associated with the DASSH archive is specified according to a nationally accepted metadata standard (MEDIN) and must be transformed to or interworked (brokered) with the format used by EMBRC;
- Re-use of DASSH data by users of EMBRC should be notified to DASSH, perhaps as citation counts, and/or data downloads, and/or data views/consultations; and,
- Transactions should be automated as much as possible, for example mapping automatically as data and metadata are exchanged between DASSH and EMBRC.

<sup>13</sup> A well-known example of this is the engineering decision to use a database rather than a files structure for storing data. A technology choice is what kind of database to use (SQL/relational, NoSQL, etc.) and the choice of specific product.

<sup>14</sup> For an introduction to the Engineering and other viewpoints of ODP, refer to [Linnington 2012].

<sup>15</sup> For example, e-Infrastructure platforms such as EGI FedCloud, EUDAT, HelixNebula Science Cloud; or purely commercial platforms such as Amazon Web Services, Microsoft Azure, etc.





We could select a third-party service from one of the independent e-Infrastructure providers (e.g., EGI.eu, EUDAT) to accomplish this integration but each service probably has its own characteristics and modes of use. Manual work is needed to set up the integration and it is difficult to change from one provider to another when, for example Service Level Agreements are not met, or when a better solution comes onto the market.

We can be more specific, using capabilities of the ENVRI RM to clearly define the exact functionalities needed for integration. Not only that, but these functionalities can be provided directly by one or both RIs in agreement with one another; or they can be outsourced to a third-party offering a service in accordance with the template or pattern. Then the choice becomes one of decision based considerations of competing value offerings, Service Level Agreements, and other non-functional aspects rather than on functional aspects. In other words, patterns help to harmonise functional capabilities.

This is illustrated in Figure 3 (seen from the perspective of a Data Archive in  $RI_A$  wanting to integrate with and be part of the Virtual Laboratory of  $RI_B$  to reach a wider user base).

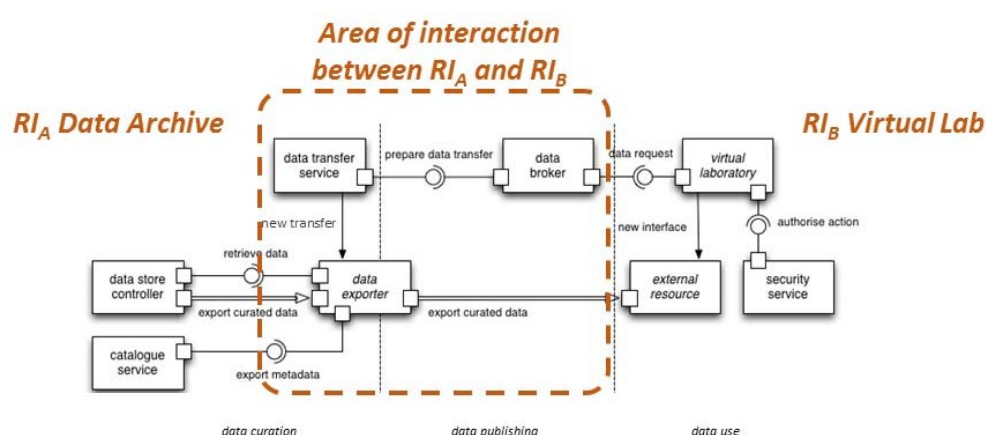


FIGURE 3 BROKERED IMMEDIATE DATA EXPORT FROM DATA ARCHIVE ON REQUEST FROM USER(S) IN A VIRTUAL LABORATORY

Alternatively, as illustrated in Figure 4 ( $RI_B$  wanting to pull in and archive data such as the data generated in a Virtual Laboratory of  $RI_A$  to enhance  $RI_B$ 's own value offering).

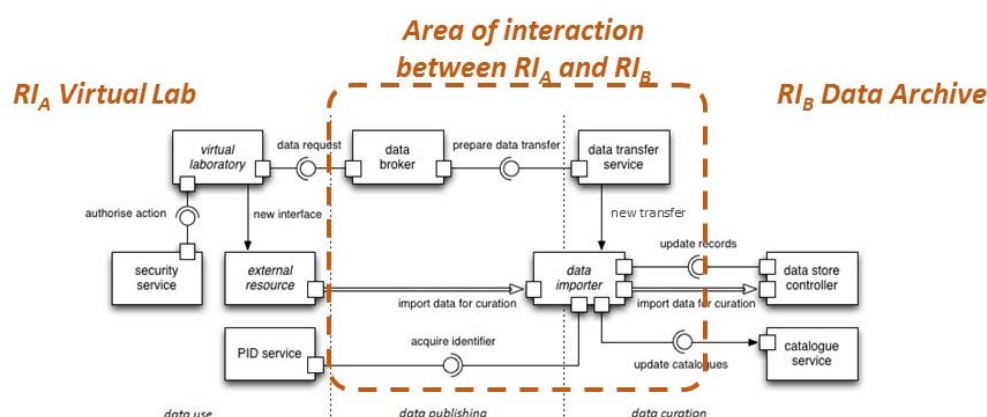


FIGURE 4 BROKERED DATA IMPORT FOR CURATION IN A DATA ARCHIVE

We see a similar pattern in both illustrations for brokering between the two pieces of infrastructure. In the first case (Figure 3)  $RI_A$  offers a brokered data export pattern composed of

three main objects (data broker, data transfer service and data exporter<sup>16</sup>) that allows RI<sub>B</sub> to make requests to use RI<sub>A</sub> data in RI<sub>B</sub>'s virtual laboratories. In the second case (Figure 4) RI<sub>B</sub> deploys a brokered data import pattern, again with three objects<sup>16</sup> that allows new data produced in virtual labs of RI<sub>A</sub> to be curated and preserved by RI<sub>B</sub>. In both patterns the broker and transfer service are common, while the importer and exporter services are likely to be mirrors of one another.

The essential aspect of this example is that there is an area of interaction, which we name as an 'interaction point' between two pieces of infrastructure at which a specific engineering pattern can be applied to achieve interoperability between the two parts.

#### **Definition of term**

**Interaction point:** An interaction point is a point between two parts of an architectural design where an engineering pattern can be applied to achieve interoperability between the two parts. An interaction point can appear within a single RI, between a pair of RIs, or between an RI and a third-party.

An engineering pattern can be:

1. Implemented jointly by agreement between the two RIs;
2. Offered by one RI to another as a service to which the other RI can bind and exploit;
3. Offered as a third-party service<sup>17</sup> that one or both RIs can bind to and exploit i.e., outsourcing.

In all three cases, physical interfaces (which must be specified in technology terms) are implied. This will usually mean a defined exchange format and specific protocol operating at an interface, most likely being presented at this level in terms of an Application Programming Interface (API). Best practice guidelines (to be found in the TV) will recommend that such APIs are simple, uniform and deterministic, with low atomicity i.e., they each represent one (micro)service that does one thing and one thing only.

In forming the set of templates or patterns for the Engineering Viewpoint, there are several considerations to keep in mind:

1. The viewpoint should recognise any engineering constraints arising from what the RIs are already committed to;
2. The need to minimise the number of objects (from the CV) that are treated as logically central (i.e., behave as a single instance with a consistent state); because that is expensive to maintain;
3. The viewpoint should focus on the interaction *between RIs* (i.e., *focus on interoperability*). Patterns should act as a roadmap for convergence around those<sup>18</sup>; and,
4. Points of interaction and integration should revolve around:
  - a. Interfaces (APIs) and protocols across interfaces (e.g., for data exchange, for initiating activity, for authentication, authorisation and accounting (AAA), for social collaboration, etc.).
  - b. Metadata definitions and descriptions of assets (data, tools, resources) as the exchange format.
5. Where reliability and elasticity are needed.

Note that the same effect or end-result can be achieved using one of several alternative patterns. Thus, an appropriate choice of pattern will need to be discussed and agreed to achieve effective interoperation between different RIs.

<sup>16</sup> Broker service negotiates query and transfer requests. Transfer service moves the data. Exporter / Importer service(s) transform data from/to structure the RI needs.

<sup>17</sup> By third party providers of digital infrastructure for research (DI4R) such as EGI.eu, EUDAT, etc.

<sup>18</sup> This criterion effectively says that it's okay to treat what goes on inside each RI either side of an interaction / integration point as 'black box' where it is expedient to do so.



### 7.1.1 The Engineering Viewpoint subsystems

From the Engineering Viewpoint perspective, the ENVRI RM identifies five subsystems. The EV subsystems are based on the notion of data life cycle phases evident from investigations of all existing RIs. Each subsystem provides 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. However, RIs can decide the coupling and distribution of the subsystems based on their individual requirements. The definition of subsystems will take components from the CV and objects from the IV and arrange them in different configurations. The existence of a component in one subsystem does not limit its use to that subsystem, but highlights the relevance of the component in the subsystem concept

**Data acquisition subsystem:** provides objects destined to collect data from registered sources to store those data within the infrastructure<sup>19</sup>.

**Data curation subsystem:** provides objects that guarantee the preservation and usability of data within the infrastructure.

**Data publishing subsystem:** provides objects which enable discovery and retrieval of scientific data subject to authorisation.

**Data processing subsystem:** provides objects for performing a variety of data processing tasks.

**Data use subsystem** (also known as community support subsystem): provides objects that support internal and external users of an infrastructure in their interactions with that infrastructure.

In addition to clearly identifying the five subsystems supporting the data life cycle, the EV will also identify supporting subsystems that provide service objects to all of them. For instance, AAA services are required by all subsystems, and consequently are candidates to be in a crosscutting subsystem that provides AAA functionalities across all the other subsystems. The objects providing provenance functionalities are also required by all subsystems, also making them a good candidate to be grouped as part of an orthogonal subsystem with respect to the five basic subsystems.

Engineering subsystems and their major components can be logically independent. However, they can be co-located in operational deployments to improve performance or reduce costs.

### 7.1.2 Interaction points

Subsystems can be composed in different ways by RIs to support different tasks across the phases of the data life cycle. Examination of the interactions between subsystems exposes a set of possible interfaces to be derived. For each of these interfaces, the interaction between the subsystems can be specified in terms of protocols that define the subsystems' behaviour in relation to one another. The EV will define a set of interoperability patterns based on the interaction points between EV subsystems, as illustrated in Figure 5.

---

<sup>19</sup> Note also that in many communities there is also a pressing need for *ad hoc* ingest of bulk data from pioneering projects, as well as temporary deployments or accumulations established by RIs operating elsewhere or outside the governance regime. These are vital for a full coverage, for historical data and for incorporating pioneering work. But they are rarely given enough thought today.

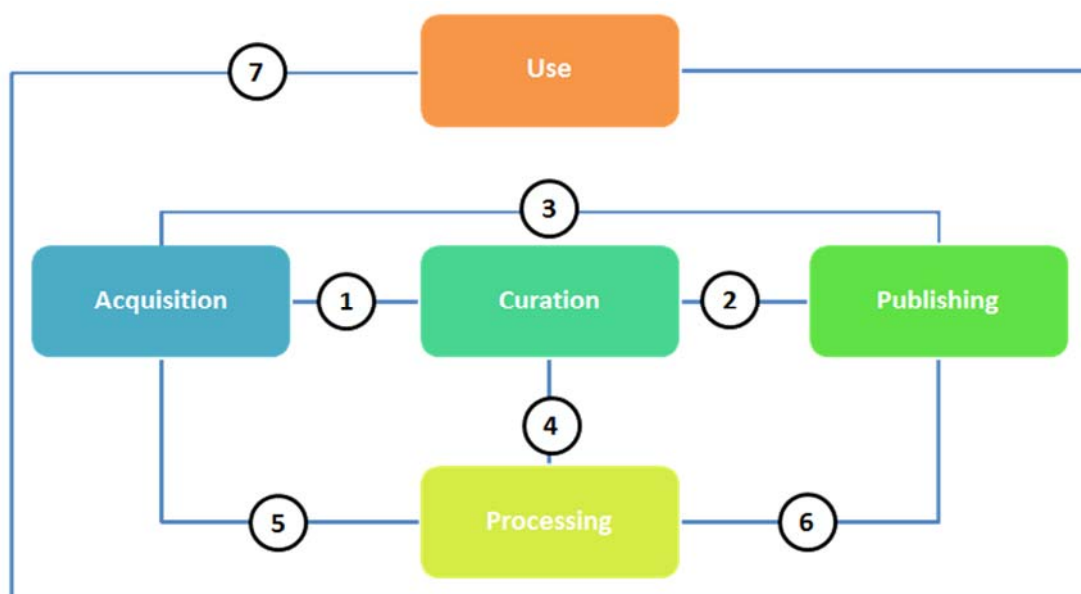


FIGURE 5 POTENTIAL INTERACTION POINTS BETWEEN EV SUBSYSTEMS

These interactions can occur between subsystems provided by a single RI, but they also allow integration of components provided by third parties. There are ten interaction points. Note that in the illustration of these points in Figure 5, points 8 – 10 appear at the same place where point 7 is shown.

1. **Acquisition - Curation:** integrate data acquisition objects with data curation objects.
2. **Curation - Publishing:** integrate data curation objects with data publishing objects.
3. **Acquisition - Publishing:** integrate data acquisition objects with data publishing objects.
4. **Curation - Processing:** integrate data curation objects with data processing objects.
5. **Acquisition - Processing:** integrate data acquisition objects with data processing objects.
6. **Processing - Publishing:** integrate data processing objects with data publishing objects.
7. **Use - Acquisition:** integrate data use objects with data acquisition objects.
8. **Use - Curation:** integrate data use objects with data curation objects.
9. **Use - Publishing:** integrate data use objects with data publishing objects.
10. **Use - Processing:** integrate data use objects with data processing objects.

The architectural style adopted is that of the brokered, service-oriented architecture (brokered SOA)<sup>20</sup>. Furthermore, the style recommended is the style of shared standardised micro-services APIs as the engineering mechanism for interfaces between RIs.

### 7.1.3 Identification of core competencies

The ENVRIplus project has analysed the requirements of twenty environmental research infrastructures [Atkinson 2016]. The core competencies of these RIs are aligned to the data life cycle and define the subsystems that each research infrastructure implements. Figure 6 shows the alignment of research infrastructures to the subsystem view supporting the data life cycle.

<sup>20</sup> Brokered SOA is style of Service Oriented Architecture in which a registry of services (the broker) acts as the point (catalogue) where all services within the scope of the architecture are registered and discoverable by users and other services.

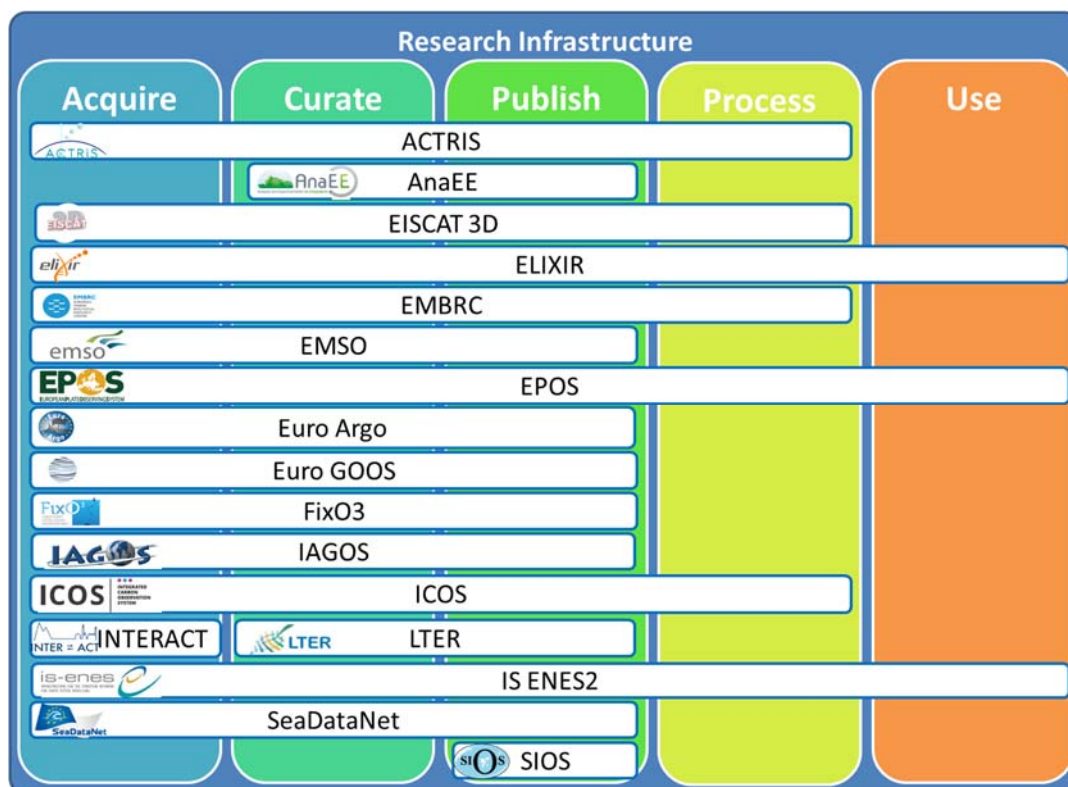


FIGURE 6 IDENTIFYING CORE COMPETENCIES OF ENVRI PLUS RIs

Within this view, the degree of support of RIs for each life cycle phase and the kind of data products they produce varies. For instance, several of the marine RIs support Acquisition, Curation and Publishing of Data. However, they focus and specialise on different phases of the data life cycle. Some are more focused on acquisition of raw data, others on collecting and curating data from different sources, yet others will collate and process data from different domains.

FixO3, EuroFleets, Jerico, ESONET, EMSO, and EuroArgo are focused on management of acquisition of raw data from different types of marine observatories, fleets and device networks. The main competencies of these RIs are in the acquisition phase. The types of data products they make available are commonly closer to raw data than to processed or derived data.

SeaDataNet focuses on integrating data from different types of observations into a consistent format to provide a more complete picture. In this case their main competencies are in the curation and publishing phases. The types of data product produced by these RIs are more complex and imply some degree of processing.

Beyond this, other RIs such as EMBRC and SIOS aim to integrate and use data from multiple disciplines. Their main competencies are processing and use.

The Engineering Viewpoint can be used to support the integration of RIs following the concept of core competencies, subsystems and interaction points.

#### 7.1.4 Alignment with existing viewpoints

Following the ODP model, there are three types of Engineering Viewpoint object: Basic Engineering Objects (EV BEO), Container Objects (EV Container) and Channel Objects (EV Channel). EV BEO objects map one-to-one with objects from the CV. EV Containers are used to group EV BEO, and Channels are used to connect EV BEOs across containers.

In the ENVRI RM, the CV objects are generic and can be used by any subsystem. In the EV, CV objects can be specialised to better describe a fully working subsystem. In practice, RIs may already have defined their systems and the division between subsystems may not be obvious. In this case

the subsystems will be viewed as a logical rather than a physical grouping of objects; and can be used to help identify practical interfaces for interoperating between subsystems in different RIs. The different types of EV Containers allow the definition of containers that are physically distributed and containers that are logically distributed. Subsystems can therefore be tightly coupled or extremely loose. Similarly, they can also identify parallelism with large shared memory (important for cross-correlation algorithms on large data) and fast-interconnected clustered essential when simulating systems with effects coupling over a wide range of different scales, e.g., those described by stiff differential equations. Thus, they capture critical domain knowledge about the critical issues in a domain's scientific methods that have to be "engineered in". The definition of channels will depend on the distribution model selected.

### 7.1.5 Recommended architectural style and sector trends

The recommended ENVRI RM architectural style is SOA, which helps in the deployment of multiple subsystems of services aimed at supporting the research data life cycle. Tiers within the SOA are already defined in the CV of the ENVRI RM and can be carried over into the EV. The Engineering Viewpoint will refine and specialise CV objects to clearly describe each subsystem, and identify interfaces by which the subsystems can be seamlessly interfaced to form a coherent RI or to support interoperation across RIs. Currently, most RIs are built using a tiered architecture, with some limited use of SOA as service clients. Outside of the ENV RI domain there are a couple of examples of infrastructures as service providers. GBIF and EMBL are the most well-known, acting as service providers through service APIs. At the same time, most e-Infrastructures are moving rapidly towards the Cloud Computing Architectural Model. RIs need to become aware of and prepared to take advantage of the services offered by e-Infrastructures when it is appropriate and cost-effective to do so. According to industry<sup>21</sup>, SOA is a good model to ease and support the transition to cloud infrastructures. The EV of the ENVRI RM will illustrate the pathways that facilitate moving the RIs into a SOA model and then use this model to provide and connect to cloud based services.

The use of a brokered SOA architecture also aligns the CV and EV of the ENVRI RM with the current work of the RDA Data Fabric Interest Group (DFIG) [RDA 2016b], as well as with GEOSS brokering principles. The DFIG envisions a Digital Object Cloud (Figure 7). In this vision, end users, developers, and automated processes access persistently identified and consistently structured digital objects. The objects are securely and redundantly managed in the Digital Object Cloud. The Digital Object Cloud is an overlay that exposes data contained in a variety of information storage systems [Lannom 2016].

---

<sup>21</sup> e.g., Cappemini/HP [Cappemini 2008], Oracle [Kress 2014], IBM [Kreger 2014].



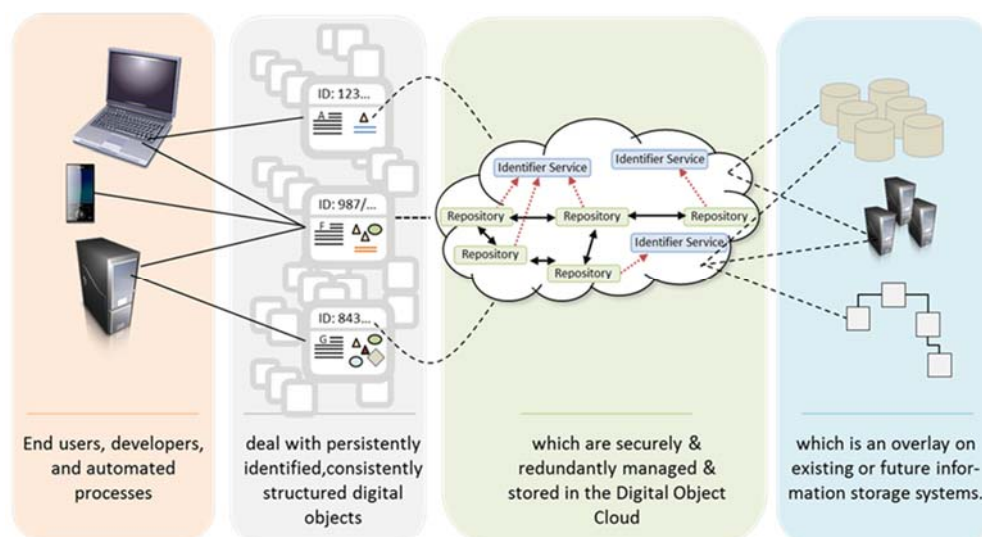


FIGURE 7 THE DIGITAL OBJECT CLOUD

The RDA DFIG is considering the role of brokers in this perspective. Case-by-case mediation has been done for many years by the diverse communities. The broker approach, which is now evolving towards being metadata driven, has been advanced as an alternative to perform mediation in an automatic, more effective (i.e., multi-disciplinary, one-to-many), sustainable, and re-usable way [Santoro 2016, Nativi 2015]. In this context, the independent RIs appear as the repositories providing access to the digital object collections envisioned within the digital object cloud (Figure 8).

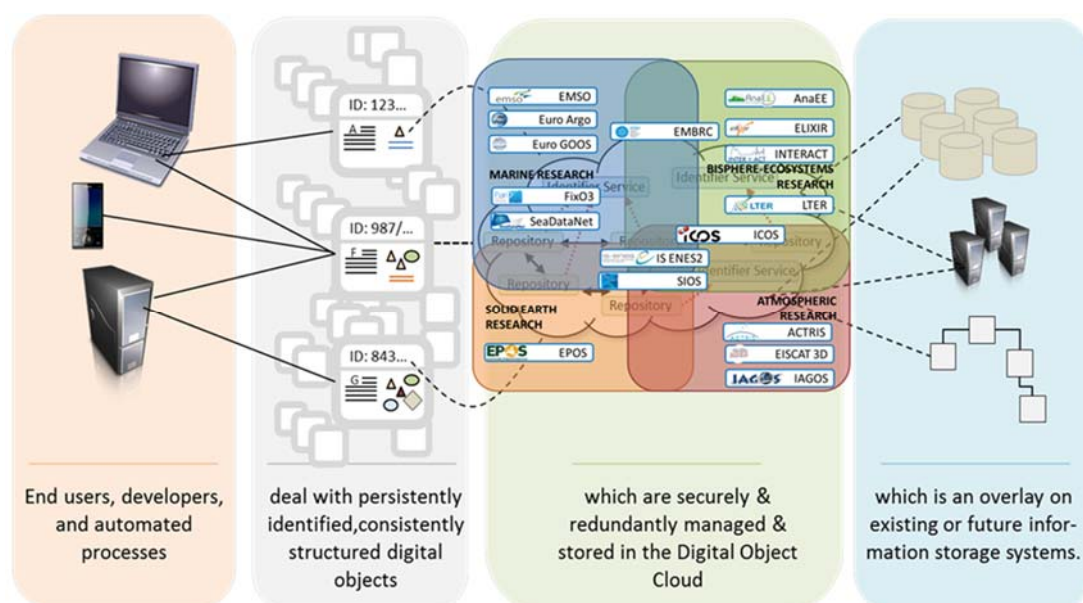


FIGURE 8 ENVRI RIs WITHIN THE DIGITAL OBJECT CLOUD

## 7.2 Approaching the Technology Viewpoint

The Technology Viewpoint (TV) provides concepts and constructs to specify the exchange formats, interface protocols, standards, hardware and software products from which RIs can be built or by which they can interoperate. The TV can also define tests to verify that such implementations comply with the specification as prescribed by all viewpoints. The TV specifies the plans and processes for selection, acquisition and evolution of the system parts during the lifetime of the RI.



The SV, IV, CV and EV provide technology independent models of the RI at different levels of abstraction. The challenge of the TV is to specify the set of technologies and standards to be used. In addition to this, in the context of the European Open Science Cloud (EOSC) [EOSC 2016] those implementations must foster the use of shared European Infrastructures and facilitate seamless integration of RIs. Such e-Infrastructures can serve as a model for the basic technologies to be considered in line with the aims of the RIs and the identified ENVRI RM competence areas. The following sub-sections outline the main objects to be defined in the TV, the correspondence of the TV with the other four viewpoints, and the relationship of the TV with the architectural and engineering models.

### 7.2.1 Main objects of the Technology Viewpoint

The Technology Viewpoint defines four main concepts: technology objects, implementable standards, implementation and IXIT (Implementation eXtra Information for Testing). The Technology Viewpoint describes the implementation of an RI in terms of a configuration of technology objects. Technology objects represent the hardware and software components used for the implementation, and the interfaces linking them.

There are multiple constraints on the selection of the technologies and standards that can be used for building an RI. The ENVRI RM must identify constraints derived from four areas: the environmental research area and its previous investments and culture; core competencies of the RI, including any commercial constraints from business related to the RI; the coupling and distribution of the subsystems; and the use of third-party components.

The targeted environmental research area constraints determine the standards that can be used. Some voluntary standards, such as those from OGC for describing geospatial information are common to many research areas, while others are specific to specific research areas. For instance, Darwin Core is used in the biodiversity field to register data about species occurrences. Requirements also arise from regulatory constraints, such as those of the INSPIRE Directive [INSPIRE 2007]. Many domains are constrained by existing legacy systems, often operating on the basis of long-established global agreements and sector specific standards e.g., domain specifications of NetCDF file content. Skills and experience with these systems are critical elements in the operational practices and innovation paths of an RI.

The definition of core competencies will determine the main components to be considered in the building of the RI. If the planned RI is aiming to support the entire data life cycle from acquisition to use, the number of components to be considered is substantial. On the other hand, when only a limited subsystem is considered for interoperation with another RI, the number of components might be quite small. The choices available for implementing each subsystem can range from design and construction of bespoke systems, through use of common off-the-shelf parts, to the outsourcing or acquisition of components or services from third-parties.

The coupling and distribution of the subsystems produce a set of constraints that dictate the location of subsystem resources. These constraints are closely linked to the decisions made in terms of developing, acquiring or outsourcing.

The use of third-party resources generates constraints that affect the selection of providers. In this case the variability can be from acquisition of assets to be maintained by the RI team to the use of shared resources in a cloud environment.

### 7.2.2 Correspondence with the other four viewpoints

The main principle in the design process from the high level of the SV to the low level of the TV is the proper identification of the phases or sub-parts of the research data life cycle of concern. As stated earlier, the SV, IV, CV and EV provide technology independent models of the RI, its parts and interfaces at different levels of abstraction. The selection of components in the TV is guided by those abstract models. The grouping of EV BEO and the architectural model provided by the CV



affect the selection of components and the criteria to determine that the selected components are suited for the needs of the RI. The definition of CV objects is in turn affected by the SV and IV models.

### 7.2.3 Relation of the TV to architectural and engineering models

The definition of the RI in the TV must interact strongly with the CV and EV specification. The selection of technologies and standards has consequences that can lead to breaking some of the CV and EV models. Specifying the TV starts a cyclical process that that can lead to redefinition of CV and EV models.

The ENVRI RM will propose the introduction of '**configuration points**' in the TV to facilitate this process.

#### **Definition of term**

**Configuration point:** A configuration point is a point in the RI design/architecture at which a designer/architect must make decisions between one technology choice or another, or between one third-party service solution or another.

Configuration points exist where alternative technological approaches can be chosen. For instance, the selection a Cloud based data storage facility for curated data will require interfaces to be provided for transferring curated data from the RI curation subsystem to the external cloud storage subsystem. The RI should also consider how to provide access to the data when it is published, for instance defining constraints on use, access, and citation.

A lot more thinking, planning and detail is needed before much of the TV can be specified. It needs a deep comprehension of the technology options and the ways in which they can be bundled together in middleware. Initially, therefore the TV will be populated in specific areas where critical community level decisions are needed.

## 8 Future enhancements to the Reference Model

The current version (v2.1, November 2016) of the ENVRI RM is in an advanced state to support the creation of high level models for designing RIs, parts of RIs and the patterns and interfaces between them to bring about further integration and interoperability. Moreover, when complemented with the definition of EV and TV, the ENVRI RM will also be a tool to facilitate the reuse of common patterns and infrastructures.

The following sub-sections describe possible further enhancements that can increase the usefulness of the ENVRI RM.

### 8.1 Pending issues from T5.1 requirements analysis

Several areas where further attention to the RM will be needed in the future have been identified by the Task 5.1 requirements analysis [Atkinson 2016] and from project internal discussions/decisions. These include, in no specific order providing support for:

- Definition of a requirement for provenance and corresponding support in the RM for provenance tracking components (e.g., in the SV, IV and CV);
- Non-functional requirements;
- Research campaigns;
- Adding support for canonical metadata; and,
- User-user interactions.



### 8.1.1 Provenance Tracking

The SV presents provenance as an activity that parallels the activities in the data life cycle, registering the outcomes of each phase as important provenance elements. Definition of the corresponding roles and behaviours in the SV, as well as the corresponding objects in the CV is still pending; although objects and related action types are already defined in the IV. In addition, better understanding of the inter-RI data provenance and trace mechanisms to be supported must be developed. This work is presently being carried out in Work Package 8 of the ENVRIplus project but is not due to conclude for some months. Support is likely to be added in a future version of the RM.

### 8.1.2 Non-Functional Requirements

Non-Functional Requirements (NFR), ranging from performance and cost, to ease of use, to rights and obligations is an important area that needs to be further addressed. Table 13 of deliverable D5.1 presents a summary of NFRs as follows:

- Cost (funding), mentioned by 10 RIs;
- Long-term data preservation, mentioned by 2 RIs;
- Privacy, mentioned by 3 RIs; and,
- Security, mentioned by 3 RIs.

NFRs need to be expressed more explicitly because D5.1 does not clearly identify them as functional requirements dominated discussion at that time.

Nevertheless, in the future the ENVRI RM needs to consider the NFR to be supported and how. Ultimately, NFRs must be expressed in a formal syntax with declared semantics so that they are machine actionable (i.e., can be taken into account when deployment choices across platforms are made).

### 8.1.3 Research campaigns

Some RIs are built to serve a research campaign; many are used by research campaigns, and some will initiate research campaigns to improve their effectiveness or efficiency. We define a ‘research campaign’ as a co-ordinated, resourced and sustained effort to achieve a recognisable research goal. Research campaigns can be conducted both within a single organisation or (of more interest from the RIs and RM perspective) can involve multiple organisations. They can encourage *independent* rival teams to validate results, as in the pursuit of the Higgs boson. More details and examples may be found in Appendix 5, page 81.

By considering research campaigns, RIs and the RM will analyse and consider topics that have and need long-term stability; thereby increasing the chances of sustained value for RIs, the RM and the recognised commonalities. Incorporating explicit support into the RM would be helpful as these are processes important to individuals, teams, and projects with extended duration, multiple participants and research data cycles.

Appendix 5 provides an initial account of the critical features of a research campaign and suggests an approach to their accommodation in the RM. The terms in Appendix 5 that are in *italics* should be considered as candidates for inclusion in the RM. Consultation with RIs and with those leading research campaigns will be needed to refine this initial view of research campaigns using the RM as a framework. Once the critical concepts of a research campaign have stabilised, many of these will relate well to existing concepts in the SV and IV of the RM. Some extensions will be necessary for those viewpoints. The impact on other viewpoints is expected to be small.



### 8.1.4 Adding support for canonical metadata

An ENVRIplus project decision (Zandvoort, May 2016) has been taken to use a canonical core metadata model as a logical point of interoperability between RIs. There has been a proposal, presently being documented in deliverable D5.4<sup>22</sup> to base this on the CERIF metadata standard for research information [CERIF 2013].

This canonical metadata model approach must be further assessed from the perspective of how to embed support for it into the ENVRI RM.

### 8.1.5 User-user interaction

A key aspect of research is intercommunication between researchers while they work on datasets, analysis tools, experiments, etc. The formal channels are publication and citation but increasingly there is rapid and fluid user-user communication arising from the use of electronic communications including social media. This should perhaps be supported. Requirements in this area are inferred but not explicitly mentioned in deliverable D5.1 [Atkinson 2016], although they are known to be identified elsewhere e.g., in recent and current Virtual Research Environment projects.

## 8.2 Other potential enhancements

### 8.2.1 Scientific Workflows

The requirement to support integration of scientific workflows within the RI ICT systems has been identified as part of the minimal model from the first version of the ENVRI RM. However, the coverage of scientific workflow support has not been adequately addressed so far in the RM. A useful recent survey on this topic is [Liew 2016]. Several further enhancements to the SV, IV, and CV are proposed:

- Add a workflow management system role to the data processing community in the Science Viewpoint, as well as the behaviours associated with workflow management: composition, validation, optimisation and execution.
- Add a workflow information object to the Information Viewpoint, as well as the corresponding actions to support the activities described in the SV.
- Add a workflow management system computation object in the Computational Viewpoint that will correspond to the workflow management system role of the IV and act on the workflow information objects of the IV.

Similarly, the definition of EV and TV will include the integration of workflow management systems within the processing subsystems.

### 8.2.2 Data management plan

Funding bodies increasingly require grant-holders to develop and implement Data Management Plans (DMPs) [DCC 2016]. Plans typically state what data will be created and how, and outline the plans for sharing and preservation, noting what is appropriate given the nature of the data and any restrictions that may need to be applied. Some RIs representatives have expressed the view that it would be helpful when the ENVRI RM provides clear support for representing DMPs and the requirements specified in them.

---

<sup>22</sup> Deliverable D5.4 “A Development Plan for Common Operations and Cross-Cutting Services based on a Network of Data Managers and Developers”.

Inclusion of support of DMPs in the RM must be considered carefully because support will be required at multiple levels – at the overall RI level, and at the level of individual projects that make use of the RI, for instance.

An overarching DMP can be applied to an RI in its entirety. In this it plays roles both in: i) the overall definition of the RI in the first place; and ii) during the daily operation of the RI.

In the first role, the RI must define a DMP that is a policy that will apply to all the processes for acquiring, curating, publishing and using data. The DMP can be defined as a control i.e., a constraining policy in the SV and IV that ensures that the acquired data meets precise standards, such as a data definition and data integrity constraints. The control process should also ensure that data will be processed, stored and identified within the RI infrastructure following a concrete procedure. The control process will also determine and enforce the access constraints for the data, including guarantees on availability, consistency, and secure transport. Finally, the control process will define clearly duration of the data in the system, and the related disposal procedures. This type of control process can define as an additional crosscutting process, which needs to be implemented, for example using the provenance subsystem.

In the second role, the DMP control process is monitored using the provenance system. This can help in facilitating the verification of compliance with the DMP.

During the operation of the RI, the integration and clear identification of the DMP control process can be extended to help users of the RI in defining the support of their independent, project level DMPs. For example, an RI can define a DMP in line with activities of the data life cycle and define data states to be stored in the provenance records that later can be used for verifying compliance.

### 8.2.3 Bulk data ingest

Autonomous studies, perhaps running over many years will themselves have acquired data internally and administered many of the data management functions associated with the typical data life cycle or similar, as depicted in Figure 2 and Table 1 (page 25 above). Bulk ingest of such data into a RI can occur in any life cycle phase, according to its state at the time.

This is not currently precluded by the RM and the RM could be enhanced to cover this case more explicitly. This is motivated by the observation that in today's system such bulk ingest is rare, and so practitioners must develop their own methods and infrastructure on each occasion they combine such data with production data.

## 9 Steps into practice – putting the RM to work

The principal target audience for the RM is the ICT experts within RIs who have responsibility for the architectural design and implementation of their RI and for its interoperations with other RIs. These persons are typically systems architects, designers, integrators and engineers. Secondary audiences include: i) research leaders steering the long-term balance of investment in each RI, who will appreciate the savings from shared solutions and the benefits of inter-operation; and ii) third-party solution or component providers wanting to understand how to shape their offerings to gain adoption.

The biggest challenge is in convincing the target audience of the value of the RM approach i.e., how it can help them and persuading them to invest their time to learn and exploit it. This can be made easier when the learning route delivers easy-to-achieve and early benefits.

Patterns, guidelines for use, training materials, learning ramps, case studies, engineering tooling are all instruments to facilitate this process. Nevertheless, making the complexity of the RM accessible and demonstrating its value at both theme expert level within ENVRIplus project and at the technical practice level in RIs is a significant challenge. Putting the RM to work in practice requires several continuing steps that include:



**Preparing e-Training / learning content:** Now that version v2.1 of the RM is published and stable, further effort will be expended (ENVRIplus Task 15.1) during the first half of 2017 to develop usable e-training / learning materials and to run a Webinar series of education activities directed towards the RIs.

**Raising awareness:** Substantial progress has been made with raising overall awareness about the RM. The RM has existed for several years already, firstly in the ENVRI project and now developed further in ENVRIplus. Feedback suggests a good level of interest and enthusiasm for the concept from senior RIs managements. Nevertheless, the level of awareness and engagement at lower levels in RIs organisations is much less.

**Demonstrating the added value of the RM:** As noted above in section 3.2, the ENVRI RM presents a unified view of the RI ICT system as a coherent whole, supporting the aims of the research community. It addresses the major types of information and communications technology (ICT) and data-oriented operations that a given RI is likely to need to support.

**Encouraging adoption and use:** Within several RIs (EISCAT-3D, EMBRC, EMSOdev, EUFAR, ICOS, LTER) a level of awareness exists and several have forayed to exploit the RM for their own design work. Perhaps the most successful of these thus far have been EMSOdev and ICOS although this has been difficult to assess. Translation into everyday working practice remains challenging.

**Creating a community of practice:** The community of people familiar with the RM is small at present. Growing the community of non-specialists users will reduce the barriers to wider adoption of the RM as a larger cohort of individuals mutually helping one another emerges. Initiatives are needed to stimulate the development of this self-supporting community of practice. In the longer-term, the groups supporting RM use and the system they establish to steer future developments needs to reach a critical mass, initiated by ENVRIplus but independent from it, so that the contribution of the RM to RI development and operations is sustained.

Appendix 6 (page 86) contains a synopsis of ideas, gathered during the third ENVRIweek project meeting that took place in Prague, 14-18<sup>th</sup> November 2016. These ideas can be used as a basis for making further progress during 2017. This will include outlining key messages to disseminate to the wider ENVRI community.

## 10 Outlook and next steps

The Environmental Research Infrastructures Strategy for 2030 aims towards a seamless holistic understanding of the Earth System through an approach that enhances technological, cultural and human capital [ERIS 2014]. Investments in each of these areas are needed to integrate instrumentation and ICTs to *“bring about interoperability between regions and between disciplines. This is crucial for systems-level science.”* To accomplish it, *“steps towards common standards are needed, such as agreeing on a joint reference model to describe infrastructure components.”* The work reported in the present document is a substantial step towards achieving these objectives, which must now find its way into the everyday parlance of the RIs.

The ENVRI RM has been derived from distillation of the functions seen today in typical RIs, and from the notion that a life cycle for research data exists and consists of multiple phases. Much of the everyday work of practitioners, perhaps as much as 70% is concerned with discovering, accessing, retrieving, checking, cleaning, adjusting, transforming data for the intended purpose of experimental analysis. To this extent, the RM formalises the elements of the methods that underpin current and intended working practices of the practitioners during such a life cycle and in using the resulting data i.e., the science dynamics an RI has or is trying to support. This must continue to be a major driver in the next generation of the RM. It can potentially be achieved via representations of the abstract workflows that implement the methods and the intent of processes i.e., the culture (training and experience) that has domain value, which must be sustained, preserved and enhanced wherever possible.





Several important areas (see section 8) regarding support for: provenance tracking, non-functional requirements, research campaigns, HES, canonical metadata, scientific workflows, data management plans, etc. imply a substantial amount of new content for future iterations of the RM. It still must be decided how best to accommodate such extensions, which expand the scope of the RM beyond the minimal model it set out to establish. Of the listed areas, support for Data Management Plans is the most valuable and beneficial area to address in the short-term.

## 11 References

- [Atkinson 2016] Atkinson, M., Hardisty, A., Filgueira, R., Alexandru, C., Vermeulen, A., Jeffery, K., Loubrieu, T., Candela, L., Magagna, B., Martin, P., Chen, Y. and Hellström, M., A consistent characterisation of existing and planned Research Infrastructures, Technical report D5.1, ENVRIplus project, 203 pages, May 2016. <http://www.envriplus.eu/wp-content/uploads/2016/06/A-consistent-characterisation-of-RIs.pdf>.
- [Candela 2011] Candela, L., Athanasopoulos, G., Castelli, D., El Raheb, K., Innocenti, R., Ioannidis, Y., Katifori, A., Nika, A., Vullo, G., Ross, G. The Digital Library Reference Model. April 2011 (PDF) Retrieved 2 November, 2016, from <http://bscw.research-infrastructures.eu/pub/bscw.cgi/d222816/D3.2b%20Digital%20Library%20Reference%20Model.pdf>.
- [Capgemini 2008] Capgemini. The Cloud and SOA. Retrieved 01/12/2016 from: [http://www.hp.com/hpinfo/analystrelations/wp\\_cloudcomputing\\_soa\\_capgemini\\_hp.pdf](http://www.hp.com/hpinfo/analystrelations/wp_cloudcomputing_soa_capgemini_hp.pdf).
- [CCSDS] Consultive Committee for Space Data Systems. Reference Model for an Open Archival Information System (OAIS). Recommended Practice, Issue 2. June 2012. Washington, DC, USA. Retrieved 13 December 2016 from: <https://public.ccsds.org/pubs/650x0m2.pdf>.
- [CERIF 2013] Eurocris. CERIF-1.5. Retrieved 2 December 2016 from: <http://www.eurocris.org/cerif/main-features-cerif>
- [Chen 2013a] Chen, Yin, et al. "Analysis of common requirements for environmental science research infrastructures." Proceedings of the International Symposium on Grids and Clouds (ISGC), Proceedings of Science (SISSA), Academia Sinica, Taipei, Taiwan. 2013.
- [Chen 2013b] Chen, Y., Martin, P., Magagna, B., Schentz, H., Zhao, Z., Hardisty, A., Preece, A., Atkinson, M., Huber, R., & Legré, Y. (2013), A Common Reference Model for Environmental Science Research Infrastructures. In: Page, B., Fleischer, A.G., Göbel, J., and Wohlgemuth, V., (Eds.) (2013), Proceedings of the 27th Conference on Environmental Informatics – Informatics for Environmental Protection, Sustainable Development and Risk Management, September 2–4, 2013, University of Hamburg, Germany. Shaker Verlag GmbH, Aachen, Germany. ISBN 978-3-8440-1676-5. p665-673.
- [DCC 2016] Digital Curation Centre, UK. Data Management Plans. Retrieved 1 December, 2016 from: <http://www.dcc.ac.uk/resources/data-management-plans>.
- [DDI 2015] Data Documentation Alliance. <http://www.ddialliance.org/training/why-use-ddi>. Accessed: 14<sup>th</sup> November 2016.
- [DFT WG – RDA 2015] DFT WG–RDA, *RDA Data Foundation and Terminology – DFT: Results RFC*. Eds. Gary Berg -Cross, Raphael Ritz, Peter Wittenburg. Date: 29/06/2015. Consulted on: 04/03/2016. Available at: <https://rd-alliance.org/system/files/DFT%20Core%20Terms-and%20model-v1-6.pdf>.
- [ENVRI RM V2.1 2016] ENVRI Reference Model V2.1, 9<sup>th</sup> November 2016. Retrieved: 10<sup>th</sup> November 2016 from <https://wiki.envri.eu/download/attachments/8553250/EC-091116-1403-27.pdf>.
- [EOSC 2016] 15. European commission. High Level Experts Group on Realising the European Open Science Cloud, Retrieved 10 November, 2016, from [http://ec.europa.eu/research/openscience/pdf/realising\\_the\\_european\\_open\\_science\\_cloud\\_2016.pdf](http://ec.europa.eu/research/openscience/pdf/realising_the_european_open_science_cloud_2016.pdf).





- [ERIS 2014] Asmi, A., Konijn, J., Pursula, A. (2014). *ERIS Environmental Research Infrastructures Strategy for 2030*. figshare. doi: [10.6084/m9.figshare.2067537.v1](https://doi.org/10.6084/m9.figshare.2067537.v1) Retrieved: 12<sup>th</sup> September 2016.
- [FAIR 2016] Force 11. Guiding Principles for Findable, Accessible, Interoperable and Re-Usable Data Publishing Version B1.0. Retrieved on 2nd December 2016 from: <https://www.force11.org/fairprinciples>.
- [Higgins 2008] Higgins, Sarah. "The DCC curation life cycle model." *International Journal of Digital Curation* 3.1 (2008): 134-140. doi: [10.2218/ijdc.v3i1.48](https://doi.org/10.2218/ijdc.v3i1.48).
- [Hodson 2016] Hodson, D. Strategies for Open Science and Research Data, Conferência: Dados de investigação e ciência aberta, Porto Portugal 22/09/2016. Retrieved 9 November, 2016, from [http://confdados.rcaap.pt/wp-content/uploads/2016/09/ConfDados\\_Simon\\_Hodson\\_web.pdf](http://confdados.rcaap.pt/wp-content/uploads/2016/09/ConfDados_Simon_Hodson_web.pdf)
- [INSPIRE 2007] EU Parliament, "Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE)," *Official Journal of the European Union*, vol. 50, no. L108, April 2007.
- [ISO/IEC 7498 1994] ISO/IEC standard 7498-1:1994. Information technology -- Open Systems Interconnection -- Basic Reference Model.
- [ISO/IEC 10746-1 1998] ISO/IEC 10746-1:1998. Information technology -- Open Distributed Processing -- Reference model: Overview.
- [ISO/IEC 10746-2 2009] ISO/IEC 10746-2:2009 Information technology -- Open distributed processing -- Reference model: Foundations.
- [ISO/IEC 10746-3 2009] ISO/IEC 10746-3:2009 Information technology -- Open distributed processing -- Reference model: Architecture.
- [ISO/IEC 10746-4 1998] ISO/IEC 10746-4:1998 Information technology -- Open Distributed Processing -- Reference Model: Architectural semantics.
- [ISO/IEC 19793 2015] ISO/IEC 19793:2015 Information technology -- Open Distributed Processing -- Use of UML for ODP system specifications.
- [Kreger 2014] Kreger, H., Brunssen, V., Sawyer, R., Arsanjani, A., High, R. The IBM advantage for SOA reference architecture Standards, Retrieved 01/12/2016 from: <http://www.ibm.com/developerworks/library/ws-soa-ref-arch/ws-soa-ref-arch-pdf.pdf>.
- [Kress 2014] Kress, J. How to Avoid the Perils of Patchwork Cloud Integration Retrieved 01/12/2016 from: [https://blogs.oracle.com/soacommunity/entry/how\\_to\\_avoid\\_the\\_perils](https://blogs.oracle.com/soacommunity/entry/how_to_avoid_the_perils).
- [Lannom 2016] Lannom, L., Wittenburg, P. The Digital Object Cloud, Retrieved 01/12/2016 from: <https://www.rd-alliance.org/group/data-fabric-ig/wiki/global-digital-object-cloud>.
- [Liew 2016] Liew, C.S., Atkinson, M.P., Galea, M., Ang, T.F., Martin, P., and van Hemert, J.I. Scientific Workflows: Moving Across Paradigms, *ACM Comput. Surv.* Vol 49, No. 4 pp 66:1-66:39
- [Linington 2012] Linington, P. F., Milosevic, Z., Tanaka, A., & Vallecillo, A. (2011). Building enterprise systems with ODP: an introduction to open distributed processing. CRC Press.
- [Martin 2015] Martin, P., Grosso, P., Magagna, B., Schentz, H., Chen, Y., Hardisty, A., Los, W., Jeffery, K., de Laat, C., Zhao, Z. (2015) Open Information Linking for Environmental Research Infrastructures. Presented at: IEEE 11th International Conference on e-Science, Munich, Germany, 31 August 2015 - 4 September 2015. *e-Science (e-Science)*, 2015 IEEE 11th International Conference on. IEEE, pp. 513-520. doi: [10.1109/eScience.2015.66](https://doi.org/10.1109/eScience.2015.66).
- [Michener 2012] Michener, William K., et al. "Participatory design of DataONE—enabling cyberinfrastructure for the biological and environmental sciences." *Ecological Informatics* 11 (2012): 5-15. doi: [10.1016/j.ecoinf.2011.08.007](https://doi.org/10.1016/j.ecoinf.2011.08.007).



- [Nativi 2015] Nativi, S., Jeffery, K.G., and Koskela, R. RDA: Brokering with Metadata. ERCIM News 2015.100 (2015). Retrieved on 9<sup>th</sup> December 2016 from <http://ercim-news.ercim.eu/en100/special/rda-brokering-with-metadata>.
- [NIST 2011] National Institute of Standards and Technology. The NIST Definition of Cloud Computing. Recommendations of the National Institute of Standards and Technology. Retrieved 8 November, 2016, from <http://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf>.
- [OASIS 2006] OASIS. Reference Model for Service Oriented Architecture 1.0. OASIS Standard, 12 October 2006. Retrieved 2 November, 2016, from <http://docs.oasis-open.org/soa-rm/v1.0/soa-rm.pdf>.
- [OGC ORM 2011] Open Geospatial Consortium. OGC Reference Model, Retrieved 7 November, 2016, from <http://www.opengeospatial.org/standards/orm>.
- [RDA 2016a] Research Data Alliance. Research Data Alliance in a nutshell. Retrieved 10 November, 2016, from [http://rd-alliance.org/sites/default/files/attachment/RDA\\_in\\_a\\_nutshell\\_November2016.pptx](http://rd-alliance.org/sites/default/files/attachment/RDA_in_a_nutshell_November2016.pptx).
- [RDA 2016b] Research Data Alliance. Data Fabric Interest Group. Retrieved 11 November, 2016, from <https://www.rd-alliance.org/group/data-fabric-ig.html>.
- [Santoro 2016] Santoro, M., Nativi, S., Mazzetti, P. Contributing to the GEO Model Web implementation: A brokering service for business processes . Environmental Modelling & Software Vol 84 October 2016. Pages 18 – 34. <http://dx.doi.org/10.1016/j.envsoft.2016.06.010>.
- [SEI 2016] Software Architecture, Getting started, Glossary. Retrieved 3 November, 2016, from <https://www.sei.cmu.edu/architecture/start/glossary/index.cfm>.
- [TOGAF 2011] The Open Group. Foundational Architecture TOGAF Reference Models, TOGAF Version 9.1, Open Group Standard. Retrieved 2 November, 2016, from <http://pubs.opengroup.org/architecture/togaf9-doc/arch/index.html>.
- [Wikipediaorg 2016]. Reference Model. Retrieved 2 November, 2016, from [https://en.wikipedia.org/wiki/Reference\\_model](https://en.wikipedia.org/wiki/Reference_model).
- [Zhao 2015] Zhao Z, Martin P, Grosso P, Los W, de Laat C, Vermeulen A, Jeffrey K, Castelli D, Hardisty A, Legré Y, Kutsch W. Reference Model Guided System Design and Implementation for Interoperable Environmental Research Infrastructures. Presented at: e-Science 2015: IEEE 11th International Conference on e-Science, Munich, Germany, 31 August - 4 September 2015. e-Science (e-Science), 2015 IEEE 11th International Conference on. IEEE, pp. 551-556. doi: [10.1109/eScience.2015.41](https://doi.org/10.1109/eScience.2015.41).



## Appendix 1: National Marine Biodiversity Data Archive Centre (DASSH) – A validation case

The Marine Biological Association (MBA)<sup>23</sup> coordinates the DASSH Data Archive Centre<sup>24</sup>, which is a UK national facility for the archival of marine species and habitat data. DASSH needs to be integrated with other European marine biological data (e.g., data curated by EMSO, SeaDataNet, JERICO and EMBRC) as a joint contribution to EMODNET Biology<sup>25</sup>, the COPENICUS provider.

The aim of this test case is to enable an RM-based description of the DASSH Data Centre, and its associated services. It will facilitate integration with the EU data infrastructure, and serve as a demonstration of how the ENVRI RM could benefit other data centres within the UK MEDIN Partnership<sup>26</sup>.

MBA would like to describe the entire data flow from acquisition to publication. Existing flow diagrams are given in Figure 9 and Figure 10 below, with both internal and external flows and process illustrated. These are the starting point.

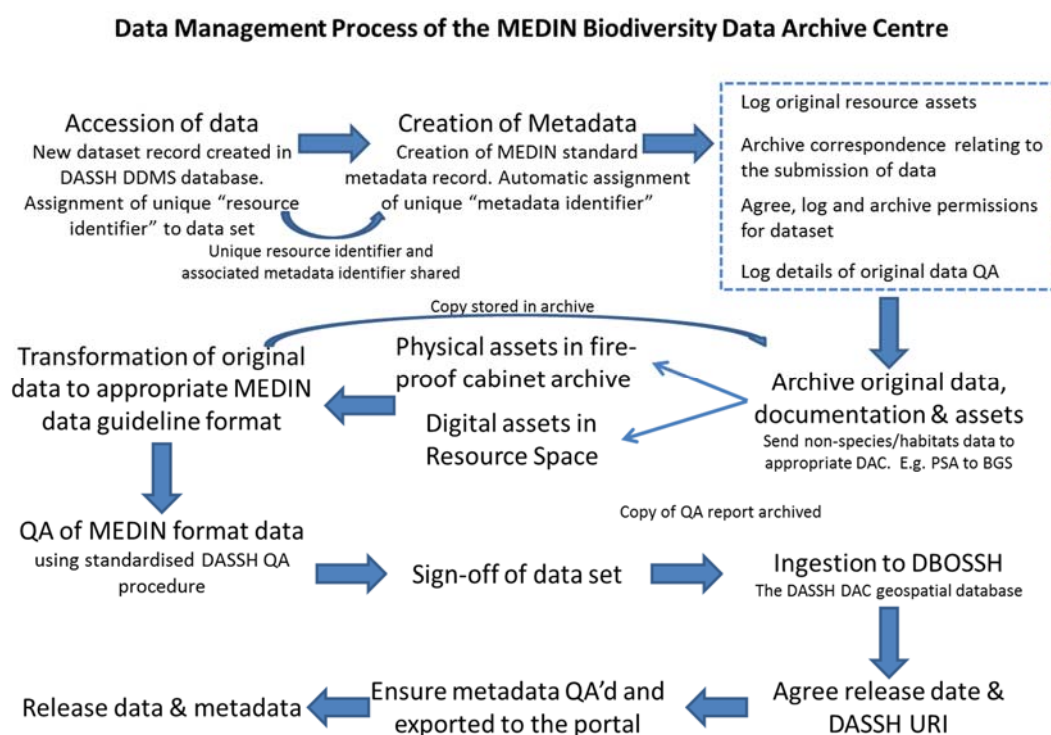


FIGURE 9 EXISTING DATA MANAGEMENT PROCESS

<sup>23</sup> <https://www.mba.ac.uk/>

<sup>24</sup> <http://www.dassh.ac.uk/>

<sup>25</sup> <http://www.emodnet-biology.eu/>

<sup>26</sup> <http://www.oceannet.org/>

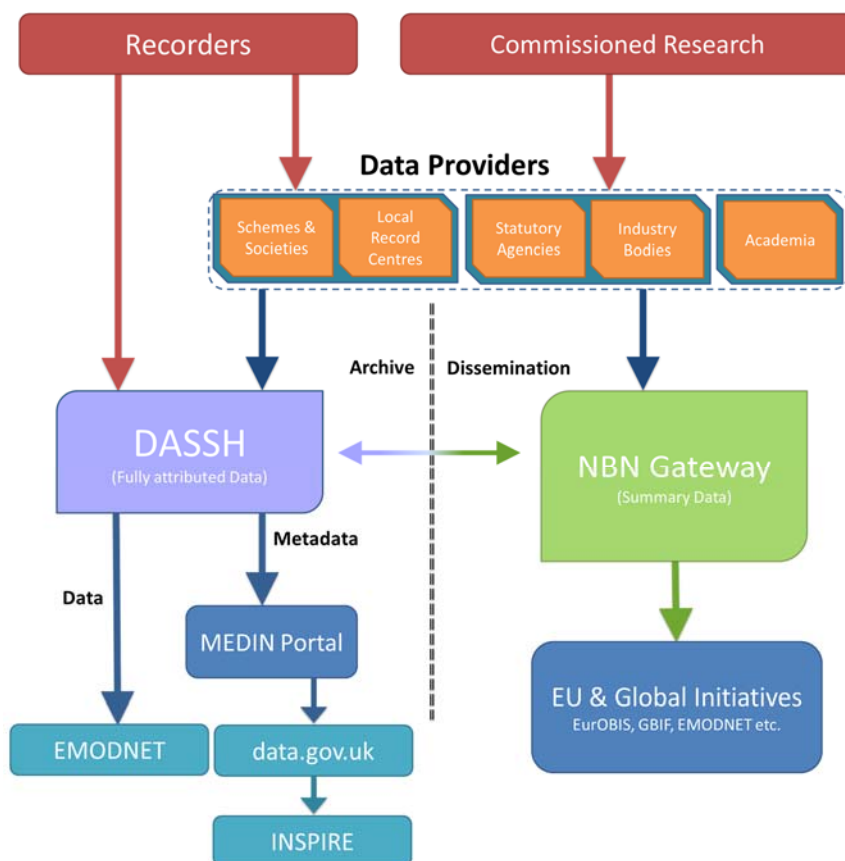


Figure 2. Recommended Data Flow within the UK

#### FIGURE 10 EXISTING DATAFLOWS AT NATIONAL LEVEL

This is a typical test case for the ENVRI RM. It will help to improve both the integration with EU data infrastructure and the Reference Model. Results from the test case help to validate the Reference Model.

A challenge in this test case is that the involved MBA staff have no prior experience with the ENVRI RM and are not ICT professionals. The implementation of the RM by domain-centric experts could highlight differences or uncertainty in definitions and terminology used within the RM.

Following in Figure 11 - Figure 16 below, are some examples of UMLet diagrams for the RM modelling of DASSH.

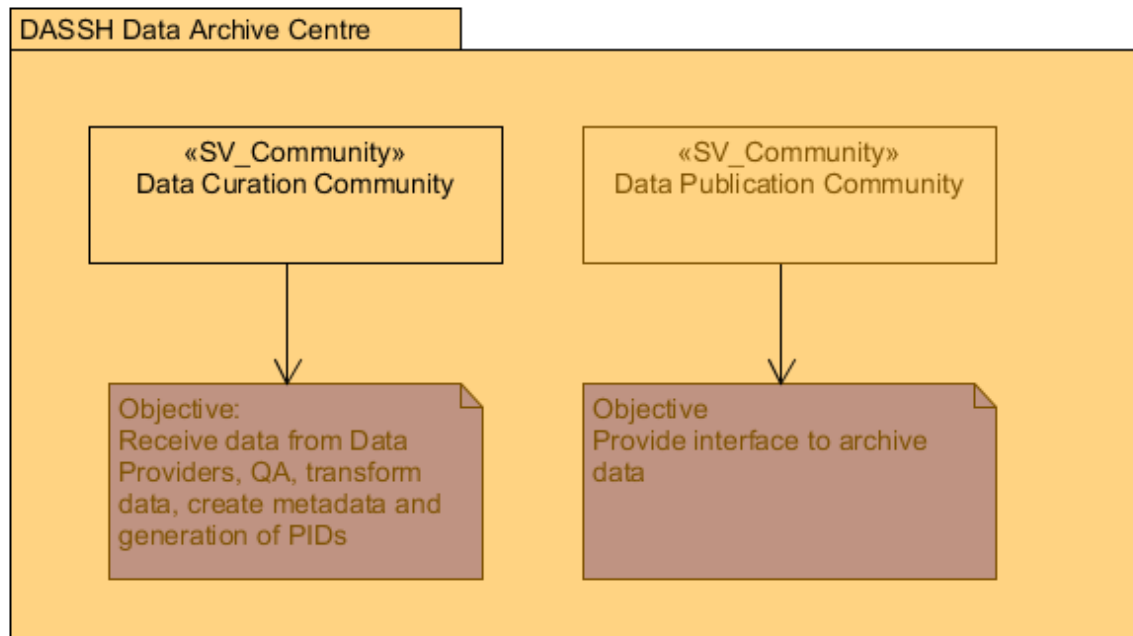


FIGURE 11 DIAGRAMS FOR THE RM MODELLING OF DASSH, EXAMPLE 1

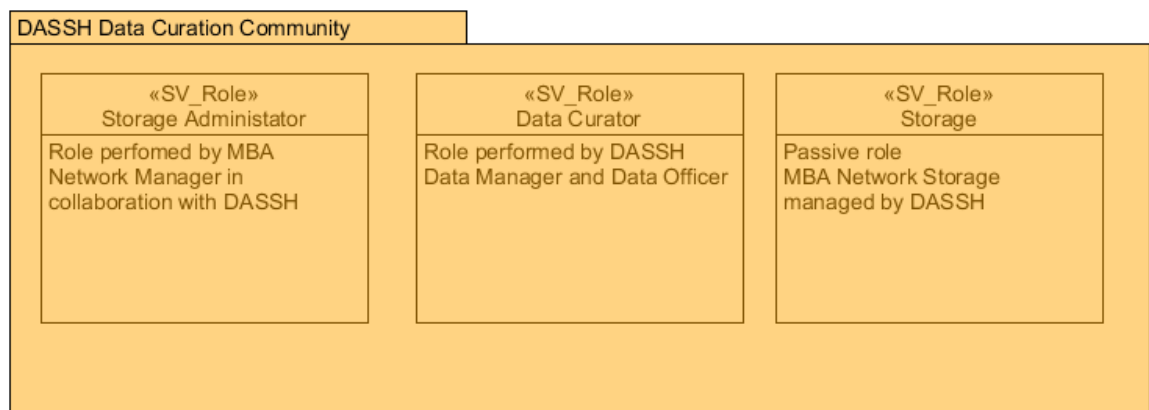


FIGURE 12 DIAGRAMS FOR THE RM MODELLING OF DASSH, EXAMPLE 2

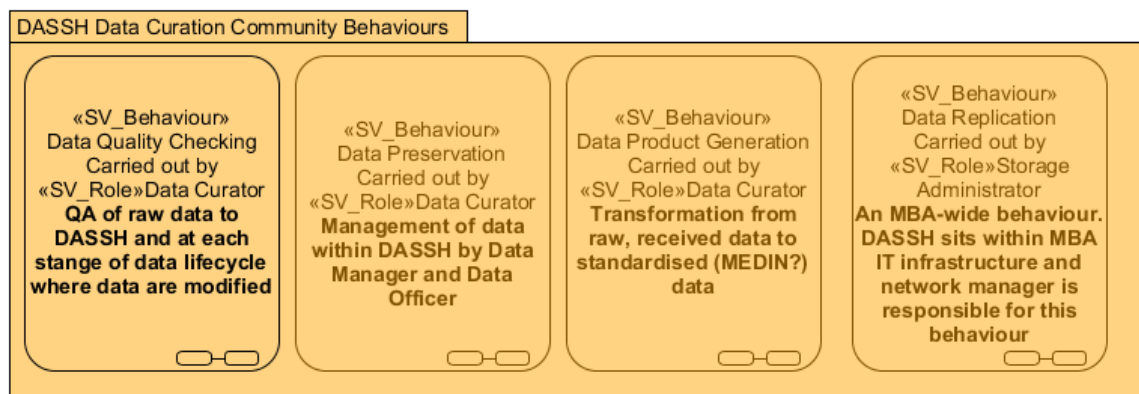


FIGURE 13 DIAGRAMS FOR THE RM MODELLING OF DASSH, EXAMPLE 3

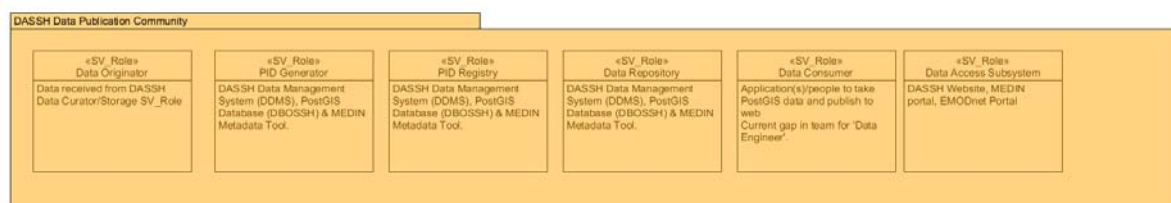


FIGURE 14 DIAGRAMS FOR THE RM MODELLING OF DASSH, EXAMPLE 4

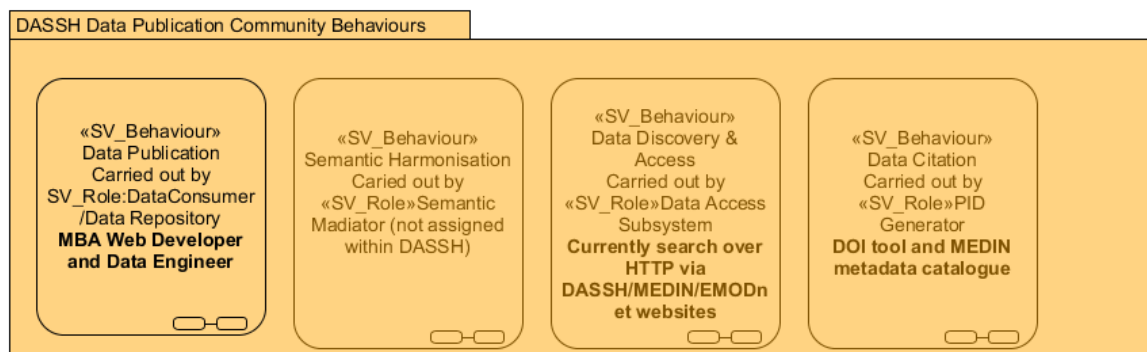


FIGURE 15 DIAGRAMS FOR THE RM MODELLING OF DASSH, EXAMPLE 5

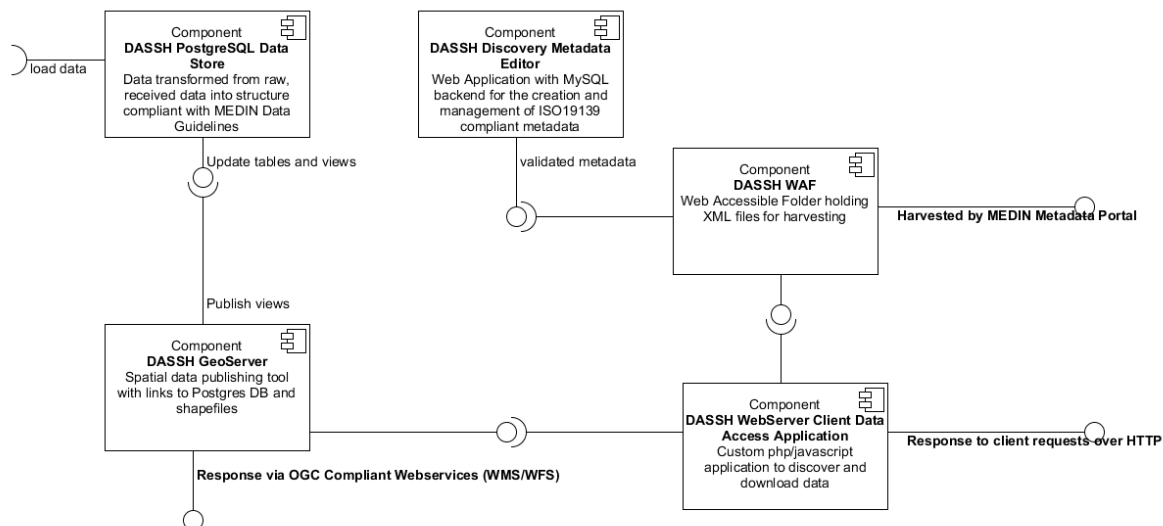


FIGURE 16 DIAGRAMS FOR THE RM MODELLING OF DASSH, EXAMPLE 6



## Appendix 2: European Facility for Airborne Research (EUFAR) – A validation case

The current EUFAR (European Facility for Airborne Research) represents a consortium of 24 European institutions and organisations involved in airborne research. It is a unique pan-European portal and network for airborne research infrastructures dedicated to environmental sciences. EUFAR works to coordinate the operation of instrumented aircraft and remote sensing instruments, exploiting the skills of experts in airborne measurements in the fields of environmental and geo-sciences, to provide researchers with the infrastructure best suited to their needs.

The aim of this test case is to describe the processes within EUFAR infrastructure using the ENVRI Reference Model (RM), to describe requirements and architectural features of the infrastructure, and serve as a common language in communication materials.

To assure the continuity of EUFAR beyond the end of its current project funded activity, EUFAR is establishing a sustainable legal structure (an International non-profit Association under Belgian law - AISBL). The AISBL is suitable for multi-national infrastructure networks and is a key step to enable EUFAR to develop more straightforward ways of accessing aircraft across Europe and to develop more efficient use of the existing facilities. The RM will be used to support these developments.

The expected output of the test case will help in the preparation of EUFAR in the process of establishment towards a sustainable legal structure.

Table 6 illustrates some relevant modelling components representing aspects of EUFAR in the Science Viewpoint.

TABLE 6 RELEVANT MODELLING COMPONENTS OF THE SCIENCE VIEWPOINT

<b>Roles Instances</b>	<b>Key Responsibilities</b>	<b>SV_Community_Behaviours</b>	<b>SV_Community</b>
Aircraft operators with PI and instrument operator/scientist	designs experiment based on instrument/aircraft limitations	Design of Measurement Model	Data Acquisition Community
instrument operator/scientist		Instrument Configuration	Data Acquisition Community
instrument operator/scientist		Instrument Calibration	Data Acquisition Community
instrument operator/scientist		Data Collection	Data Acquisition Community
instrument operator/scientist	data is quality checked before uploading to archive	Data Quality Checking	Data Curation Community
data archive manager		Data Preservation	Data Curation Community
data processor/processing team	data is formatted before it is uploaded to the archive	Data Product Generation	Data Curation Community
data archive manager		Data Replication	Data Curation Community
data archive manager		Data Publication	Data Publication Community

<b>Roles Instances</b>	<b>Key Responsibilities</b>	<b>SV_Community_Behaviours</b>	<b>SV_Community</b>
S&P engineer	ASMM and EMC tools to create common metadata /descriptions of the data	Semantic Harmonisation	Data Publication Community
Ceda Metadata catalogue search		Data Discovery and Access	Data Publication Community
CEDA archive staff	DOIs will be allocated for datasets	Data Citation	Data Publication Community
EUFAR is not a service provider for data analysis/modelling			Data Service Provision Community
data archive manager	data download stats recorded by data archive	User Behaviour Tracking	Data Usage Community
EUFAR office staff	tracks EUFAR members when logging into website	User Behaviour Tracking	Data Usage Community
EUFAR office staff	Users can have different profiles in EUFAR website allowing different access	User Profile Management	Data Usage Community
	not requested for EUFAR	User Working Space Management	Data Usage Community
Expert Working Groups and user gesiration (incl. fileds of expertise)	User DB is searchable on the webpage	User Working Relationships Management	Data Usage Community
EUFAR webpage		User Group Work Supporting	Data Usage Community

## Appendix 3: Alignment of ENVRI RM with RDA terms and definitions

The Data Foundations and Terminology Working Group (DFT WG) of the Research Data Alliance (RDA) has completed its initial working cycle. The outputs of the DFT WG are summarised in Table 7 below and its notes. In the following text references to the relevant documents are given as references to the notes in Table 7.

The main output of the group is contained in the Results RFC (Note 6), which is the current snapshot and contains definitions of terms and their relationships. This document is provided with the suffix RFC, meaning that it is in its final phase for review by the community and to become a standard. The document contains 14 terms, their definitions and relationships. The document also presents a diagram describing the relationships of the terms.

TABLE 7 DFT WG SUMMARY OF OUTPUTS

Title	Description	Comment
RDA Data Foundation and Terminology DFT 1: Overview (Note 4)	Presents 22 Data Models considered as basis for building a common model	ENVRI model is discussed
RDA Data Foundation and Terminology DFT 2: Analysis & Synthesis (Note 7)	Presents an analysis and synthesis of the models described in the first document	Presents the first version of the data model.
RDA Data Foundation and Terminology DFT 3: Snapshot of DFT Core Terms (Note 8)	Presents the first definition of core terms.	Highlights this as a snapshot of work in progress towards real, working agreements on terminology within RDA and across the worldwide data community
RDA Data Foundation and Terminology DFT 4: Use Cases (Note 5)	Presents a series of use cases and how they use the core terms to describe them.	Assert that EUDAT is largely aligned and influenced by RDA work Describes CLARIN as an early DFT adopter Talks about data streams but calls them “gappy dynamic data”
RDA Data Foundation and Terminology DFT 3: Term Tool Description <sup>27,28</sup> (Note 1)	Presents an online tool for collecting terms provided by the adopting community	Online tool contains 159 term definitions, interesting to see what they are.
RDA Data Foundation and Terminology - DFT: Results RFC (Note 6)	First RFC on the terms identified by the DFT	This contains the definitions of 14 concepts and their relationships. This is a basic ontology for classifying data objects and repositories
DFT WG Products Metadata (Note 1)	Metadata about documents	
Data Foundation and Terminology Work Group Products Maintenance and Retirement plan. (Note 9)	How to pass on the product from the WG to an IG	Retirement plan contemplates transition to DFT IG to continue the development of the model

<sup>27</sup> Error the title says 3 and it should be 5, according to the outline included in most of the DFT WG documents.

<sup>28</sup> The RDA’s term definition tool can be found at: [http://smw-rda.esc.rzg.mpg.de/index.php/Main\\_Page](http://smw-rda.esc.rzg.mpg.de/index.php/Main_Page).



Title	Description	Comment
RDA Data Foundation and Terminology DFT: Adoption Note (Note 3)		Acts as main entry point in addition to the other documents. Points to interaction with EUDAT and ESFRI RIs as key for further development
<p>Note 1. DFT WG – RDA (2014). DFT WG Products Metadata. Creator: Gary Berg -Cross, Raphael Ritz, Peter Wittenburg. Date 12/10/2014. Consulted on 04/03/2016. Available at: <a href="https://rd-alliance.org/system/files/DFT%20WG%20%20Products%20metadata.pdf">https://rd-alliance.org/system/files/DFT%20WG%20%20Products%20metadata.pdf</a></p> <p>Note 2. DFT WG – RDA (2014). RDA Data Foundation and Terminology DFT 3: Term Tool Description. Eds. Thomas Zastrow, Gary Berg-Cross, Raphael Ritz. Date: 07/12/2016. Consulted on: 04/03/2016. Available at: <a href="https://rd-alliance.org/system/files/DFT%20term%20tool-dec-1-0.pdf">https://rd-alliance.org/system/files/DFT%20term%20tool-dec-1-0.pdf</a></p> <p>Note 3. DFT WG – RDA (2014). RDA Data Foundation and Terminology DFT: Adoption Note. Eds. Gary Berg -Cross, Raphael Ritz, Peter Wittenburg. Date 31/12/2014. Consulted on 04/03/2016. Available at: <a href="https://rd-alliance.org/system/files/adoption-notes-1-2.pdf">https://rd-alliance.org/system/files/adoption-notes-1-2.pdf</a></p> <p>Note 4. DFT WG – RDA (2014). RDA Data Foundation and Terminology DFT 1: Overview. Date 31/12/2014. Eds. Gary Berg-Cross, Karen Green, Peter Wittenburg. Consulted on 04/03/2016. Available at: <a href="https://rd-alliance.org/system/files/RDA%20DFT%20Data%20Models-v1-6.pdf">https://rd-alliance.org/system/files/RDA%20DFT%20Data%20Models-v1-6.pdf</a></p> <p>Note 5. DFT WG – RDA (2014). RDA Data Foundation and Terminology DFT 4: Use Cases. Date 31/12/2014. Eds. Gary Berg, Karen Green, Peter Wittenburg. Consulted on 04/03/2016. Available at: <a href="https://rd-alliance.org/system/files/Use%20case%20v1%206.pdf">https://rd-alliance.org/system/files/Use%20case%20v1%206.pdf</a></p> <p>Note 6. DFT WG – RDA (2015). RDA Data Foundation and Terminology - DFT: Results RFC. Eds. Gary Berg -Cross, Raphael Ritz, Peter Wittenburg. Date: 29/06/2015. Consulted on: 04/03/2016. Available at: <a href="https://rd-alliance.org/system/files/DFT%20Core%20Terms-and%20model-v1-6.pdf">https://rd-alliance.org/system/files/DFT%20Core%20Terms-and%20model-v1-6.pdf</a></p> <p>Note 7. DFT WG – RDA (2015). RDA Data Foundation and Terminology DFT 2: Analysis &amp; Synthesis. Eds. Gary Berg-Cross, Karen Green, Raphael Ritz, Peter Wittenburg. Date: 31/07/2015. Consulted on: 04/03/2016. Available at: <a href="https://rd-alliance.org/system/files/RDA%20DFT%20Data%20Models-analysis-v1-5.pdf">https://rd-alliance.org/system/files/RDA%20DFT%20Data%20Models-analysis-v1-5.pdf</a></p> <p>Note 8. DFT WG – RDA (2015). RDA Data Foundation and Terminology DFT 3: Snapshot of DFT Core Terms. Eds. Gary Berg-Cross, Keith Jeffery, Bob Kahn, Larry Lannom, Raphael Ritz, Herman Stehouwer, Peter Wittenburg, Thomas Zastrow, Zhu Yunqiang. Date: 31/07/2015. Consulted on: 04/03/2016. Available at: <a href="https://rd-alliance.org/system/files/DFT%20Core%20Terms-dec-1-5.pdf">https://rd-alliance.org/system/files/DFT%20Core%20Terms-dec-1-5.pdf</a></p> <p>Note 9. DFT WG – RDA. Data Foundation and Terminology Work Group Products Maintenance and Retirement plan. Date: Not available. Consulted on: 04/03/2016. Available at: <a href="https://rd-alliance.org/system/files/Data%20Foundation%20and%20Terminology%20Work%20Group%20Products%20Maintenance%20and%20Retirement%20plan.pdf">https://rd-alliance.org/system/files/Data%20Foundation%20and%20Terminology%20Work%20Group%20Products%20Maintenance%20and%20Retirement%20plan.pdf</a></p>		

## ENVRI RM Alignment to DFT Core Term Definitions

The terms defined in the ENVRI RM Information Viewpoint (IV) are the closest match to those described by the DFT WG. Table 8 includes a comparison of the terms in the IV and then the corresponding term in the DFT Core Term Definitions (DFTWG-CTD) (Note 8).

TABLE 8 IV TERMS COMPARED TO DFT WG TERMS

IV Term	DFTWG-CTD	Reasoning
specification of investigation design	Metadata	Describes the reason for collecting the digital object
specification of measurements or observations	Metadata	Describes the method for collecting the digital object
measurement result	Bitstream	Direct mapping, in this case the bit stream is the digitalised form of a reading. How do we deal with physical and analogue recordings?
Concept	Metadata	Assigns meaning and links it to other metadata or to a digital object
conceptual model	Digital Collection	Is an aggregation of concepts, thus an aggregation of metadata
QA notation	Digital Collection	Is an aggregation of concepts related to the quality of data and metadata
Metadata	Metadata	Direct link but the meaning given by DFTWG is more strict and brief
metadata state	state information	Describes state of metadata

IV Term	DFTWG-CTD	Reasoning
metadata catalogue	Digital Collection	Aggregation of digital objects
Citation	Persistent Identifier	Long lasting ID that links to a digital object
persistent data	Digital Object	Direct link
data state	state information	Describes state of digital object
unique identifier (UID)	Persistent Identifier	Long lasting ID that links to a digital object
Backup	Digital Object	
mapping rule	Metadata	Describes allowed operations on data
data provenance	Metadata	Describes the digital object
Service		No correspondence directly but PID resolver conversely could be seen as a service
service description	Metadata	Describes the digital object

The DFT WG definitions and the model are quite simple in concept, so that it is not hard to map to concepts in the IV. Using the DFT WG terminology treats the IV terms as a set of data and metadata terms.

Table 9 compares the terms from the DFT WG to terms in the ENVRI RM

TABLE 9 DFT WG TERMS COMPARED TO IV TERMS

DFTWG-CTD	IV Term(s)	Reasoning
Digital Object	All information objects, with the exception of service, can be characterised as digital objects	The definitions of information object and metadata allow for this
Persistent Identifier (PID)	unique identifier (UID)	Long lasting ID that links to a digital object
PID Record	metadata, data provenance	The PID record is a set of attributes describing a Digital Object
PID Resolver	Service	A service (or system) globally available for resolving a PID
Metadata	specification of investigation design, specification of measurement or observation, conceptual model, QA notation, metadata, metadata catalogue citation, concept, data state, metadata state, back up, data provenance, mapping rule, service description.	All metadata and collections in the IV can be seen as types of metadata
Aggregation	specification of investigation design, specification of measurement or observation, conceptual model, QA notation, metadata catalogue citation, concept, data state, metadata state, back up, data provenance, service description.	All the collections in the IV are aggregations of Digital Objects
Digital Collection	specification of investigation design, specification of measurement or observation, conceptual model, QA notation, metadata catalogue citation, concept, data state, metadata state, back up, data provenance, service description.	All the aggregations are digital collections as well
Digital Entity	All information objects can be characterised as digital entities	By definition

<b>DFTWG-CTD</b>	<b>IV Term(s)</b>	<b>Reasoning</b>
Repository	Service	By definition
Bitstream	persistent data	By definition
State information	Data state, metadata state	By definition
Property	Metadata	By definition
Metadata Repository	Metadata catalogue, service	By definition
Checksum	QA Notation, Metadata	By definition

The IV does not have a definition of Semantic Annotation. Neither does the DFTWG-CTD.

## Appendix 4: Grouping of Analysis of results from D5.1

Deliverable 5.1 of ENVRIplus contains a review of RI requirements, and a review of the state of the art technologies provided by data and computational infrastructures. The RI requirements section is divided in two subsections. The first subsection contains generic information on the 20 RIs. This section describes each RI and requirements not aligned with the core topics of ENVRIplus. The second subsection covers the specific requirements according to the seven topics of ENVRIplus: (1) Identification and Citation, (2) Curation, (3) Cataloguing, (4) Processing, (5) Provenance, (6) Optimisation, and (7) Community Support.

The requirements document contains 180 requirements, 102 requirements from the generic section and 78 from the topics section. The analysis of the requirements involved trying to map the requirements to functionalities included in the ENVRI RM and to the data life cycle phases. Table 10 table shows the results of those mappings compared to the subsections of the requirements document.

TABLE 10 DISTRIBUTION OF REQUIREMENTS GROUPED ACCORDING TO D5.1 CATEGORIES

Type of Requirement	Count	Mapped	Not Mapped
Generic	100	96	4
Identification and Citation	9	9	0
Curation	2	2	0
Cataloguing	7	7	0
Processing	11	11	0
Provenance	12	12	0
Optimisation	23	22	1
Community Support	14	14	0
<b>Grand Total</b>	<b>178</b>	<b>173</b>	<b>5</b>

This mapping will guide the efforts to update and complement the RM to better reflect the actual requirements of the RIs participating in ENVRIplus. Table 11 presents a summary of the same requirements as aligned to the phases of the data life cycle

TABLE 11 DISTRIBUTION OF REQUIREMENTS ALIGNED TO THE DATA LIFE CYCLE PHASES

Phase of the Data Life cycle	Count
Acquisition	7
Curation	45
Publishing	52
Processing	41
Use	28
Phase independent	5
<b>Grand Total</b>	<b>178</b>

### Non-mapped requirements

The first requirements to analyse are those which cannot be aligned with ENVRI RM functionalities or specific data life cycle phases. Four of these requirements belong to the generic section and the remaining one belongs to the optimisation section (Table 12).

TABLE 12 UNMAPPED REQUIREMENTS

RI	Requirement	Page in D5.1	Type
EISCAT-3D	Ensure that the tools that they select are well documented and open, in order to minimise the risk of lock in to proprietary systems.	31	Generic



RI	Requirement	Page in D5.1	Type
EPOS	Taking already existing software and make it available and scalable across communities.	33	Generic
SeaDataNet	Enhance the cross-community expertise on observation networks, requirements support and data management expertise by participating in ENVRIplus.	36	Generic
EMBRC	Maintenance of software and their integration into a single platform.	32	Generic
N/A	Every level of the system needs to be well enough described to support automated management and optimisation.	56	Optimisation

## Data acquisition requirements

Acquisition requirements mostly fall within the area of sensor networks setup and management which are covered in the ENVRI RM by functionalities listed in the last column of Table 13. The ENVRI RM already acknowledges all these requirements.

TABLE 13 ACQUISITION REQUIREMENTS

RI	Requirement	Page in D5.1	Type	ENVRI RM Functionality
ACTRIS	Planning and managing the activity of sensors.	30	Generic	A.2
ACTRIS	Developing understanding of how instruments work in extreme conditions.	30	Generic	A.3
ACTRIS	Improving the capabilities of small sensors.	30	Generic	A.6
LTER	Real-time availability of data.	36	Generic	A.8, A.11, A.15
SIOS	How observational networks are to be designed and implemented.	36	Generic	All A
N/A	Processing input data taking into account: dataset(s) typologies, volume, velocity, variety/heterogeneity, and access methods.	50	Processing	A.11, A.12, A.13, A.14, A.15, A.16
N/A	Optimising the acquisition of data from data providers is required to maximise the range and timeliness of datasets made available to researchers and to increase data security (by ensuring that it is prepared for curation with minimal delay, reducing the risk of data corruption or loss).	56	Optimisation	

## Data curation requirements

There are 45 requirements which can be linked to the curation phase of the data life cycle. This high count is due to the fact that three of the core topics of ENVRIplus (Identification, Cataloguing, Curation) can be mapped to the curation phase. The most important requirements related to the curation phase were those mapped to data identification (10 occurrences), data cataloguing (12 occurrences), and data product generation (8 occurrences), see tables Table 18, Table 19, and Table 20 respectively.

TABLE 14 REQUIREMENTS WHICH CAN MAP TO ANY CURATION FUNCTIONALITY

RI	Requirement	Page in D5.1	Type
FixO3	Harmonise data curation	44	generic
LTER	Support on data curation	45	generic
N/A	the requirements for curation were not made explicit, for example, none of the RIs (who responded) has appropriate metadata and processes for curation	48	Curation



N/A	curation often underpins validation of the quality of scientific decisions and since environmental sciences observe phenomena that do not repeat in exactly the same form, the profile of curation needs raising	48	Curation
-----	--	----	----------

**TABLE 15 CURATION REQUIREMENTS WHICH ARE SINGLE INSTANCES**

RI	Requirement	Page in D5.1	Type	ENVRI RM Functionality
AnaEE	the quality control of data produced by platforms	30	Generic	B.1, B.2
EuroGOOS	From a technological perspective, getting recommendations about the design of common data system, including formats or data platforms and data treatments.	34	Generic	B.1, B.3, B.4, B.5, B.8
ICOS	Metadata curation (including “recipes” for cataloguing and storage)	35	Generic	B.4, B.6, B.8
IS-ENES2	Handling volume and distribution of data: Replication, Versioning.	45	Generic	B.6
	High availability and long-term durability requires effective replication procedures across multiple sites.	56	Optimisation	B.9
	Minimise the cost of synchronising replicas	56	Optimisation	B.10

**TABLE 16 CURATION REQUIREMENTS RELATED TO WORKFLOW ENACTMENT (D.6)**

RI	Requirement	Page in D5.1	Type
EISCAT-3D	define workflows for data	31	Generic
LTER	integration of common data repositories into their workflow system (including metadata documentation with LTER Europe DEIMS	36	Generic

**TABLE 17 CURATION REQUIREMENTS RELATED TO DATA STORAGE AND PRESERVATION (B.8)**

RI	Requirement	Page	Type
Euro-ARGO	The relevant data are pushed from the RI to the ENVRI cloud	33	Generic
EMBRC	Backup system	32	Generic
ICOS	Data and Metadata storage.	44	Generic

**TABLE 18 CURATION REQUIREMENTS FOR DATA IDENTIFICATION (B.3)**

RI	Requirement	Page in D5.1	Type
AnaEE	Identification	30	Generic
IAGOS	Citation and DOI management	34	Generic
ICOS	Data object identification and citation	35	Generic
AnaEE	Homogenous approach on Identification and citation and on cataloguing across RIs	45	Expectations
LTER	data object identification especially on the aspect of dynamic data series and identification of results from data queries (e.g. data services)	45	Expectations
Survey*	having as much of the data identification and citation automated	46	Identification and citation
Survey*	Support identification of data objects stored as files (using file names as identifiers)	46	Identification and citation

Survey*	identifier systems used based on handles: DOIs from DataCite most common, followed by ePIC PIDs	46	Identification and citation
Survey*	use of persistent and unique identifiers for both data and metadata objects throughout the entire data life cycle	47	Identification and citation
N/A	ensure availability of identification services	56	Optimisation
* Reported as responses to survey from: ACTRIS, AnaEE, EISCAT-3D, EMBRC, EMSO, EPOS, Euro-ARGO, EuroGOOS, IAGOS, ICOS, IS-ENES2, LTER, SeaDataNet, and SIOS			

**TABLE 19 CURATION REQUIREMENTS FOR DATA CATALOGUING (B.4)**

RI	Requirement	Page in D5.1	Type
AnaEE	cataloguing	30	Generic
IAGOS	cataloguing	34	Generic
N/A	Catalogues of observation systems and lab equipment, if possible using a standardised approach (OGC/SWE, SSN)	48	Cataloguing
N/A	Catalogues of physical samples	49	Cataloguing
N/A	Catalogues of data products and/or results. Currently done by existing systems (EBAS, EARLINET, CLOUDNET, CKAN, MADrigal, DEIMS). Widely standardised (ISO/IEC 191XX)	49	Cataloguing
N/A	Catalogues of publications, Few RIs manage the publications on their own. Links for provenance between publications and datasets are required	49	Cataloguing
N/A	Reference catalogues: for objects used by RIs and researchers such as: people and organisations, publications, research objects, features of interest	49	Cataloguing
N/A	Federated catalogues: for data objects produced by research activities such as: data products and results, observation systems and lab equipment, physical samples, data processing procedures, systems and software, metadata	49	Cataloguing
N/A	Processing catalogues: to record activities, usage, events. These catalogues support provenance and management of the RI	49	Cataloguing
N/A	automatically supply the entire set of metadata characterising the task, e.g., through the provenance framework	50	Processing
N/A	Optimise the management of provenance data streams during data processing.	56	Optimisation
N/A	Preserving the link between data and their provenance metadata is important when metadata are not packaged with their corresponding datasets	56	Optimisation

**TABLE 20 CURATION REQUIREMENTS FOR DATA PRODUCT GENERATION (B.5)**

RI	Requirement	Page in D5.1	Type
EMBRC	Developing and learning about new standards and best practices in terms of standards. Developing new standards within INSPIRE [8], which can be used for other datasets.	32	Generic
EMBRC	Developing and learning about new standards and best practices in terms of standards.	32	Generic
EMBRC	Developing new standards within INSPIRE [8], which can be used for other datasets.	32	Generic
IAGOS	Metadata standardisation	34	Generic
INTERACT	Broad European standards exposed to NordGIS	35	Generic
INTERACT	Recommendations about how to turn primary data into data products need to be adopted.	45	Generic
INTERACT	Recommendations about how to turn primary data into data products need to be adopted.	45	Generic

INTERACT	Metadata and data standardisation at all levels.	45	Generic
----------	--	----	---------

## Data publishing requirements

There are 53 requirements which can be mapped to the publishing phase of the data life cycle. The high count in this case is due to many RIs with requirements related to publishing. The functionality which was mapped to most of these requirements was semantic harmonisation (25 occurrences), see Table 24. Other important requirements were those mapped to data publication (11 occurrences) and data citation (9 occurrences), see

Table 22 and Table 23 respectively.

**TABLE 21 PUBLISHING REQUIREMENTS WITH LOW COUNT**

RI	Requirement	Page in D5.1	Type	ENVRI RM Functionality
ACTRIS	Data visualisation	44	Generic	C.15
EISCAT-3D	Data access with searching and visualisation	31	Generic	
LTER	technical support on optimisation of data annotation (e.g. integrating of a data repository, data integration portal)	45	Generic	C.3

**TABLE 22 PUBLISHING REQUIREMENTS MAPPED TO DATA PUBLICATION (C.11)**

RI	Requirement	Page in D5.1	Type
EuroGOOS	Getting inspiration from RIs about ways to distribute the data to end users using applications which are more focused in this respect.	1	Generic
ACTRIS	Data provision	1	Generic
IS-ENES2	Providing related information for data products (provenance, user comments, usage, detailed scientific descriptions needed for usage).	1	Generic
SeaDataNet	Data policy to involve data providers in the publication of their own datasets	1	Generic
N/A	need to make output data available in a systematic way, including information on the entire process leading to the resulting data	5	Processing
N/A	output resulting from a data processing task should be “published” to be compliant with Open Science practices	5	Processing
N/A	offer an easy to use approach for having access to the datasets that result from a data processing task	5	Processing
N/A	provide the appropriate protection for cases where aspects of the information are sensitive, could jeopardise privacy, or have applications that require a period of confidentiality.	5	Processing
N/A	location of data such that they can be easily identified, retrieved and analysed	7	Optimisation
N/A	Optimise accessibility and availability of research assets (but not exclusively data).	7	Optimisation

**TABLE 23 PUBLISHING REQUIREMENTS MAPPED TO DATA CITATION (C.12)**

RI	Requirement	Page in D5.1	Type
AnaEE	Citation	30	Generic
IAGOS	Citation and DOI management	34	Generic
ICOS	Data object identification and citation	35	Generic

RI	Requirement	Page in D5.1	Type
AnaEE	Homogenous approach on Identification and citation and on cataloguing across RIs	45	Generic
Survey*	having as much of the data identification and citation automated	46	Identification and citation
Survey*	necessary to allow unambiguous references to be made to specified subsets of datasets, preferably in the citation	46	Identification and citation
Survey*	Ensuring that credit for producing (and to a lesser extent curating) scientific data sets is “properly assigned”	47	Identification and citation
Survey*	have strategies for collecting usage statistics for their data products, i.e., through bibliometric searches (quasi-automated or manual) from scientific literature, but thus often rely on publishers indexing also data object DOIs	47	Identification and citation
Survey*	standards of data citation supporting adding specific sub-setting information to a basic (DOI-based) reference.	47	Identification and citation
* Reported as responses to survey from: ACTRIS, AnaEE, EISCAT-3D, EMBRC, EMSO, EPOS, Euro-ARGO, EuroGOOS, IAGOS, ICOS, IS-ENES2, LTER, SeaDataNet, and SIOS.			

TABLE 24 PUBLISHING REQUIREMENTS MAPPED TO SEMANTIC HARMONISATION (C.13)

RI	Requirement	Page in D5.1	Type
ACTRIS	to improve their interoperability so as to make their data as accessible and understandable as possible to others,	30	Generic
ACTRIS	to link with others RIs, because there are many points in common (technologically and scientifically), and	30	Generic
ACTRIS	Semantic Interoperability	30	Generic
AnaEE	it would be useful to synchronise their approach with other RIs	30	Generic
AnaEE	the interoperability between models and data	30	Generic
EISCAT-3D	ensure interoperability with other RIs and instruments via virtual observatories	31	Generic
ELIXIR	establish a closer collaboration with environmental Research Infrastructures (RIs) and improve their access to life science data	31	Generic
ELIXIR	enhanced interaction, a better insight into data structures and relevant data standards widely adopted across environmental RIs can facilitate an effective evaluation of areas of collaboration for development of new tools, services and training	31	Generic
ELIXIR	Ultimately, this can lead to better interoperability and discoverability of environmental and life science data by users across atmospheric, marine, solid earth and biosphere domains.	31	Generic
EMBRC	Establishing collaborations with the environmental community, which would benefit from their environmental and ecological data.	32	Generic
EMSO	better mechanisms for ensuring harmonisation of datasets across their distributed networks	32	Generic
FixO3	better mechanisms for ensuring harmonisation of datasets across their distributed networks	34	Generic
IAGOS	Interoperability	34	Generic
INTERACT	homogenisation with other infrastructures	35	Generic
IS-ENES2	better understanding of interdisciplinary use cases and end-user requirements	35	Generic
LTER	harmonised data documentation	36	Generic
LTER	harmonisation of data and data flows	36	Generic
SeaDataNet	interoperability services and standards	36	Generic
ACTRIS	Interoperability between data centre nodes	44	Generic
EMBRC	Harmonisation of data between labs	32	Generic

RI	Requirement	Page in D5.1	Type
FixO3	harmonise data access	44	Generic
INTERACT	handling actual data concerning 76 active field-stations.	44	Generic
IS-ENES2	Share best practices as fast as new nodes integrate the RI federation.	45	Generic
FixO3	Heterogeneous data formats to enhance cross-community collaboration	45	Generic
FixO3	Harmonisation of datasets across distributed networks	45	Generic

TABLE 25 PUBLISHING REQUIREMENTS MAPPED TO DISCOVERY AND ACCESS (C.14)

RI	Requirement	Page in D5.1	Type
EMSO	Improved search is also desirable; currently expert knowledge is required, for example to be able to easily discover data stored in the MyOcean environment	32	Generic
IAGOS	Improve data discovery	34	Generic
EISCAT-3D	Data access with searching and visualisation	31	Generic
EMSO	Improving search is also desirable	46	Generic
FixO3	Improved search is also desirable	45	Generic

## Data processing requirements

There are 41 requirements which can be mapped to the publishing phase of the data life cycle. The high count in this case is due processing and workflow enactment requirements reported. The functionalities which were mapped to most of these requirements were workflow enactment (20 occurrences), and data processing control (19 occurrences), see Table 27 and Table 28. The high count related of workflow enactment is because it is the only functionality that mentions “provenance” which was an important requirement reported by RIs.

TABLE 26 PROCESSING REQUIREMENTS WHIT LOW COUNT

RI	Requirement	Page in D5.1	Type	ENVRI RM Functionality
ACTRIS	Data visualisation	44	Generic	D.7
EuroGOOS	Data assimilation	44	Generic	D.1

TABLE 27 PROCESSING REQUIREMENTS MAPPED TO SCIENTIFIC WORKFLOW ENACTMENT (D.6)

RI	Requirement	Page in D5.1	Type
EISCAT-3D	define workflows for data	31	Generic
EMBRC	Exploring new data workflows, which make use of marine biological and ecological data.	32	Generic
EMBRC	Exploring new data workflows, which make use of marine biological and ecological data.	32	Generic
EMSO	Collaborations with data processing infrastructures such as EGI for providing resources for infrastructure-side data processing.	32	Generic
IAGOS	provenance	34	Generic
ICOS	Collection and handling of provenance information	35	Generic
LTER	integration of common data repositories into their workflow system (including metadata documentation with LTER Europe DEIMS	36	Generic

RI	Requirement	Page in D5.1	Type
LTER	technical support on optimisation of data flows (e.g. integrating of a data repository, data integration portal)	45	Generic
N/A	knowing the evolutionary history of data – and at very different time scales – is important for any use and reuse of data	52	Provenance
N/A	Inter alia provenance to avoid undetected duplication (production or storage) of datasets.	52	Provenance
N/A	provenance data is essential and requires using a provenance recording system	53	Provenance
N/A	some provenance information about observation and measurement methods used is embedded within metadata but a real tracking tool still needs to be implemented	53	Provenance
N/A	Provenance system should enable following the data ‘back in time’ and see all the steps that happened from raw data collection, via quality control and aggregation to a useful product	53	Provenance
N/A	Provenance system should enable tracking the usage of the data, including information about users in order to understand the relevance of the data and how to improve their services	53	Provenance
N/A	Controlled vocabularies used for the descriptions of the steps for data provenance are required, some RIs already use research specific reference tables and thesauri like EnvThes and SeaDataNet common vocabularies	53	Provenance
N/A	Drawing an explicit line between metadata describing the ‘dataset’ and provenance information	53	Provenance
N/A	support on automated tracking solutions and or provenance management APIs to be applied in the specific e-science environments	53	Provenance
N/A	good overview of the existing vocabularies and ontologies that are ready to use or that need to be slightly adapted for specific purposes	53	Provenance
N/A	strong relationship between the task of identification of data and the provenance task, there must be a direct link between the data and its lineage that can be followed by the interested user	53	Provenance
N/A	defining a minimum information set that has to be tracked, finding a conceptual model for provenance which conforms to the needed information, maps existing models to the common model and finds a repository to store the provenance information	53	Provenance

TABLE 28 PROCESSING REQUIREMENTS MAPPED TO DATA PROCESSING CONTROL (D.9)

RI	Requirement	Page in D5.1	Type
EMSO	collaborations with data processing infrastructures such as EGI for providing resources for infrastructure-side data processing.	32	Generic
Euro-ARGO	The data may be converted/transformed on the ENVRI computation grid	33	Generic
EuroGOOS	Learning about other European RIs and getting inspiration from them for deciding on the general objectives and services that they could provide at European level	34	Generic
SIOS	lay the foundation for better-coordinated services for the international research community with respect to access to infrastructure, data and knowledge management, sharing of data, logistics, training and education	36	Generic
IAGOS	Data processing	44	Generic
IS-ENES2	Data near processing.	45	Generic
LTER	access to distributed data services	45	Generic
N/A	allow scientists to easily plug-in and experiment with their algorithms and methods without bothering with the computing platform	50	Processing
N/A	allow service managers to configure the platform to exploit diverse computing infrastructures	50	Processing



RI	Requirement	Page in D5.1	Type
N/A	perform analysis on data without substantial additional preparation	54	Optimisation
N/A	Take into account the overhead in time and effort required to prepare the data for processing	55	Optimisation
N/A	data staging, whereby data is placed and prepared for processing on some computational service	54	Optimisation
N/A	Given constraints on compute capacity, network bandwidth, and quality of service, the most important aspect to optimise is where should the data be processed	54	Optimisation
N/A	outputs of data processing need to be considered if the curation of results is within the scope of a given research infrastructure, and fold back into the domain of data curation	54	Optimisation
N/A	optimisation should be based on provenance because it allows predicting how data will be used and what infrastructure elements should provide access and processing capability over those data	55	Optimisation
N/A	Allow the investigator to configure the system based on their own experience and knowledge of the infrastructure	56	Optimisation
N/A	Allow the provider of a service or process to embed their own understanding in how the infrastructure operates	56	Optimisation
N/A	Allow encoding experts' knowledge within the system so it can then be accessed and applied automatically within the infrastructure.	56	Optimisation
N/A	to provide an abstraction layer over a number of individual research infrastructures and a number of shared services that they interact with	56	Optimisation

## Data use requirements

There are 28 requirements which can be mapped to the use phase of the data life cycle. The functionalities which were mapped to most of these requirements were authentication, authorisation, and accounting (8 occurrences), see Table 30. The majority of use requirements reported in the community support section of D5.1 (19 occurrences) cannot be mapped to existing functionalities (Table 31).

TABLE 29 USE REQUIREMENT MAPPED TO DATA VISUALISATION

RI	Requirement	Page in D5.1	Type	ENVRI RM Functionality
ACTRIS	Data visualisation	44	Generic	

TABLE 30 USE REQUIREMENTS MAPPED TO AUTHENTICATION, AUTHORISATION, AND ACCOUNTING (E.1, E.2, E.3)

RI	Requirement	Page in D5.1	Type	ENVRI RM Functionality
EPOS	improve the Interoperable AAI system (federated & distributed)	33	Generic	E.1, E.2, E.3
Euro-ARGO	The cloud account of the user is updated regularly with the new data provided	33	Generic	E.1, E.2
Euro-ARGO	An accounting of data provision and data delivery is performed	33	Generic	E.3
N/A	Access Control: AAI (Authentication and Authorisation Infrastructure) management is requested by many RIs. For example, IS-ENES2 currently uses OAuth2, OpenID, SAML and X.509 for AAI management.	57	Community Support	E.1, E.2
N/A	Accounting: tracking of user activities	57	Community Support	E.3

RI	Requirement	Page in D5.1	Type	ENVRI RM Functionality
N/A	Provide shared spaces (files and image repositories ) where members and stakeholders can upload/download and exchange files	58	Community Support	E.1, E.2, E.3
N/A	RI's need robust, fast-reacting systems, which offer security and privacy. Moreover, they need good performance for high data volumes	59	Community Support	E.1, E.2, E.3
N/A	Enforcement of data policy and licensing constraints that restrict access to a certain group of users	59	Community Support	E.1, E.2, E.3

**TABLE 31 USE REQUIREMENTS WHICH CANNOT BE MAPPED TO EXISTING FUNCTIONALITIES**

RI	Requirement	Page in D5.1	Type
Euro-ARGO	to design and pioneer access to and use of a cloud infrastructure with services close to European research data to deliver data subscription services	33	Generic
IS-ENES2	better understanding of interdisciplinary use cases and end-user requirements	35	Generic
LTER	developing a Data Integration Portal	36	Generic
N/A	allow third-party service providers to programmatically invoke the analytics methods	50	Processing
N/A	support scientists executing existing analytic tasks eventually customising/tuning some parameters without requiring them to install any technology or software	50	Processing
N/A	integration of tools and statistical methods, and their mapping onto platforms, should be supported in an appropriate virtual research environment or science gateway	51	Processing
N/A	anticipate user demand to minimise network congestion	56	Optimisation
N/A	optimisation of querying and data retrieval is of concern	56	Optimisation
N/A	direct users to the best replicas of a given dataset that ensures effective use of the underlying network	56	Optimisation
N/A	improved usability and easier to understand systems	56	Optimisation
N/A	Data portals provide (a single point of) access to the system and data products both for humans and machines (via APIs)	57	Community Support
N/A	Discovery of services and data facilities: metadata-based discovery mechanisms are commonly used	57	Community Support
N/A	use of a wiki to organise community information, and as a blackboard for collaborative work for community members	58	Community Support
N/A	facilitate communication to and from groups of community members using mailing lists, twitter & forums	58	Community Support
N/A	Shared calendars keep track and disseminate relevant events for community members	58	Community Support
N/A	Tools to organise meetings, events and conferences	58	Community Support
N/A	Website to disseminate community relevant information to all stakeholders	58	Community Support
N/A	Communication with all stakeholders (internal and external) supported through teleconferencing	58	Community Support
N/A	Helpdesk & Technical support for the use of complex data products which often require experience of, and detailed knowledge about, the underlying methods and science to be used in an optimal way	58	Community Support

## Appendix 5: Research campaigns

Some RIs are built to serve a research campaign, e.g., those measuring anthropogenic climate change, its impacts and the effectiveness of mitigation. Many are used by research campaigns, and some will initiate research campaigns to improve their effectiveness or efficiency. We define a *research campaign* as a *co-ordinated, resourced* and sustained effort to achieve a recognisable research *goal*. Research campaigns can be conducted both within a single organisation or (of more interest from the RIs and RM perspective) can involve multiple organisations. They can encourage *independent rival teams* to validate results, as in the detection of gravity waves. The anatomy and illustrative examples below, refine this definition and introduce concepts, highlighted in *italics*, that need to be understood in order to assess the changes necessary to the RM. The current RM will provide a framework for discussions with RIs and research campaigners to facilitate and record that refinement. Research campaigns have and need long-term stability; thereby increasing the return on investment for RIs and the RM. By considering research campaigns, we may also expose additional commonalities. Some critical features of a research campaign are introduced below and we initiate the discussion as to how they may be accommodated in the RM. Terms in *italics* are candidates for inclusion in the RM.

### Defining research campaigns and introducing their properties

A *research campaign* has a *goal*, for example, to better understand or better describe a natural phenomenon. Often the goal will be so ambitious that it needs to be broken down into a number of *sub-goals*, that may be pursued concurrently or sequentially. Some of these *sub-goals* will themselves require research campaigns; hence the concept is *recursive*. A research campaign will be organised, i.e.,

1. Its goal will be proposed, typically by a proposer or proposers, and agreed as significant by a research community that includes intellectual leaders, funders and publishers.
2. It will be addressed by sustained efforts, steered by research leaders who are supported by research organisations.
3. The leaders and organisations marshal research resources, schedule or direct effort and resources, and identify critical issues which require focused attention.
4. Quality controls are developed and agreed particularly with respect to the criteria for claiming a success.
5. A typically multi-disciplinary team builds with intensive internal communication, sharing and refining methods, code, data and equipment; but sometimes with intense rivalry between teams, which may be carefully kept separate as part of the success criteria.
6. Progress is reported in papers, and achievement of goals receives particular attention.
7. The information sharing among individuals and teams and procedures for publication and citation, are governed by agreed and enforced rules.

Research campaigns can be sustained for many years (more than 50 to confirm the Higgs boson and more than 100 years to detect gravity waves) and some require *global alliances*, e.g., establishing incontrovertible and convince evidence of the anthropogenic element of climate change. Such longevity and scope provides stability in the working practices and requirements that warrants attention from providers such as *RIs, technology R&D and supply*, and *e-Infrastructure services*.

### Anatomy of a research campaign

Many research campaigns develop and refine models of the phenomena of interest. They may use *numerical simulations* to expose the implications of *mathematical models*, often with substantial community investment in building sophisticated suites of *simulation code* and in establishing *specialist facilities*, e.g., major HPC and data centres. These simulation runs depend on the *collection and analysis of observations* to provide initial or boundary conditions, e.g., the global seismic network complying with FDSN agreements. They may be followed by analysis of the



differences between synthetic and observed values – so called "*misfit analysis*". The quantified misfit then implies *corrections* that need to be *propagated into the model*. Such *methodological patterns* are pursued repeatedly for sets of initial conditions and multiple observation comparisons until the model reaches stability or the limit of resolution in the context being studied.

Alternative reasons for terminating the iteration include: (a) enough evidence has been obtained, or (b) a resource, e.g., available time to guide emergency responders, is exhausted. The whole *iterative process* may then be applied in a new context where similar phenomena are believed to exist, or for new occurrences of the phenomena in the original context or both. The frequent repetition of such methods motivates investment in *optimisation* and *automation* of as much of each method as possible, often by adopting *workflow technologies*.

Many research campaigns interact with *governance* and *standards bodies* to establish *quality standards* and to facilitate *global collaboration*. These interactions are vital; the research often poses *new issues* that need a prompt *agreed response* to sustain *progress*. *Rules governing* a specific campaign or task within it often include a *commitment to adopt revisions* emerging from such bodies. Examples include, IPCC for climate change, IVOA for astronomy and FDSN for seismology. The agility of those bodies and the speed of adoption of their revisions is often critical to sustain momentum and to achieve credibility.

## Effective output from research campaigns

A crucial element of many campaigns is to convince identified *decision-makers* to believe in and appropriately *respond to the results*. This requires carefully tuned *presentations* of the *evidence* and of the *implications*. This communication also needs refinement and *optimisation*, which warrants research investment, particularly in contexts pertaining to natural hazards or societal challenges. This is a combination of *socio-political effort*, preparing *relevant citizens* so that they are aware of the risks, and understand the available *alerts* and their *interpretation*. It is also necessary to prepare local authorities and regional governments, as well as emergency responders, in such cases, so that they understand the information made available during an emergency, and there is a joint investment in tuning it to match needs and capabilities. This will depend on *mapping information* to popular *media channels*, as well as developing appropriate *visualisations*. These are specific examples of the need to go beyond the academic boundaries and develop *information to guide actions and decision makers*.

## Data-driven research-campaign patterns

An alternative pattern of research campaign elements is becoming prevalent as our wealth of data grows. Such research campaigns still have similar high-level goals and outputs, but the *collection of observations* dominates the agenda. Then various forms of *pattern recognition* are used over the collected data, such as *machine learning* or *citizen-science classification* campaigns<sup>29</sup>. Here again refinement and development of methods, e.g., the pattern-recognition techniques may require advances in statistics and advances in the machine learning algorithms, or in volunteer recruitment, motivation and management. Once again, the methods are often repeated as new data or new aspects of the data are investigated. Hence automation and optimisation becomes a worthwhile investment.

These two research campaign patterns interwork. The second yields observational input or comparators for the first; or it exposes evidence that provokes the formulation of new models for new forms of phenomena. For example, regular high-altitude meteorological observations revealed changes in ozone in the atmosphere above Antarctica. This provoked a recognition that the ozone hole was expanding, and led to an understanding of the relevant atmospheric chemistry and its anthropogenic origin. *Successful communication* then led to a global reduction in the release of chlorofluorocarbons (CFCs). The substitute refrigerants, hydrofluorocarbon (HFCs) have

---

<sup>29</sup> [www.zooniverse.org](http://www.zooniverse.org)



been found to have a significant GHG effect, and as a result there has been a recent agreement to stop their use.

Many research campaigns generate the need for data-intensive federations (DIF), see Section 4.2.3 of D5.1.

## Preliminary view of RM modelling research campaigns

The significant differences between research campaigns, and other research, e.g., individual efforts, projects, etc. is their scale and longevity, which imposes the need for more organisational structure, and co-ordination via explicit mechanisms and recorded rules. Consequently, the impact is expected to be significant in the Science and Information Viewpoints. To explore this further, two tables for each viewpoint are presented below (Table 32 - Table 39). The first in each pair (Table 32, Table 34, Table 36, Table 38) contains the entities that are needed by research campaigns; the second (Table 33, Table 35, Table 37, Table 39) the relationships between these. Each table has a column for concept name, and a column describing how it may be handled in the RM. The tables are populated by scanning the text above for italics denoting putative RM entity, and recording an entry for each that is significantly different – possibly naming it more abstractly. The refinement of these is future work. It will involve:

1. Rendering as a graph in the RM style to facilitate use as a framework for discussions.
2. Collection of examples and case studies from RIs.
3. Discussion with RIs and with those (recently) engaged in research campaigns.

TABLE 32: POTENTIAL CONCEPTS IN THE SCIENCE VIEWPOINT<sup>30</sup>

Name	Mapping to RM
Res. campaign	Long-lived evolving
Goal	Developed and agreed by the community
Team	Often multi-disciplinary, and many times multi-organisational
Coordinator	The role of an organisation such as IPCC
Community	The community addressing the goal via many relationships
Practitioner	Covering all the roles from proposer to technical support of many kinds including intellectual leaders, funders and publishers
Effort	A commitment of expertise, facilities and support for a significant period
Global Alliance	A formalised body related to the RC
Math Models	Formalisation of theories explaining processes and phenomena
Method	A scientific, observational or data handling procedure supported by workflows and working practises that achieves a specific (often repeated) step in the RC
Pattern	A repeated combination of methods that can be seen in different contexts
Decision makers	People not within the RC that the RC needs to influence
Preparation	Outreach, communication and socio-political effort to prepare decision makers

TABLE 33: POTENTIAL RELATIONSHIPS IN THE SCIENCE VIEWPOINT

Name	From	To	Mapping to RM
RCparticipantOrgs	Res. Camp.	Orgs	One to many organisations and institutions
RCgoal	Res. Camp.	Goal	One to one
RCcoord	Res. Camp.	Org.	1:1
RCresourcedBy	Res. Camp.	Org.	n:m organisations and institutions providing effort, facilities and funds; they become stakeholders
Independent	Team	Team	Symmetric 1:1 attribute denoting form of isolation
Subgoal	Goal	Goal	1:n
RCproposer	Res. Camp.	Practitioner	1:n
RCcommunity	Res. Camp.	Community	n:1 agree and endorse the RC(s)

<sup>30</sup> At present the order of entries in these tables has not been considered; read them as sets.

CommittedEffort	Res. Camp.	Effort	m:n commitment to RC; attributes duration & quantity
Contribution	Effort	Org./Pract.	m:n promise of sustained engagement
RCledBy	Res. Camp.	Practitioner	1:n leads and steers the campaign or m:n where leadership is shared
RCregime	Res. Camp.	Rule	The set of rules that govern behaviour of those conducting the campaign
Collaboration	Pract. or Team	Pract. or Team	Commitment to combine and share assets, such as thinking, software, methods, information, etc.
RCalliances	Res. Camp.	GlobalAllian	m:n global alliances shaping the RC's context
InSilicTest	MathModel	Simulation	m:n assessment of match between model and reality
Workhorse	pattern	model	m:n a proven effective pattern for refining models
InfWorkhorse	Pattern	influence	m:n a proven effective pattern for achieving influence

TABLE 34: POTENTIAL CONCEPTS IN THE INFORMATION VIEWPOINT

Name	Mapping to RM
Success Crit.	Definition of how to recognise Goal has been met
Rule	Agreed by community under leader's guidance, communicated to and adopted by the community
Publication	External public and formal persistent records
Announcement	Output to media and stakeholders on significant achievements
Internal Reps.	Reports whose circulation is limited by prevailing rules
Observation	Primary data or any derivative and composition thereof
SimResults	Collected and possibly post processed output from simulations
Corrections	Data to shape adjustments to a mathematical model
ActionableInfo	Information with sufficient evidence, preparation and presentation to trigger appropriate reactions in decision makers

Note: Digital representations of organisations, persons, etc. must be considered but this should be taken care of by support for CERIF.

TABLE 35: POTENTIAL RELATIONSHIPS IN THE INFORMATION VIEWPOINT

Name	From	To	Mapping to RM
RCcriteria	RC	Succ. Crit	
QArequired	Rule	QAMethod	1:n the conditions to be passed before proceeding
PublicationAdjudic	Publication	Rule	m:n progress through adjudication to approval
AnnouncementAgr	Announcement	Rule	m:n progress through adjudication to agreement
AnnouncePub	Announcement	Publication	1:n these publications back up and follow up an announcement
TechRepRules	InternalRep	Rule	m:n controls on internal info.
Influence	ActionableInfo	DecisionMk	m:n Series of messages tuned to recipient decision makers

TABLE 36: POTENTIAL CONCEPTS IN THE COMPUTATION VIEWPOINT

Name	Mapping to RM
QAMethod	Workflows and procedures that must be completed before such things as using data, publishing or declaring success
TechProviders	Producers of required technical capabilities and operational resources
Simulations	Processes running simulation suites with particular parameters
Platform	Any computational platform with specified relationships to volatile and persistent memory, its software stack and specialised facilities
Workflow	A mainly automated and optimised procedure that usually includes computation, data handling and inter-site communication.

MetaWorkflow	A workflow generator by whatever mechanism to implement a pattern
Classification	Means of categorising observations or outputs
Machine learning	Computational means of performing classification, inference and pattern matching
CitizenClassification	Volunteer procedures for classification, inference and pattern matching
D-I platform	A platform giving high performance, or high throughput for data handling and data analysis
Visualisation	Methods for transforming data into forms easily understood by target groups

TABLE 37: POTENTIAL RELATIONSHIPS IN THE COMPUTATION VIEWPOINT

Name	From	To	Mapping to RM
Dependence	RC	TechProv	m:n the commitments to provide, performance, persistence, sustained operation, bandwidth, etc. all quantified for bounded periods
SimReq	Simulation	Platform	m:n with quantified duration and share
Misfit	Observation	SimResults	m:n analytic or human comparison
CorrectionGen	Misfit	Corrections	m:n the corrections to the model
MetodWF	Method	Workflow	m:n a software dominated workflow that implements the method with specified forms of practitioner input
PatternGen	MetaWF	Pattern	A means of using the meta workflow to generate instances of the pattern

TABLE 38: POTENTIAL CONCEPTS IN THE ENGINEERING VIEWPOINT

Note: This table to be filled in when further work on the Engineering Viewpoint has been completed.

Name	Mapping to RM

TABLE 39: POTENTIAL RELATIONSHIPS IN THE ENGINEERING VIEWPOINT

Note: This table to be filled in when further work on the Engineering Viewpoint has been completed.

Name	From	To	Mapping to RM



## Appendix 6: Ideas to facilitate RIs engagement

This appendix contains a list of idea, generated during the third ENVRIweek project meeting that took place in Prague, 14-18<sup>th</sup> November 2016 for engaging the interest and commitment of the RIs technical experts to the ENVRI RM.

This appendix is organised into several categories of ideas, as follows:

- Primers, tutorials and practical examples
- Training, consulting, and helpdesk support
- Community building and self-help
- Tools for design
- Interaction more directly with the RIs
- KISS (Keep It Simple Stupid)
- Experiential reporting and case studies
- Focus
- Improve the RM to make it more relevant

### Primers, tutorials and practical examples

Easy to understand ‘primer’ for the RM, to understand the basic concepts, with practical examples. = “How to start with it, simply”. Easy to understand guides. Example based. Use of actual problems to illustrate.

More examples and use cases; illustrations, tools, linked to examples

Have a detailed list (glossary) with explanations of terms and acronyms.

Mapping the use cases to the concepts of RM.

Hold 30 minutes conversation with RM experts and see how RI business or community (n RI) fits in RM.

Designate some basic ‘starting points’ that allow users to see the correspondences arising between viewpoints for some core concepts without having to tackle the entire model at once.

### Training, consulting, and helpdesk support

Consulting style training which shows targeted solutions.

Training workshops (physical or webinars) to introduce key concepts, terminology and barriers

Workshop for discussions.

Attend domain meeting to make sure developments are aligned.

Motivate the interest of the RI to the RM. If time is needed from them they must be convinced that they will benefit.

Training program, perhaps leading to certification

Supported from a helpdesk (ideally available 24/7).

Develop/deliver webinars that help people to understand patterns, tools, etc.

### Community building and self-help

RM champions, within each RI one person with an understanding and experience.

Meetings (within RI staffs) to discuss how to make it easy to get started.

Make it easy to find and express the issues for broad range of users.

Online discussion forum. Google group to help build community and knowledge base / FAQ.



Centres of excellence e.g., this RI is good at acquisition and curation so it would be the poster boy for acquisition/curation integration.

Apply RM to an RI and gain immediate feedback from the group when problems pop up.

### **Tools for design**

Use a tool with options to choose for each viewpoint. Visual tools.

Good example, visually enhanced on-line.

Share developed tools in github and try to converge to common code-base.

Share developed models in github and try to converge to common models.

In addition to UML, suggest to use less technical diagrams e.g., concept maps to avoid the message/impression that this is for IT people.

Provide ready-made templates that you think cover the kinds of topics they care about. Support these with tools.

Use linguistic notation as well as diagrams.

Good GUI plus case studies.

Develop a logic that is computable from the linguistic form of the RM.

Develop visualisation and consistency checking aids that depend on that logic, and cope with scale. Focus over the integrated viewpoints. People need to understand the integrated whole from the starting point they are at and to be able to drill down.

Create tools that allow RIs to build their descriptions incrementally, based on some underlying model.

Easy tool.

### **Interaction more directly with the RIs**

Directly interact with the right people inside the different RIs.

Carry out a skills appraisal within RIs to identify where and whether system architects exist, thus to establish contacts and/or understand the gap.

Sit/lock the architects of different RIs in one room and collect their common ideas.

Speak to the domain specific experts and not only to the computer scientists.

Find out where RIs have problems with their current work and introduce tools, patterns and procedures that will be helpful.

### **KISS (Keep It Simple Stupid)**

Keep it simple and abstract; details are for the implementation of RIs.

Focus on the simple but fundamental issues.

Set of basic 'lego' building blocks that connect with one 'snap'.

Keep the documentation manageable to read and understand for groups from broad background.

Work towards a clean, simple structural RM with examples. It should be inviting.

Provide as many good examples as possible of applying the RM in different typical contexts; based on actual problems.

Describe processes / aspects that are critical/important to the people you are attempting to communicate with. In such descriptions, the terms link to concepts in the RM but you don't need



to say that. Then develop a dialogue refining what they do, and see how vocabulary and definitions develop.

Provide step-wise examples as technology will advance and examples will age/decay.

Template with different colours for each viewpoint. User must fill everything with help of online guidance. Compare with, for example DMPonline tool as an approach.

### **Experiential reporting and case studies**

Ask non-RM 'insiders' to report on their experiences.

Cookbook of recipes for common scenarios.

How the RM can fit with already existing components of the RIs. The AtlantOS project/case is a very nice use case for this purpose. Similar to what DASSH use case is looking at.

AtlantOS use case is a typical data integration example/challenge that RIs are facing. In the near future, RI are asked to offer services on the aggregated data (beyond search and browse).

Explain different ways in which RM can be used e.g., to design new RI, when upgrading an RI, or other purposes. Practical guide to these possibilities is important.

### **Focus**

Focus on the appropriate interfaces and stop considering that RIs have nothing in place.

Invite the RI to design the interface.

Tackle the system of RI services.

Push RIs data and metadata on EU cloud infrastructures and build inter-disciplinary operations.

Link RM with OIL-E with help of B2ANNO service. Annotate while reading RI data management documents.

Establish wiki platform of RM users.

### **Improve the RM to make it more relevant**

Improve the RM from D5.4, D8.1, D8.3.

Consider that RI data system already exists/works and therefore focus on the publishing interfaces to facilitate interoperability with other RIs.

Make clear the functional and non-functional aspects of the interface, especially for interoperation.

## Appendix 7: ENVRI Reference Model version 2.1, November 2016 – Snapshot

This appendix (118 pages following the present one) is a snapshot (export) from the ENVRI RM section of the wiki <http://envri.eu/rm>.

It is the record of the state of version 2.1 of the ENVRI RM on 5<sup>th</sup> January 2017.

Note 1: The difference between this snapshot and that of version 2.1 of 9<sup>th</sup> November 2016, referred to in the main body of the present document is the removal of Example 6 from the section 'Guidelines for using the Reference Model'. This has been done at the request of the relevant RI to avoid out of date information being given.

This deletion of a guidance example is not a material change to the RM *per se* and thus no change of version number has been needed.

Note 2: Page numbers in the bottom right corner of the following 118 pages of this appendix refer to the Table of Contents that appears on the next page, and not to the Table of Contents of the main document.



1. ENVRI Reference Model	2
1.1 Download of ENVRI Reference Model	3
1.2 Getting started with the ENVRI RM	4
1.3 The ENVRI and ENVRIplus Projects	5
1.4 Introduction	5
1.5 Model Overview	9
1.6 The ENVRI Reference Model	14
1.6.1 Science Viewpoint	14
1.6.1.1 SV Communities	15
1.6.1.2 SV Community Roles	15
1.6.1.3 SV Community Behaviours	20
1.6.2 Information Viewpoint	24
1.6.2.1 IV Components	24
1.6.2.1.1 IV Information Objects	25
1.6.2.1.2 IV Information Action Types	33
1.6.2.2 IV Information Objects Lifecycle	36
1.6.2.2.1 IV Lifecycle Overview	36
1.6.2.2.2 IV Lifecycle in Detail	38
1.6.2.3 IV Information Management Constraints	42
1.6.3 Computational Viewpoint	44
1.6.3.1 CV Objects	44
1.6.3.1.1 CV Presentation Objects	45
1.6.3.1.2 CV Broker Objects	47
1.6.3.1.3 CV Service Objects	48
1.6.3.1.4 CV Component Objects	50
1.6.3.1.5 CV Back End Objects	52
1.6.3.2 CV Objects and Subsystems	53
1.6.3.2.1 CV Data Acquisition	53
1.6.3.2.2 CV Data Curation	55
1.6.3.2.3 CV Data Publishing	57
1.6.3.2.4 CV Data Processing	59
1.6.3.2.5 CV Data Use	62
1.6.3.3 CV Integration points	64
1.6.3.3.1 CV Brokered Data Export	64
1.6.3.3.2 CV Brokered Data Import	65
1.6.3.3.3 CV Brokered Data Query	66
1.6.3.3.4 CV Instrument Integration	66
1.6.3.3.5 CV Citation	67
1.6.3.3.6 CV Raw Data Collection	68
1.6.3.4 How to read the Model (Computational Viewpoint)	68
1.6.3.5 How to use the Model (Computational Viewpoint)	71
1.7 Conclusions and Future Work	72
1.8 Appendix A Common Requirements of Environmental Research Infrastructures	73
1.9 Appendix B Terminology and Glossary	75
1.10 Appendix C Notation	81
1.10.1 Notation of Science Viewpoint Models	81
1.10.2 Notation of Information Viewpoint Models	85
1.10.3 Notation of Computational Viewpoint Models	96
1.10.4 UML4ODP Graphical Notation	99
1.11 Bibliography	102
1.12 Guidelines for using the Reference Model	103
1.12.1 Example 1: Using the Reference Model to Guide Research Activities (EISCAT 3D - EGI)	104
1.12.2 Example 2: Using the Reference Model as an Analysis Tool (EUDAT)	109
1.12.3 Example 3: Using the Reference Model in documentation (EMSO)	110
1.12.4 Example 4: Using the Reference Model as design reference (EPOS)	112
1.12.4.1 EPOS/ENVRI Modelling	114
1.12.5 Example 5: Using the Reference Model to explain the technology details of common services (WP4 practices)	114

# ENVRI Reference Model

This is the home of the ENVRI Reference Model v2.1, published 09th November 2016 and guidelines on how to use it. Click on the navigation links to the left, or search using the search box above.

This space is under active development. If you find something incorrect or missing, or something that is not clearly explained, please tell us by emailing us at [<envri-rm@list.uva.nl>](mailto:envri-rm@list.uva.nl).

- [Analysis of Common Requirements](#)
- [Guideline for Using the Reference Model](#)
- [Video Tutorials](#)
- [Publications](#)
- [Award](#)
- [Articles, Posters and Presentations](#)
- [ENVRI Reference Model Flyer](#)

## Analysis of Common Requirements

The ENVRI Reference Model is originally based on a pre-study of 6 ESFRI Environmental Research Infrastructures (RI), carried out as part of the ENVRI project. It has been updated during the ENVRIplus project from the results of a study of these original 6 and a further 13 RIs. The reports of these studies can be downloaded as follows:

Requirements studies	Notes	Date	Authors	Download
ENVRIplus deliverable D5.1: A consistent characterisation of existing and planned RIs	A version of the study carried out during the ENVRIplus project, with minor editorial corrections beyond the version submitted to the European Commission.	24 May 2016	Malcolm Atkinson (UEDIN) et al.	<a href="#">[.docx]</a> <a href="#">[.pdf]</a>
ENVRI deliverable D3.3: Analysis of Common Requirements For ENVRI Research Infrastructures V1.0 (Final)	A final version report of the study carried out during the ENVRI project, as submitted to the European Commission.	01 May 2013	<a href="#">Yin Chen</a> (CU)	<a href="#">[.doc]</a> <a href="#">[.pdf]</a>

## Guideline for Using the Reference Model

Versions	Notes	Date	Authors	Download
Guideline for Using the Reference Model (Final)	A final version submitted to the European Commission. These are the original guidelines, produced during the ENVRI project. They are still relevant but have been supplemented with other materials more recently.	30/09/2013	Yin Chen (CU), Barbara Magagna (EAA), Paul Martin (UEDIN), Alex Hardisty(CU), Alun Preece(CU), Herbert Schentz(EAA), Zhiming Zhao(UvA), Robert Huber(UniHB), Ingemar Haggstrom (EISCAT), Ville Savolainen(CSC), Malgozata Krakowian(EGI.eu)	<a href="#">[.doc]</a> <a href="#">[.pdf]</a>

## Video Tutorials

- ENVRI Reference Model: an Overview. [\[.ppt\]](#)
- Main Processes of the ENVRI Reference Model – Corresponding Viewpoint [\[.ppt\]](#)

## Publications

- Martin P, Chen Y, Hardisty A, Jeffery K, and Zhao Z. (2016) Research data infrastructures for environmental related societal challenges. In: Terrestrial Ecosystem Research Infrastructures: Challenges and Opportunities. Eds. Chabbi A, and Loescher HW. *October 1, 2016 Forthcoming* by CRC Press ISBN 9781498751315.
- Zhao Z, Martin P, Grosso P, Los W, de Laat C, Vermeulen A, Jeffrey K, Castelli D, Hardisty A, Legre Y, Kutsch W. (2015) Reference Model Guided System Design and Implementation for Interoperable Environmental Research Infrastructures. Presented at: e-Science 2015: IEEE 11th International Conference on e-Science, Munich, Germany, 31 August - 4 September 2015. e-Science (e-Science), 2015 IEEE 11th International Conference on. IEEE, pp. 551-556. doi: [10.1109/eScience.2015.41](https://doi.org/10.1109/eScience.2015.41) Near-final text: [\[.pdf\]](#)
- Martin, P., Grosso, P., Magagna, B., Schentz, H., Chen, Y., Hardisty, A., Los, W., Jeffery, K., de Laat, C., Zhao, Z. (2015) Open Information Linking for Environmental Research Infrastructures. Presented at: IEEE 11th International Conference on e-Science, Munich, Germany, 31 August 2015 - 4 September 2015. e-Science (e-Science), 2015 IEEE 11th International Conference on. IEEE, pp. 513-520. doi: [10.1109/eScience.2015.66](https://doi.org/10.1109/eScience.2015.66) Near-final text: [\[.pdf\]](#)
- Chen, Y., Martin, P., Schentz, H., Magagna, B., Zhao, Z., Hardisty, A., Preece, A., Atkinson, M., Huber, R. & Legre, Y. (2013), "A Common Reference Model for Environmental Science Research Infrastructures", in the *Proceedings of the 27th Conference on Environmental Informatics* 2013, p665-673, 2013. [\[.pdf\]](#)

- Chen, Y., Hardisty, A., Preece, A., Martin, P., Atkinson, M., Zhao, Z., Magagna, B., Schentz, H. & Legre, Y. (2013). "Analysis of Common Requirements for Environmental Science Research Infrastructures", in the *Proceeding of Science (PoS) SISSA, PoS(ISGC 2013)032* [[pdf](#)]
- Zhao, Z., Grosso, P. & Laat, C. de (2012). "OEIReference Model: An Open Distributed Processing based Interoperability Reference Model for e-Science", *Cloud&Grid interoperability workshop*, Gwangju, Korean, 2012.
- Zhao, Z., van der Ham, J., Taal, A., Koning, R., Dumitru, C., Wibisono, A., Grosso, P., de Laat, C. (2012). "Planning data intensive workflows on inter-domain resources using the Network Service Interface (NSI)", *the 7th Workshop on Workflows in Support of Large-Scale Science, in the context of Supercomputing*, Salt Lake City, 2012;
- Zhao, Z., Dumitru, C., Grosso, P. & Laat, C. de (2012). "Network resource control for data intensive applications in heterogeneous infrastructures", *26th IEEE International Parallel and Distributed Processing Symposium*, Shanghai, 2012.
- Jiang, W., Zhao, Z., Grosso, P., de Laat, C., (2013) Dynamic workflow planning on programmable infrastructure, *IEEE Network Architecture Storage*, China 2013.

## Award

- 1 of 3 [Lightening talks in the EGI Community Forum 2014](#), Helsinki, Finland, 19-23 May 2014. [[pdf](#)]

## Articles, Posters and Presentations

- Nieva de la Hidalga, A., and Hardisty, A. (2016), "How the ENVRI Reference Model helps to design Research Infrastructures", *ENV RIplus Newsletter No.2, May 2016*. [[link](#)] [[pdf](#)]
- Hardisty, A. (2015). "Reference Models: What are they and why do we need them?", *Blog post, 8th July 2015* <https://alexhardisty.wordpress.com/2015/07/08/reference-models-what-are-they-and-why-do-we-need-them/>.
- Chen, Y., Hardisty, A. (2014), "A Common Reference Model for Environmental Research Infrastructures", *iLEAPS newsletter, Special issue, September 2014, page 17-19* [[pdf](#)]
- Chen, Y. ""Using the Reference Model in ICOS Research Infrastructure Design Study -- Updates on Science Viewpoint", *ICOS Interim Scientific Advisory Board, Sep 2014*. [[pdf](#)]
- Chen, Y., B. Magagna, P. Martine (2014), "Using the Reference Model in ICOS Research Infrastructure Design Study", *ICOS Community, Jun 2014*. [[pdf](#)]
- Chen, Y., (2013), "ENVRI, Common Operations of Environmental Research Infrastructure", *Data Science Symposium 2013*. [[link](#)]
- Chen, Y., Häggström, I., Mann, I., Heinselman, C., (2013), "EISCAT 3D incoherent scatter radar system", *Data Science Symposium 2013*. [[link](#)]
- Chen, Y., Häggström, I., Hardisty, A., Sipos, G., Krakowian, M., Ferreira, N. L., Savolainen, V. (2013). "Towards the Big Data Strategies for EISCAT-3D", *EISCAT International Symposium 2013*, Lancaster, the UK, 2013. [[pdf](#)]
- Häggström, I., Chen, Y., Hardisty A., Sipos, G., Krakowian, M., Ferreira, N., & Savolainen, V. (2013). "Towards the Big Data Strategies for EISCAT-3D", *Radiovetenskap och Kommunikation 2013: Generation, Real-Time Processing, Transport, Distribution and Management of Large Raw Data Volumes in the Physical Sciences*.11 - 12 November 2013, KVA, Royal Academy of Sciences, Frescati, Stockholm, 2013. [[link](#)]
- Preece, A. (2013). "The ENVRI Reference Model", *Bulding Global Partnerships - RDA Second Plenary Meeting*, Washington DC, US, 16-18 Sep 2013. [[Poster](#)]
- Zhao, Z., Grosso, P., Los, W., de Laat, C., Chen, Y., Hardisty, A., Martin, P., Herbert, S. & Barbara, M., " OEILM: a semantic linking framework for environmental research infrastructures", *9th IEEE International Conference on eScience 2013*, Beijing, China, 2013. [[Poster](#)]
- Zhao, Z., Grosso, P., Los, W., de Laat, C., Chen, Y., Hardisty, A., Martin, P., Herbert, S. & Barbara, M., " OEILM: a semantic linking framework for environmental research Infrastructures", *Supercomputing 2013*, Dutch exhibition booth. [[Poster](#)]
- Zhao, Z., Grosso, P., Los, W., de Laat, C., Chen, Y., Hardisty, A., Martin, P., Herbert, S. & Barbara, M., " OEILM: a semantic linking framework for environmental research infrastructures", *Dutch ICT 2013*. [[Poster](#)]
- Legre, Y. (2013). "Contributions of Environmental Research Infrastructure to GEOSS", *Presentation in GEO European Projects Workshops 2013*, Barcelona, Spain, 2013. [[ppt](#)]

## ENVRI Reference Model Flyer

- [[pdf](#)] HD
- [[pdf](#)] For Professional Printing Service

*Research data infrastructures for environmental related societal challenges.*

*Martin P, Chen Y, Hardisty A, Jeffery K, and Zhao Z. In: Terrestrial Ecosystem Research Infrastructures: Challenges and Opportunities. Eds. Chabbi A, and Loescher HW. October 1, 2016 Forthcoming by CRC Press ISBN 9781498751315.*

## Download of ENVRI Reference Model

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

From time to time new versions of the document will be released which are the snapshots of development milestones.

Versions	Notes	Date	Authors	Download



ENVRI Reference Model V2.1	Version 2.1, incorporating further changes arising from ENVRIplus requirements analysis and the new larger community of Environmental Research Infrastructures represented in the ENVRIplus project. Full details of all changes are documented in deliverable D5.2 (linked to be provided).	09 Nov 2016	Abraham Nieva de la Hidalga (CU), Barbara Magagna (EAA), Markus Stocker (UniHB), Alex Hardisty (CU), Paul Martin (UvA), Zhiming Zhao (UvA), Malcolm Atkinson (UEDIN)	[.pdf]
ENVRI Reference Model V2.0	Version 2.0, incorporating changes arising from ENVRIplus requirements analysis and realignment along lines of data lifecycle model	27 July 2016	Abraham Nieva de la Hidalga (CU), Barbara Magagna (EAA), Markus Stocker (UniHB), Alex Hardisty (CU), Paul Martin (UvA), Zhiming Zhao (UvA), Malcolm Atkinson (UEDIN)	[.pdf]
ENVRI Reference Model V1.1	The second major version	30 Aug 2013	Yin Chen (CU), Paul Martin (UEDIN), Herbert Schentz (EAA), Barbara Magagna (EAA), Zhiming Zhao (UvA), Alex Hardisty (CU), Alun Preece (CU), Malcolm Atkinson (UEDIN)	[.doc][.pdf]
ENVRI Reference Model V1.0 (Final)	A final version ready to submit to the European Commission	01 May 2013	Yin Chen (CU), Paul Martin (UEDIN), Herbert Schentz (EAA), Barbara Magagna (EAA), Zhiming Zhao (UvA)	[.doc][.pdf]
ENVRI Reference Model V1.0 (Draft)	A draft version ready for Internal review	01 Apr 2013	Yin Chen (CU), Paul Martin (UEDIN), Herbert Schentz (EAA), Barbara Magagna (EAA), Zhiming Zhao (UvA), Alex Hardisty (CU), Alun Preece (CU), Malcolm Atkinson (UEDIN)	[.doc][.pdf]

## Getting started with the ENVRI RM

**The ENVRI Reference Model (ENVRI RM, RM) exists to illustrate common characteristics of environmental science research infrastructures in order to provide a common language and understanding, promote technology and solution sharing and improve interoperability.**

## About

Independent development of research infrastructures leads to unnecessary replication of technologies and solutions whilst the lack of standard definitions makes it difficult to relate experiences in one infrastructure with those of others. The ENVRI Reference Model (ENVRI RM) uses Open Distributed Processing (ODP) in order to model the "archetypical" environmental research infrastructure. The use of the ENVRI RM to illustrate common characteristics of existing and planned European Environmental Research Infrastructures from a number of different perspectives provides a common language for and understanding of those infrastructures, promotes technology and solution sharing between infrastructures, and improves interoperability between implemented services.

## Intended Audience

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 conceptual architecture that underpins the construction of such research infrastructures. In particular, the intended primary audience the Reference Model includes [33]:

- Research Infrastructures 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 Reference Model is also intended for research infrastructure leaders and service centre staffs.

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

## Document Structure

**Introduction** introduces the motivation and background knowledge of the ENVRI RM.

**Model Overview** presents an overview of the ENVRI RM against the backdrop of a typical lifecycle for research data.

**The ENVRI Reference Model** is a detailed description of the ENVRI RM from the Open Distributed (ODP) Viewpoints perspectives.

**Conclusions and Future Work** concludes this work.

Appendices are not formally part of the reference model. They provide additional information that may be helpful and for the convenience of the reader.

**Appendix A** presents the full list of the required functionalities that is the result of the investigations of the common requirements of Research Infrastructures.

**Appendix B** is a glossary of terms, and consists of concepts and terms defined throughout the ENVRI RM.

## How to Read

- The primary audience of the ENVRI RM should generally read the whole documentation, starting with the **Introduction** and **Model Overview**. Such readers then should proceed to the **Science Viewpoint** and the **Information Viewpoint** before looking at the **Computational Viewpoint**. It is not necessary to read everything nor to read in order. The tutorials given below are useful entry points. Elsewhere (link to be provided) we give detailed guidance on how best to engage with the Reference Model for different purposes.
- The leaders of research infrastructures, and service centre staff may want to read the introduction and background knowledge in **Introduction** and **Model Overview**.
- Readers who have general interests in the ENVRI RM may want to read **Introduction**.

## Tutorial

- **Tutorial one**: ENVRI Reference Model: an Overview
- **Tutorial two**: Main Processes of the ENVRI Reference Model – Corresponding Viewpoint

## The ENVRI and ENVRIplus Projects

Frontier environmental research increasingly depends on a wide range of data and advanced capabilities to process and analyse them. The original ENVRI project, "Common Operations of Environmental Research infrastructures" (2011 - 2014) was 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 are intended to speed up the construction of these Environmental Sciences research infrastructures and to allow scientists to use the data and software from each facility to enable multi-disciplinary science. The work is continuing (2015 - 2019) as part of the **ENVRIplus project "Environmental Research Infrastructures Providing Shared Solutions for Science and Society"**.

The focus is on developing common capabilities including software and services for environmental and e-infrastructure communities. While the Environmental Sciences research 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. Common standards, deployable services and tools will be adopted by each infrastructure as it progresses into its construction phase.

The ENVRI and ENVRIplus projects deliver a common reference model, the "ENVRI Reference Model" or "ENVRI RM" created by capturing the functional and other capabilities of each ESFRI-ENV infrastructure. This model and the development driven by the testbed deployments result in ready-to-use systems that can be integrated into the environmental research infrastructures.

The projects put 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.

## Introduction

- **Purpose and Scope**
- **Rationale**
- **Basis**
- **Approaches**
- **Conformance**
- **Related Work**
  - **Related Concepts**
  - **Related Reference Models**
    - **Committee Reference Models**
    - **Consensus Reference Models**
    - **Consultation Reference Models**
  - **Other Related Standards**

## Purpose and Scope

All research infrastructures for environmental sciences (the so-called 'ENVRI's') although very diverse, have some common characteristics, enabling them potentially to achieve a greater level of interoperability through the use of common standards and approaches for various functions. The objective of the ENVRI Reference Model is to develop a common framework and specification for the description and characterisation of computational and storage infrastructures. This framework can support the ENVRI's to achieve seamless interoperability between the heterogeneous resources of their different infrastructures.

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

- to provide a way for structuring thinking that helps the community to reach a common vision;
- to provide a common language that 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.

The present wiki / document describes the ENVRI Reference Model which:

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

The Reference Model provides an abstract logical conceptual model. It does not impose a specific architecture. Nor does it impose specific design decisions or constraints on the design of an infrastructure.

The *initial* model (versions 1.0 and 1.1) focused on the urgent and important issues prioritised for ENV research infrastructures including data preservation, data discovery and access, and data publication. It defines a minimal set of functionalities to support these requirements. The *initial* model does not cover engineering mechanisms or the applicability of existing standards or technologies.

Version 2.x of the model incrementally extends these core functionalities:

- Version 2.0 is a simplification of the way the Reference Model is presented, to make it easier to understand and become familiar with. Version 2.0 explicitly aligns the RM with a lifecycle oriented view of research data management.

## Rationale

Environmental issues will dominate the 21<sup>st</sup> century [2]. Research infrastructures that provide advanced capabilities for data sharing, processing and analysis enable excellent research and play an ever-increasing role in the environmental sciences as well as in solving societal challenges. The **ENVRIplus project** and its predecessor ENVRI project gathers many of the EU ESFRI and other environmental infrastructures (**ICOS**, **EURO-Argo**, **EISCAT-3D**, **LifeWatch**, **EPOS**, **EMSO**, etc.) to find common solutions to common problems, including use of common software solutions. The results, including the ENVRI Reference Model will accelerate the construction of these infrastructures and improve interoperability among them. The experiences gained will also benefit 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 infrastructures 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.

## Basis

The ENVRI Reference Model is built on top of the Open Distributed Processing (ODP) framework [4, 5, 6, 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 that 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 serves as 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; For better communication with 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.

## 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 deliverables:

- D3.2: 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.

## Conformance

A conforming environmental research infrastructure should support the common functionalities described in **Model Overview** and the functional and information model described in **The ENVRI Reference Model**.

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 functions 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.

## Related Work

### 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 [13]. 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 **knowledge base**), 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 model 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 of building the model and technologies the model captures.

## Related Reference Models

### *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].

There are no reference model standards yet for environmental science research infrastructures.

### *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.

### 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.

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.

### 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].

## Model Overview

- [The Research Data Lifecycle within Environmental Research Infrastructures](#)
  - [Data Acquisition](#)
  - [Data Curation](#)
  - [Data Publishing](#)
  - [Data Processing](#)
  - [Data Use](#)
- [Lifecycle Support Inter- and Intra- Research Infrastructure Relationships](#)
- [Common Functions within a Common Lifecycle](#)

### The Research Data Lifecycle within Environmental Research Infrastructures

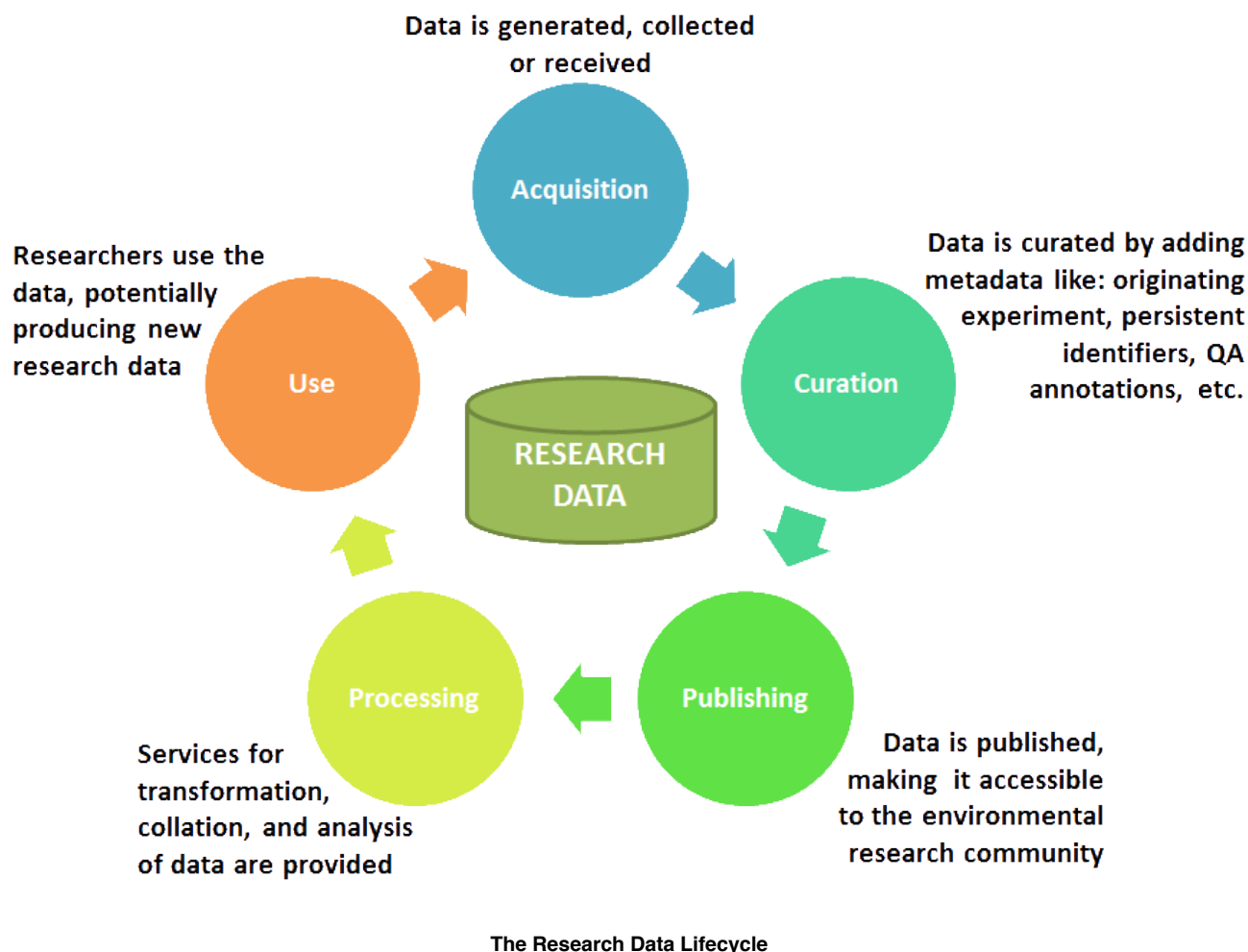
The ENVRI and ENVRIplus project investigated a collection of more than 20 representative environmental research infrastructures (RIs) from different areas. By examining these research infrastructures and their characteristics, a common data lifecycle was identified. The data lifecycle is structured in five phases: *Data Acquisition*, *Data Curation*, *Data Publishing*, *Data Processing* and *Data Use*

The fundamental reason of the division of the data lifecycle is based on the observation that all applications, services and software tools are designed and implemented around five major activities: acquiring data, storing and preserving data, making the data publicly available, providing services for further data processing, and using the data to derive different data products. This data lifecycle is fairly general and all research infrastructures investigated exhibit behaviour that aligns with its phases. Consequently, the ENVRI-RM is structured in line with the five phases of the data life-cycle.

This lifecycle begins with the acquisition of data from a network of integrated data collecting entities (seismographs, weather stations, robotic buoys, human observers, or simulators) which is then registered 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 parties external to the infrastructure, as well as to services within the infrastructure. This results in a natural partitioning into data acquisition, curation and publishing. In addition, RIs may provide services for

processing data, the results of this processing can then produce new data to be stored within the infrastructure. Finally, the broader research community outside of the RI can design experiments and analyses on the published data and produce new data, which in turn can be passed to the same RI or to other RI for curation, publishing and processing, restarting the lifecycle.

The activities of each research infrastructure can align with this lifecycle. However, research infrastructures will tend to optimise and concentrate more on some phases. For instance, some research infrastructures concentrate mostly on the acquisition of data, while others focus their expertise on curation or publishing. ENVRI RM assumes that the research infrastructures can complement and integrate with each other to support the entire data lifecycle. Integration is achieved through providing a set of capabilities via interfaces invoked within systems (or subsystems) which can be used within the infrastructures but also across boundaries. In the ENVRI RM, an interface is 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.



## Data Acquisition

*In the **data acquisition phase** the research infrastructure collects raw data from registered sources to be stored and made accessible within the infrastructure.*

The data acquisition phase supports collecting raw data from sensor arrays and other instruments, as well as from human observers, and brings those data into the data management part (ie., ICT sub-systems) of the research infrastructure. Within the ENVRI-RM, the acquisition phase is considered to begin upon point of data entry into the RI systems. The acquisition phase as modeled in the ENVRI RM starts from the design of the experiment. Acquisition is typically distributed across networks of observatories and stations. The 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 ready for curation. Control software is often deployed to manage and schedule the execution and monitoring of data flows. Data collected during the acquisition phase ultimately enters the data curation phase for preservation, usually within a specific time period.

## Data Curation

*In the **data curation phase** the research infrastructure stores, manages and ensures access to all persistent data-sets produced within the infrastructure.*

The data curation phase facilitates quality control and preservation of scientific data. The data curation functionalities are typically



implemented across one or more dedicated data centres. Data handled at this phase 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, supported by provenance data. 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 independent data access mechanisms. There is usually an emphasis on non-functional requirements for data curation satisfying availability, reliability, utility, throughput, responsiveness, security and scalability criteria.

## Data Publishing

*In the **data publishing phase** the research infrastructure enables discovery and retrieval of scientific data to internal and external parties.*

The data publishing phase enables discovery and retrieval of data housed in data resources managed as part of data curation. Data publishing often provide mechanisms for presenting or delivering data products. Query and search tools allow users or upstream services to discover data based on metadata or semantic linkages. Data handled during publishing need not be homogeneous. When supporting heterogeneous data, different types of data (often pulled from a variety of distributed data resources) can be converted into uniform representations with uniform semantics resolved by a data discovery service. Services for harvesting, compressing and packaging data and metadata, as well as encoding services for secure transfer can be provided. Data publishing is controlled using rights management, authentication, and authorisation policies.

## Data Processing

*In the **data processing phase** the research infrastructure provides a toolbox of services for performing a variety of data processing tasks. The scope of data processing is very wide.*

The data processing phase enables the aggregation of data from various sources, as well as conduct of experiments and analyses upon that data. During this phase data tends to be manipulated, leading to both/either derived and/or recombined data. To support data processing, a research infrastructure is likely to offer service operations for statistical analysis and data mining, as well as facilities for carrying out scientific experiments, modelling and simulation, and visualisation. Performance requirements for processing scientific data during this phase tend to be concerned with scalability, which can be addressed at the level of engineering and technical solutions to be considered (e.g., by making use of Cloud computing services). The data products generated during processing may themselves be curated and preserved within the RI.

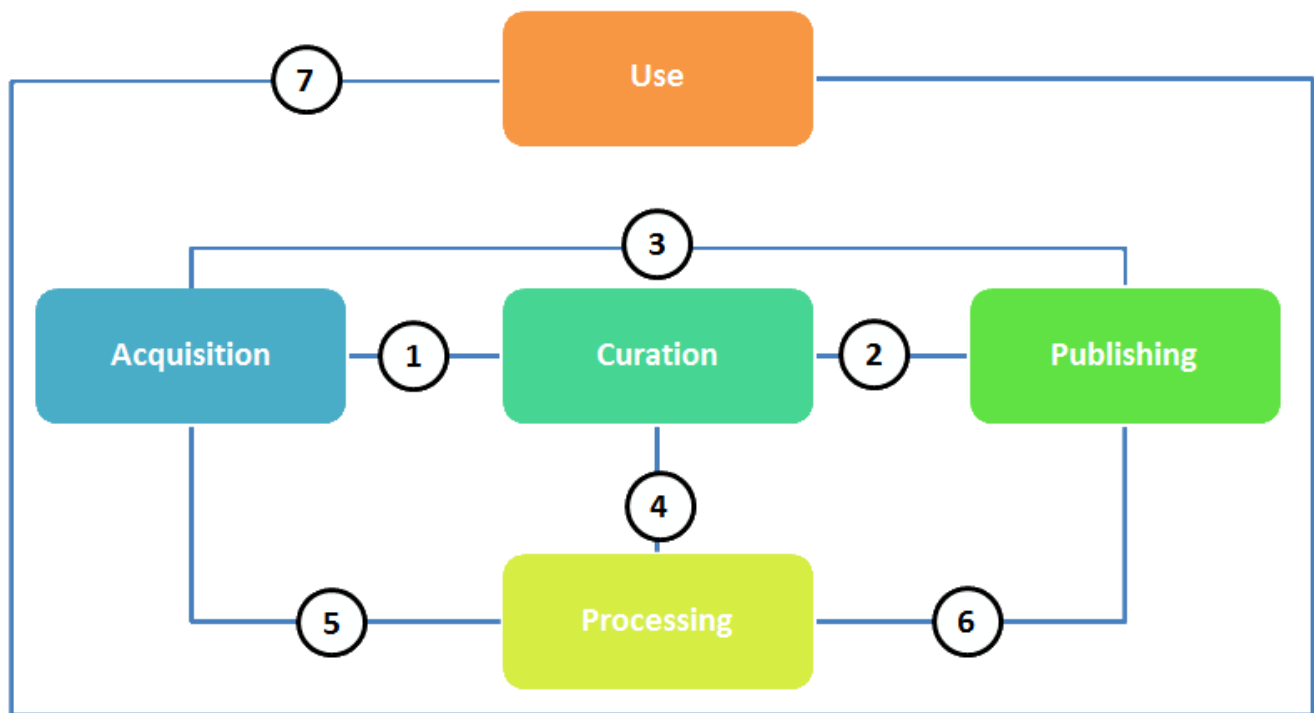
## Data Use

*In the **data use phase** the research infrastructure supports users of an infrastructure in gaining access to data and facilitating the preservation of derived data products.*

The data use phase provides functionalities that manage and track users' activities while supporting the users to conduct their research activities which may result in the creation of new data products. Data 'handled' and produced at this phase are typically user-generated data and communications. The data use phase requires supporting activities such as interactive visualisation, standardised authentication, authorisation and accounting protocols, and the use of virtual organisations. This is the most advanced form of data processing, at this phase the research infrastructure implements an interface with the wider world in which it exists.

## Lifecycle Support Inter- and Intra- Research Infrastructure Relationships

Each research infrastructure supports the data lifecycle to a different degree. According to the scope of a particular research infrastructure, some core activities align strongly with some of the phases while other phases are not so comprehensively supported. In this case, the integration of the research infrastructures and their external supporting systems and services help in the overall fulfilment of the research data lifecycle. For these cases, the major integration points are those at the transition between phases of the data lifecycle. These integration points are important to build the internal subsystems of the research infrastructure, as well as to integrate the research infrastructure with other research infrastructures.



**Illustration of the major integration (reference) points between different phases of the data lifecycle.**

The integration points described as follows refer to the components supporting a phase of the data lifecycle. However, the components being integrated can be within the same research infrastructure or in different research infrastructures.

1. **Acquisition/Curation** by which components specialized in data acquisition are integrated with components which manage data curation.
2. **Curation/Publishing** by which components specialized in data curation are integrated with components which support data publishing.
3. **Acquisition/Publishing** by which components specialized in data acquisition are integrated components which support data publishing.
4. **Curation/Processing** by which components specialized in data curation are integrated with components which support data processing.
5. **Acquisition/Processing** by which components specialized in data acquisition are integrated with components which support data processing.
6. **Processing/Publishing** by which the components specialized in data processing are integrated with components which support data publishing.
7. **Use/All** by which entities outside the research infrastructure may be allowed to provide, access, or use data at different phases of the data lifecycle.

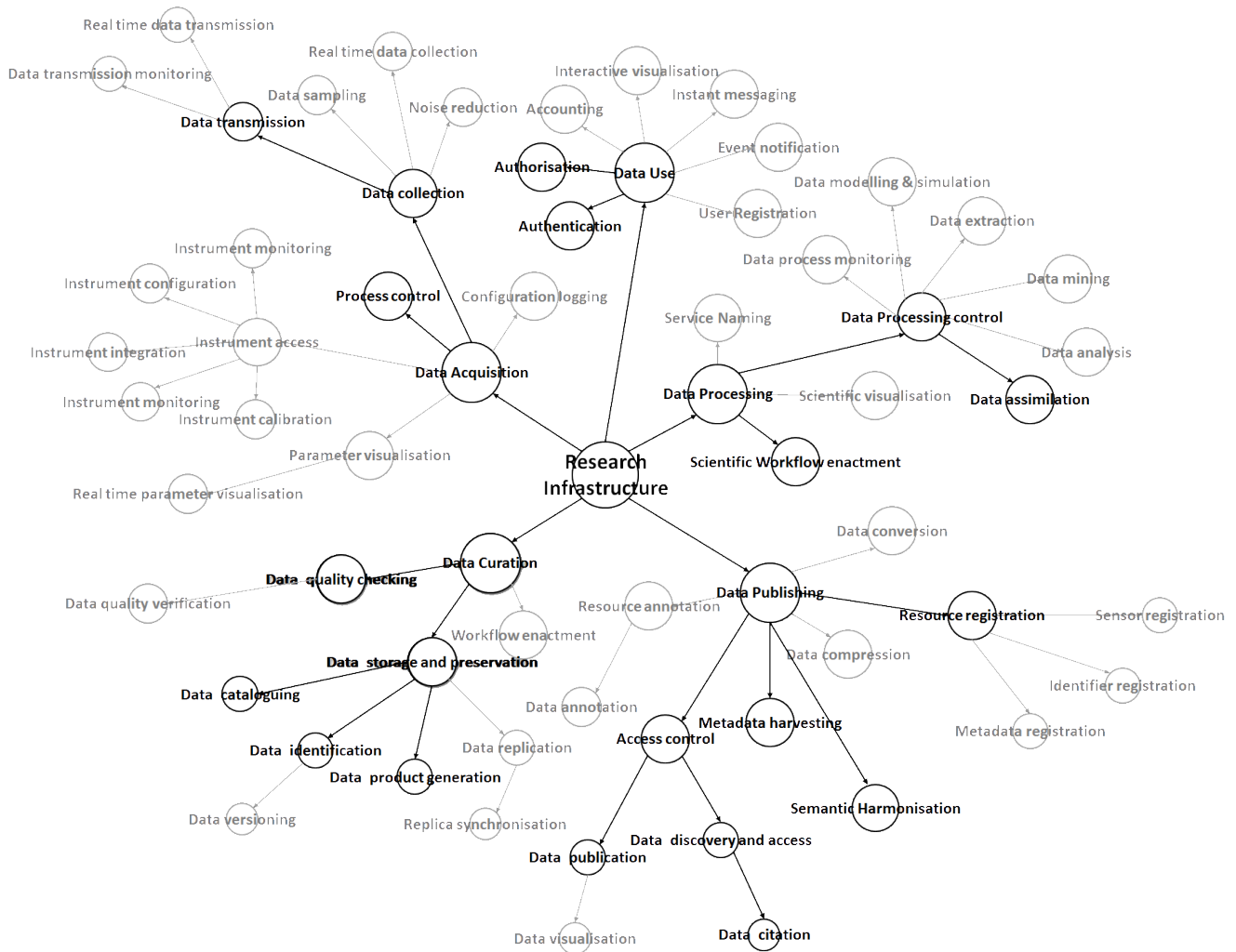
No notion of direction is implied in the definition of these points of reference. Relations with direction only appear when interfaces are superimposed on reference points, and then they can be unidirectional in either or both directions, or bidirectional - according to the nature of the interface(s).

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

## Common Functions within a Common Lifecycle

Analysis of requirements of environmental research infrastructures during the ENVRI and ENVRIplus projects has resulted in the identification of a set of common functionalities. These functionalities can be classified according to the five phases of the data lifecycle. The requirements 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).

In order to better manage the range of requirements, and in order to ensure rapid verification of compliance with the ENVRI-RM, a *minimal model* has been identified which describes the fundamental functionality necessary to describe an environmental research infrastructure. The minimal model is a practical tool to produce a partial specification of a research infrastructure which nonetheless reflects the final shape of the complete infrastructure without the need for significant refactoring. Further refinement of the models using the ENVRI-RM allow producing more refined models of designated priority areas, according to the purpose for which the models are created.



**Radial depiction of ENVRI-RM requirements with the minimal model highlighted.**

The definitions of the minimal set of functions are given as follows (a full list of common functions is provided in [Appendix A](#)):

#### **(A) Data Acquisition**

**Process Control:** 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:** Functionality that obtains digital values from a sensor instrument, associating consistent timestamps and necessary metadata.

**Data Transmission:** Functionality that transfers data over a communication channel using specified network protocols.

#### **(B) Data Curation**

**Data Quality Checking:** Functionality that detects and corrects (or removes) corrupt, inconsistent or inaccurate records from datasets.

**Data Identification:** Functionality that assigns (global) permanent unique identifiers to data products.

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

**Data Product Generation:** Functionality that processes data against requirement specifications and standardised formats and descriptions.

**Data Storage & Preservation:** 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 Publishing**

**Access Control:** Functionality that approves or disapproves of access requests based on specified access policies.

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

**Resource Registration:** 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:** 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:** Functionality that assigns an accurate, consistent and standardised reference to a data object, which can be cited in scientific publications and/or from other data collections.

**Semantic Harmonisation:** Functionality 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:** Functionality that retrieves requested data from a data resource by using suitable search technology.

#### **(D). Data Processing**

**Data Assimilation:** 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:** Functionality that inspects, cleans, and transforms data, providing data models which highlight useful information, suggest conclusions, and support decision making.

**Data Mining:** Functionality that supports the discovery of patterns in large datasets.

**Data Extraction:** 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:** Functionality that supports the generation of abstract, conceptual, graphical or mathematical models, and to run an instances of those models.

**(Scientific) Workflow Enactment:** Functionality provided as a specialisation of Workflow Enactment supporting the composition and execution of computational or data manipulation steps in a scientific application. Important processing results should be recorded for provenance purposes.

**Data Processing Control:** Functionality that initiates calculations and manages the outputs to be returned to the client.

#### **(E) Data use**

**Authentication:** Functionality that verifies the credentials of a user.

**Authorisation:** Functionality that specifies access rights to resources.

## The ENVRI Reference Model

The ENVRI Reference Model (ENVRI RM) defines an archetypical environmental research infrastructure. The ENVRI RM is structured according to the Open Distributed Processing (ODP) standard, ISO/IEC 10746-n, and as such, is defined from five different perspectives.

The Science, Information and Computational viewpoints take particular priority. These viewpoints allow expression of the complex concerns of the research infrastructures at a high level of abstraction. When building a research infrastructure, these viewpoints are important during the design and conceptualisation phases. These viewpoints have been defined previously by the ENVRI project, and enhanced by the ENVRIplus project.

The Engineering and Technology viewpoints complement the high level abstractions of the other three viewpoints by describing elements for physically building research infrastructures. When building a research infrastructure, these viewpoints are more relevant in the implementation and operational phases. These viewpoints are being defined as part of the ENVRIplus project.

- **Science Viewpoint**
- **Information Viewpoint**
- **Computational Viewpoint**
- Engineering Viewpoint
- Technology Viewpoint

## Science Viewpoint

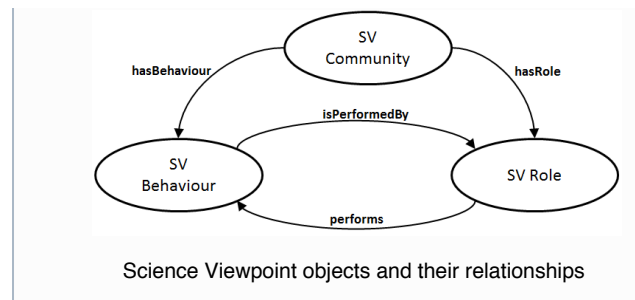
The Science Viewpoint (SV) of the ENVRI RM captures the requirements for an environmental research infrastructure from the perspective of the people who perform their tasks and achieve their goals as mediated by the infrastructure. Modelling in this viewpoint derives the principles and properties of model objects through the analysis of the structure and functionality of organisations, people interacting within and around those organisations, and rules governing the interactions.

Two requirements engineering efforts in the ENVRI and ENVRIplus projects revealed the existence of a common lifecycle for the data produced, shared, and processed by research infrastructures. The five phases of the data lifecycle are *Data Acquisition*, *Data Curation*, *Data Publishing*, *Data Processing* and

The Science Viewpoint defines communities with their community behaviours and roles. The diagram below shows the main elements of the science viewpoint and their relationships. Each ellipse contains a concept. The arrows connecting the concepts are directed and indicate the relationship between to concepts. The label of the link indicates the type of relationship. From this, the diagram indicates that SV behaviours are performed by SV roles. This is represented by two relationships, **isPerformedBy** and **performs**.

*Data Use*. Correspondingly, activities that support these five phases 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 called *communities* in ODP. The Science Viewpoint examines what those communities are, what kind of roles they have, and what main behaviours they act out.

- **Communities**
- **Community Roles**
- **Community Behaviours**



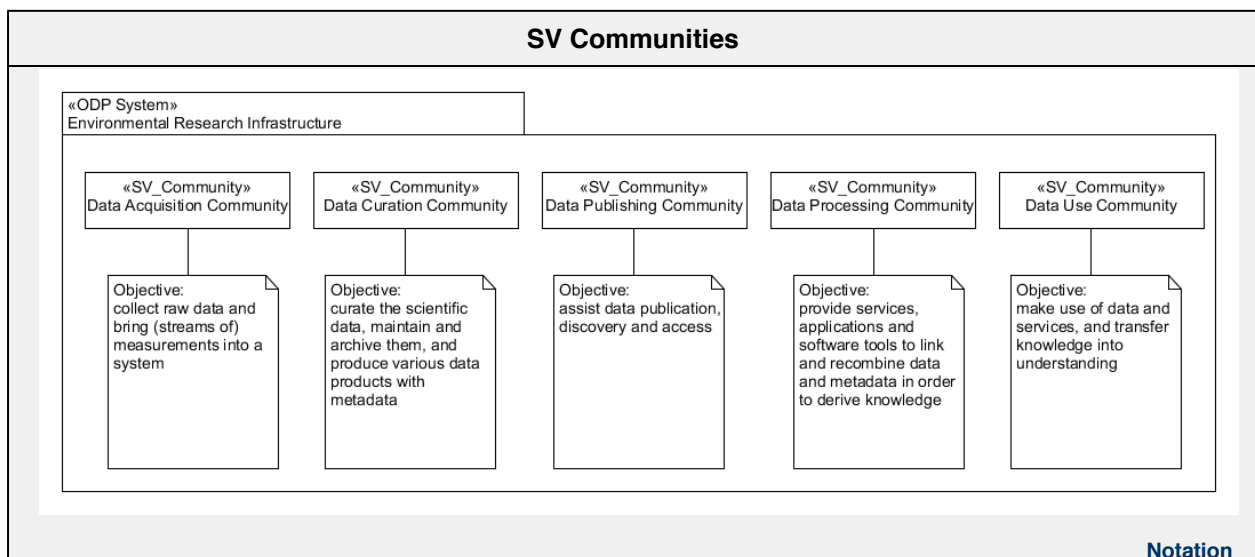
## SV Communities

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

The ENVRI RM distinguishes five groups of behaviours and roles, seen as communities which by design align with the five phases of the data lifecycle.

The five communities are, *data acquisition*, *data curation*, *data publication*, *data service provision*, and *data use*. The definitions of the communities are based on their objectives.

- **Data Acquisition Community**, collect raw data and bring (streams of) measurements into a system.
- **Data Curation Community**, curate the scientific data, maintain and archive them, and produce various data products with metadata.
- **Data Publishing Community**, assist data publication, discovery and access.
- **Data Processing Community**, provide various services, applications and software/tools to link and recombine data and metadata in order to derive knowledge.
- **Data Use Community**, make use of data and service products, and transfer knowledge into understanding.



The community roles and behaviours are described at the following pages:

- **Community Roles**
- **Community Behaviours**

## SV 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, e.g. software or hardware).

components).

**Active roles** are identified 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 by undertaking different roles.

**Passive roles** are identified with subsystems, subsystem components, and hardware facilities. Active roles interact with passive roles to achieve their objectives.

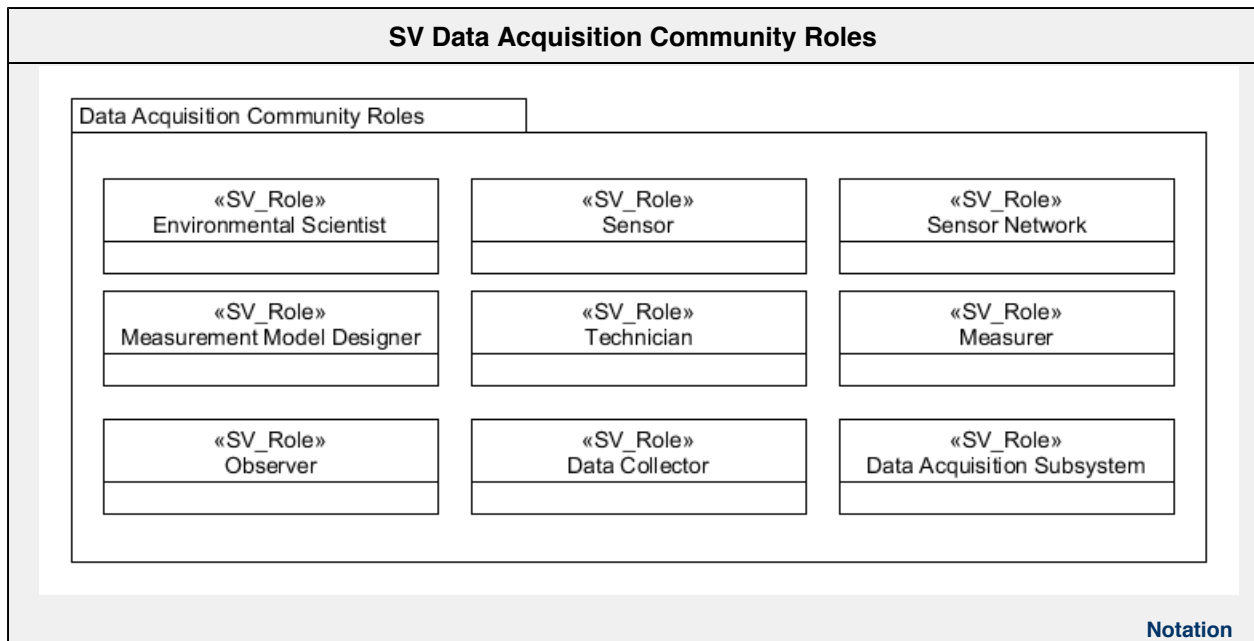
- [Roles in the Data Acquisition Community](#)
- [Roles in the Data Curation Community](#)
- [Roles in the Data Publishing Community](#)
- [Roles in the Data Processing Community](#)
- [Roles in the Data Use Community](#)

### Roles in the Data Acquisition Community

The main objectives of the data acquisition community is to bring measurements into the system. Consider a typical data acquisition scenario: A measurement and monitoring model is designed by *designers* 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, e.g., 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 subsystem to prepare the data or control the flow of data in order to automatically collect and transmit the data.

The following roles are identified in a data acquisition community:

- **Environmental Scientist:** An active role, which is a person who conducts research or performs scientific investigations. Using knowledge of various scientific disciplines, they may collect, process, analyse, synthesize, study, report, and/or recommend action based on data derived from measurements or observations of (for example) air, rock, soil, water, nature, and other sources.
- **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.
- **Measurement Model Designer:** An active role, which is a person who designs the measurements and monitoring models based on the requirements of environmental scientists.
- **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 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 the Science Viewpoint, the data acquisition subsystem is passive role of the data acquisition community. It is the part of the research infrastructure providing functionalities to automate the process of data acquisition.



The behaviours of the data acquisition community is described at [Community Behaviours](#).

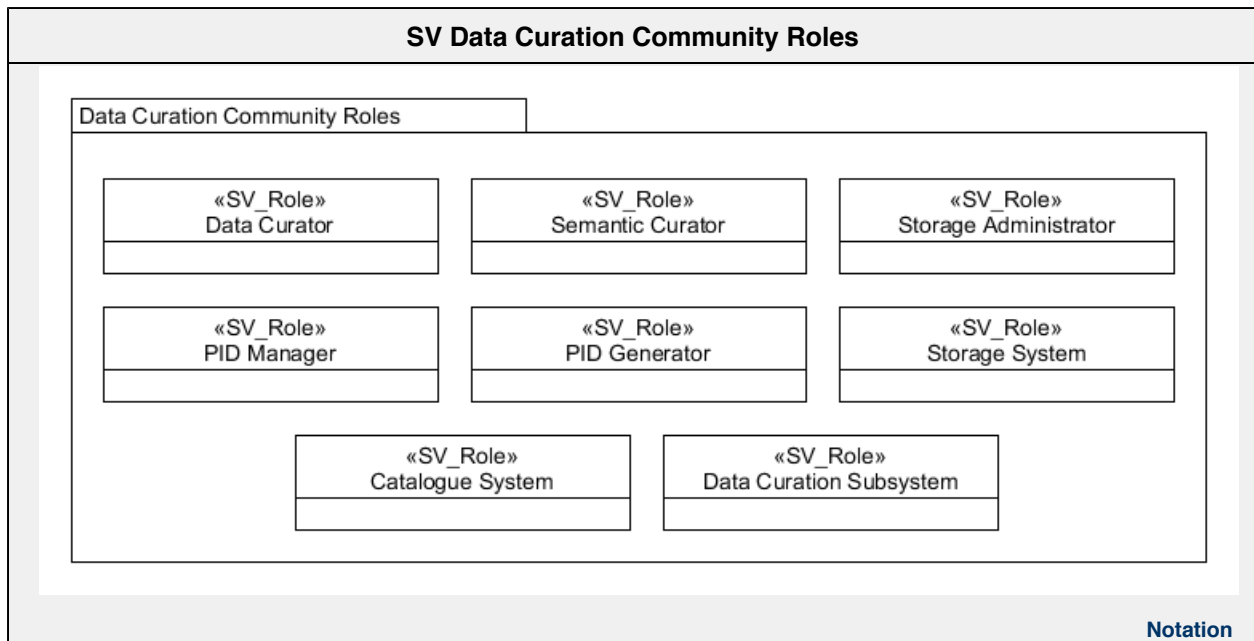
### Roles in the Data Curation Community

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. Upon registering 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. Here, we use a more general term "PID" instead of "*handle*" (thus, "PID manager" instead of "*handle servers*"), and identify the key roles involved in the data curation process.

We identified the following roles in this community:

- **Data Curator:** An active role, which is a person who verifies the quality of the data; annotates the data; catalogues, preserves and maintains the data as a resource; and prepares various required data products.
- **Semantic Curator:** An active role, which is a person who designs and maintains local and global conceptual models and uses those models to annotate the data and metadata.
- **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.
- **PID Manager:** A passive role, a system or service that assigns persistent global unique identifiers to data and metadata products. The Manager invokes a external entity, the PID Service, to obtain the PIDs. The manager maintains a local catalogue of PIDs that are being used to reference data and metadata. If the data or metadata in the RI change location or are removed, the PID manager updates this information locally and informs the PID Service.
- **PID Generator:** A passive role, a public system or service which generates and assigns persistent global unique identifiers (PIDs) to sets of digital objects. The PID Generator also maintains a public registry of PIDs for digital objects.
- **Storage System:** A passive role, which includes memory, components, devices and media that retain data and metadata for an interval of time.
- **Catalogue System:** A passive role, a catalogue system is a special type of storage system designed to support building logical structures for classifying data and metadata.
- **Data Curation Subsystem:** the data curation subsystem is a passive role of the data curation community. It is the part of the research infrastructure which stores, manages and ensures access to all persistent data and metadata produced within the infrastructure.





The behaviours of the data curation community are described at [Community Behaviours](#).

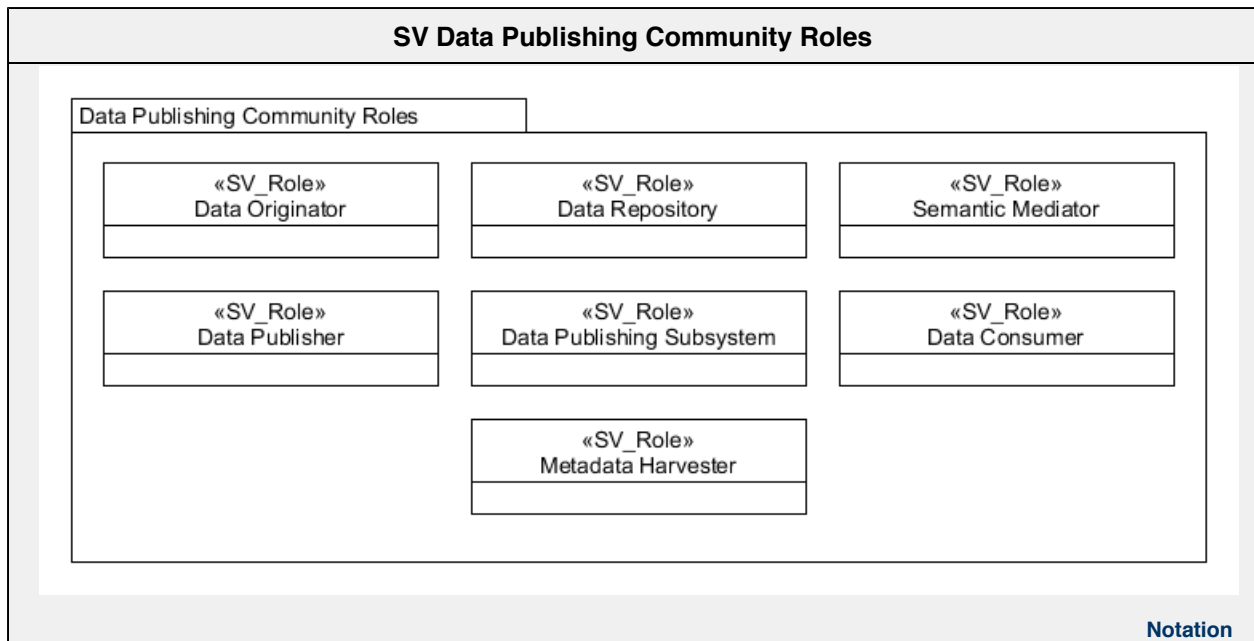
### Roles in the Data Publishing Community

The objectives of the data publishing 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 authorised 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.).

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, 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.
- **Data 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 (i.e., executing mapping and translation rules), discovery and integration of heterogeneous data.
- **Data Publisher:** An active role, is a person in charge of supervising the data publishing processes.
- **Data Publishing Subsystem:** In the Science Viewpoint, the data access subsystem represents a passive role of the data publication community. It is the part of the research infrastructure enabling the discovery and retrieval of scientific data. The access to this subsystem could require authorisation at different levels for different roles.
- **Data Consumer:** Either an active or a passive role, which is an entity who receives and uses the data.
- **Metadata Harvester:** A passive role, which is a system or service collecting metadata which supports the construction/selection of a global conceptual model and the production of mapping rules

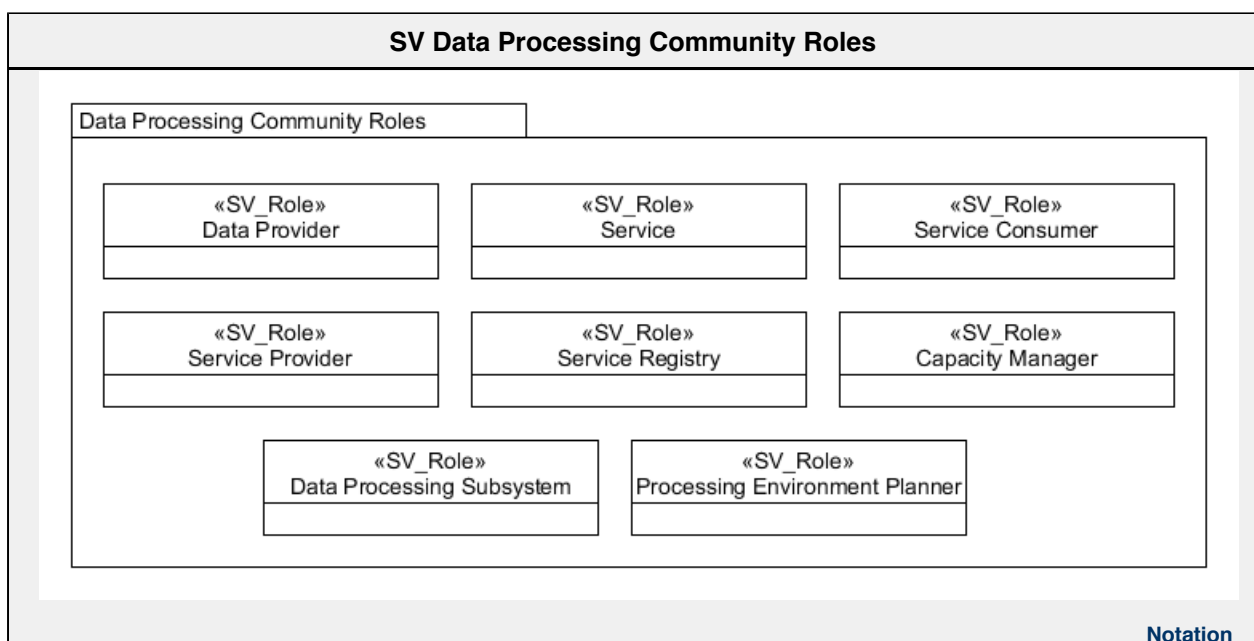


The behaviours of the data publishing community are described at [Community Behaviours..](#)

### Roles in the Data Processing Community

The data processing 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.

- **Data Provider:** Either an active or a passive role, which is an entity providing the data to be used.
- **Service:** A passive role, in which a functionality for processing data is made available for general use.
- **Service Consumer:** Either an active or a passive role, which is an entity using the services provided.
- **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.
- **Data Processing Subsystem:** In the Science Viewpoint, the data processing subsystem represents a passive role of the data processing community. It is the part of the research infrastructure providing services for data processing. These services could require authorisation at different levels for different roles.
- **Processing Environment Planner:** An active agent that plans how to optimally manage and execute a data processing activity using RI services and the underlying e-infrastructure resources (handling sub-activities such as data staging, data analysis/mining and result retrieval).

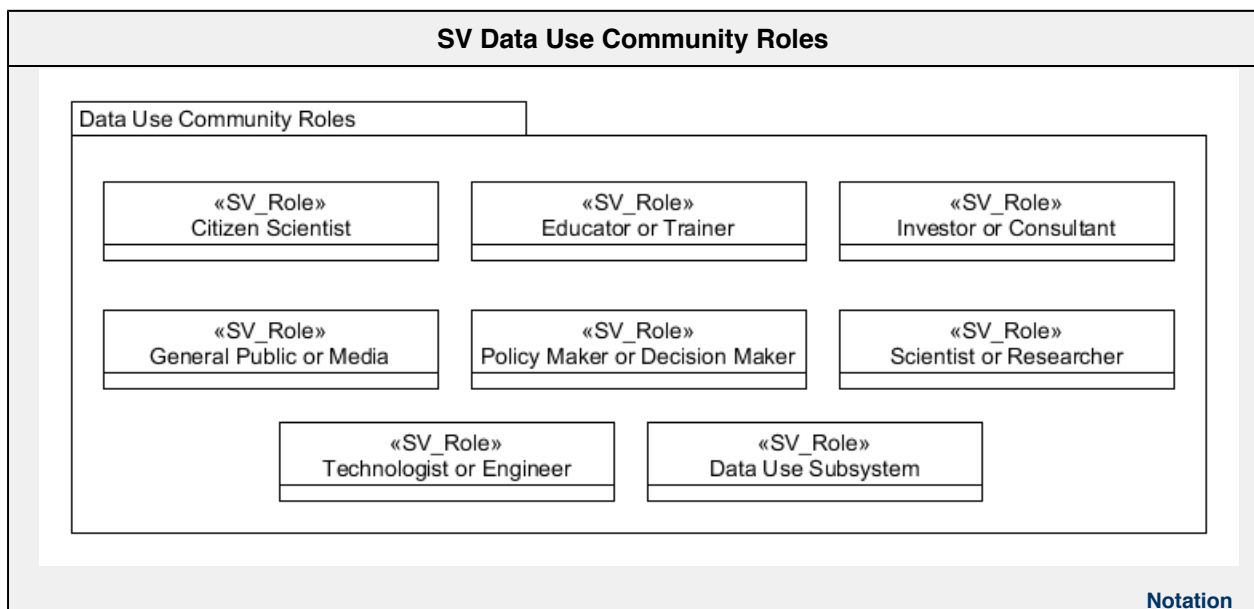


The behaviours of the data processing community are described at [Community Behaviours](#).

## Roles in the Data Use Community

The main role in the data use community is a *user* who is the ultimate consumer of data, applications and services. Depending on the purposes of use, 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.
- **Educator or Trainer:** An active role, which is a person who makes use of the data and application services for education and training purposes.
- **Policy Maker or Decision Maker:** An active role, which is a person who makes decisions based on the data evidence.
- **Investor or consultant (Private Sector):** 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:** An active role, which is a person or organisation interested in understanding the knowledge delivered by an environmental science research infrastructure, or discovering and exploring the [knowledge base](#) enabled by the research infrastructure.
- **Citizen Scientist:** An active role, member of the general public who engages in scientific work, often in collaboration with or under the direction of professional scientists and scientific institutions (also known as amateur scientist).
- **Data Use Subsystem:** In the Science Viewpoint, the data use subsystem represents a passive role of the data use community. It is the part of the research infrastructure supporting the access of users to an infrastructure. The data use subsystem manages, and tracks user activities and supports users to conduct their roles in different communities.



The behaviours of the data use community are described at [Community Behaviours](#).

## SV Community Behaviours

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

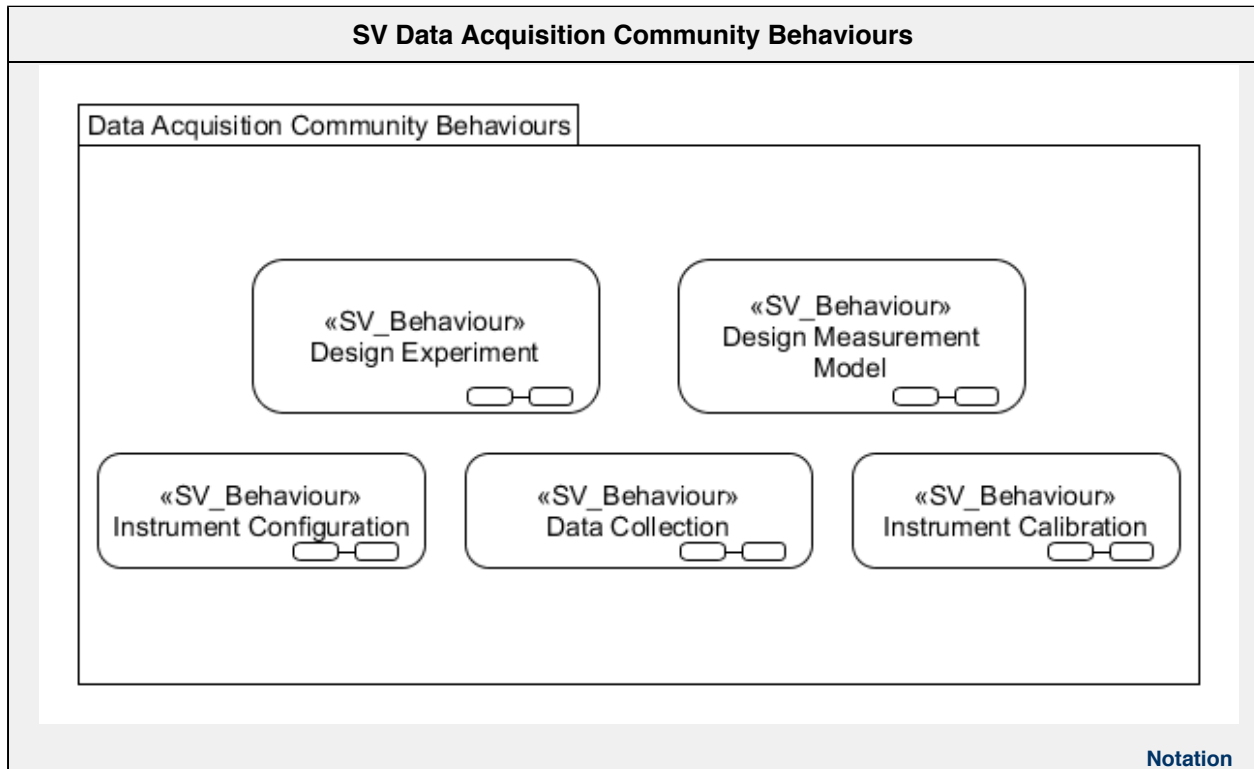
- [Behaviours of the Data Acquisition Community](#)
- [Behaviours of the Data Curation Community](#)
- [Behaviours of the Data Publishing Community](#)
- [Behaviours of the Data Processing Community](#)
- [Behaviours of the Data Use Community](#)

### Behaviours of the Data Acquisition Community

The key behaviours of the data acquisition community through the interaction of the community roles include:

- **Design Experiment:** A behaviour performed by a *Environmental Scientist* that designs the scientific experiment which motivates the data acquisition activities.
- **Design 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 *sens**r* against dependable standards or specified verification processes.
- **Data Collection:** A behaviour performed by a *Data Collector* that control and monitor the collection of the digital values from a *sens**r* instrument (or a human sensor such as a *Measurer* or a *Observer*), associating consistent timestamps and necessary metadata.



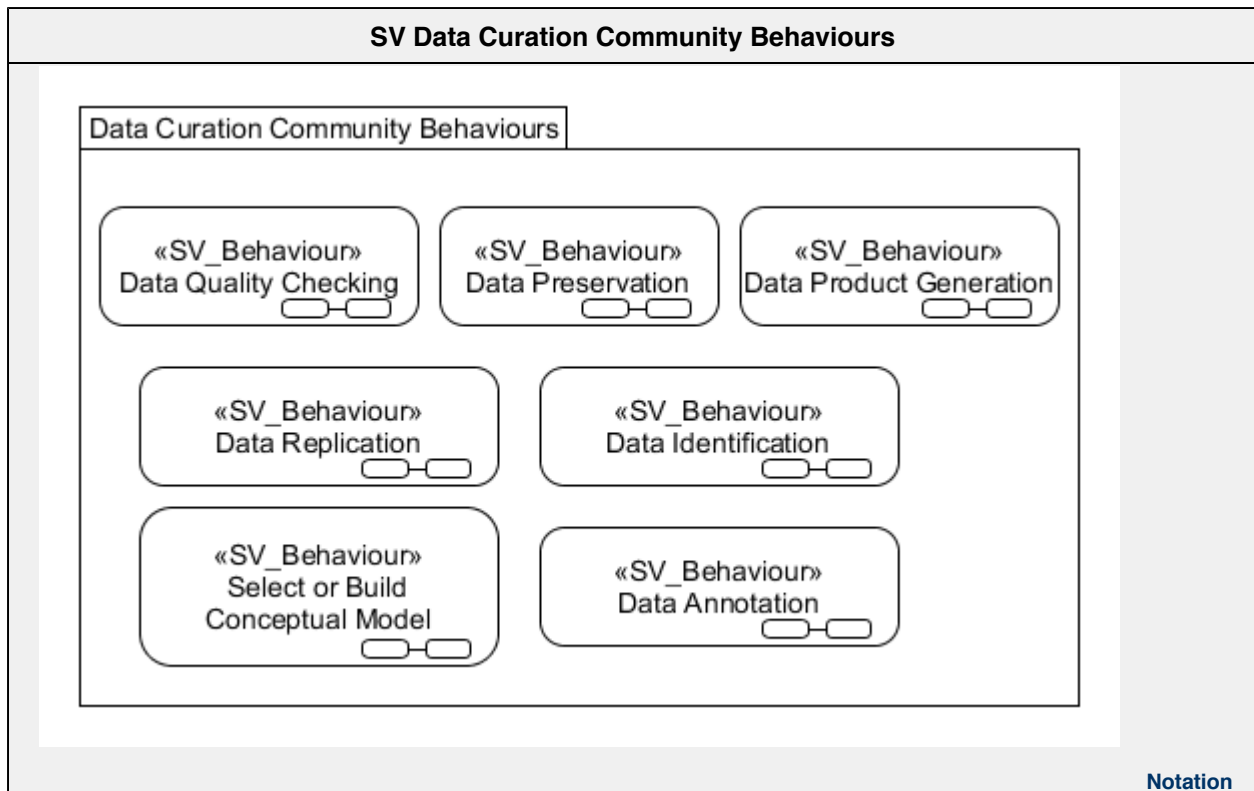
Notation

The roles of the data acquisition community are described in [Community Roles](#).

### Behaviours of the Data Curation Community

The main behaviours of the data curation community 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.
- **Data Identification:** A behaviour performed by a *PID manager* which provides a unique PID for data and metadata being curated.
- **Select or Build Local Conceptual Model:** A behaviour performed by a *Semantic Curator* which supports the annotation of data and metadata.
- **Data Annotation:** A behaviour performed by a *Semantic Curator* which supports the linking of data and metadata with a local conceptual model.

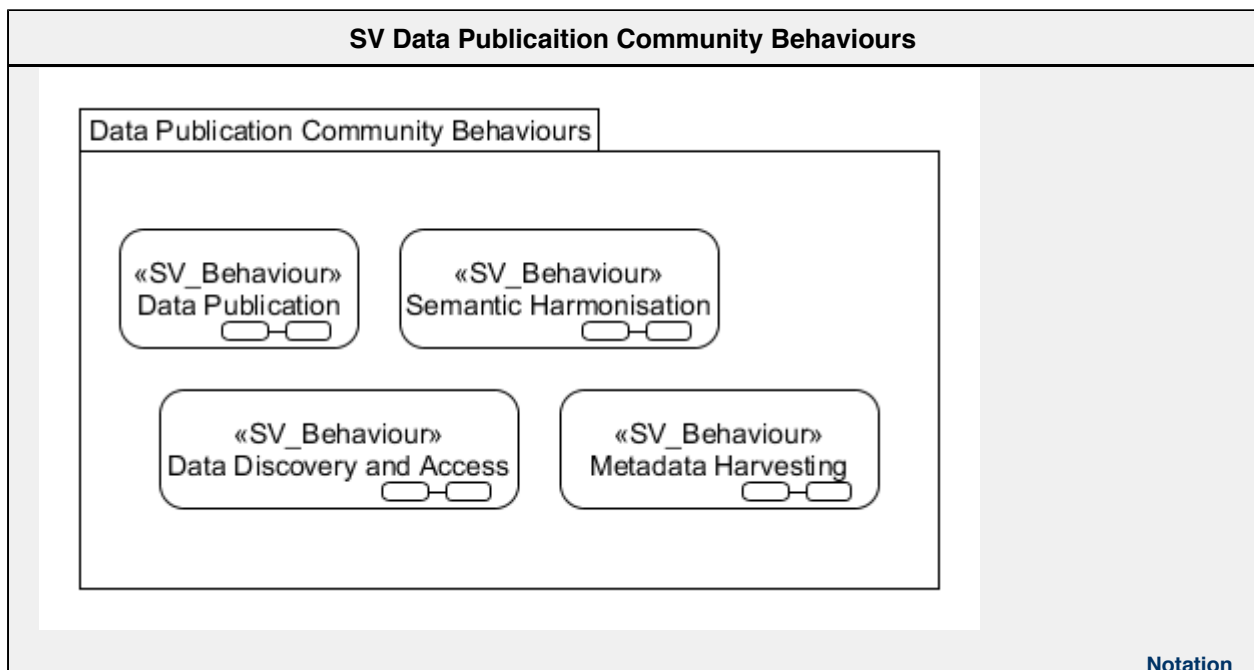


The roles of the data curation community which are described at [Community Roles](#).

#### Behaviours of the Data Publishing Community

The data publishing 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 *PID Manager* that assigns an accurate, consistent and standardised reference to a data object, in the same way as researchers routinely provide a bibliographic reference to printed resources. The RI publishing the data can define the citation contents such as authors, and dates for different citation styles.



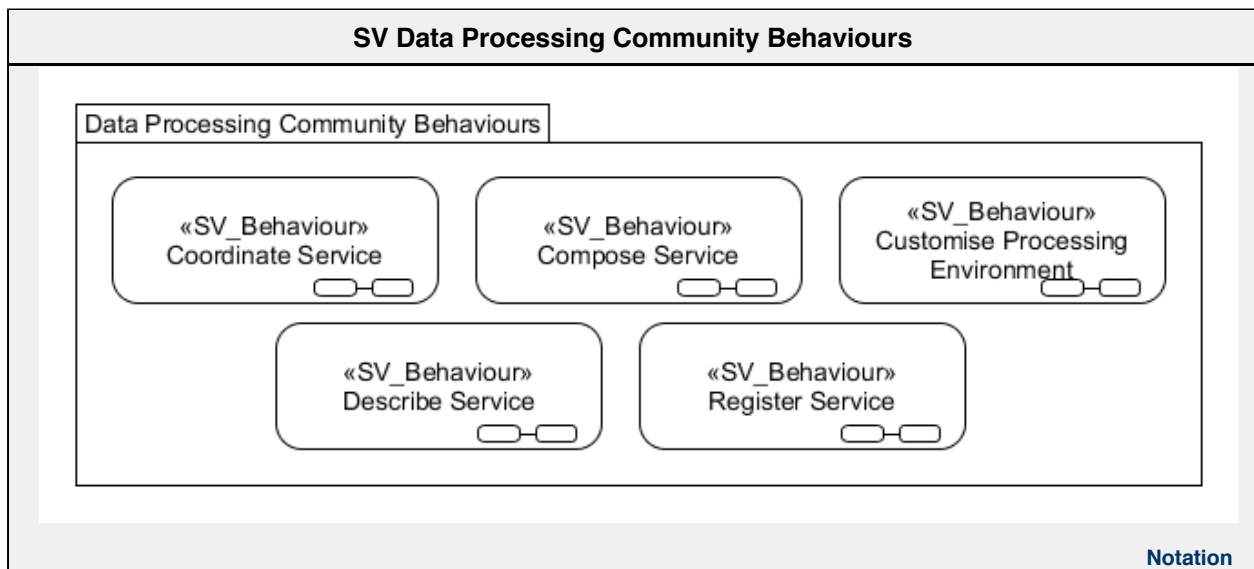
The roles of the data publication community are described at [Community Roles](#).

## Behaviours of the Data Processing Community

The following behaviours of the data processing community are modelled:

- **Coordinate Service:** 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.
- **Compose Service:** 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].
- **Customise Processing Environment:** A behaviour performed by a processing environment planner to enable a Data Processing Subsystem to prepare customised infrastructure and service platforms for managing specific data processing applications optimally, including the planning, provisioning and deployment sub-activities.
- **Describe Service:** A behaviour performed by a *Service Provider* to provide the information needed in order to use a service [8].
- **Register Service:** A behaviour performed by a *Service Provider* to make the service visible to *Service Consumers* by registering it in a service registry [8].

These are general behaviours of a service-oriented computing model. In the context of environmental science research infrastructures, a data processing 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 [Terminology and Glossary](#) for the definitions of these functionalities).



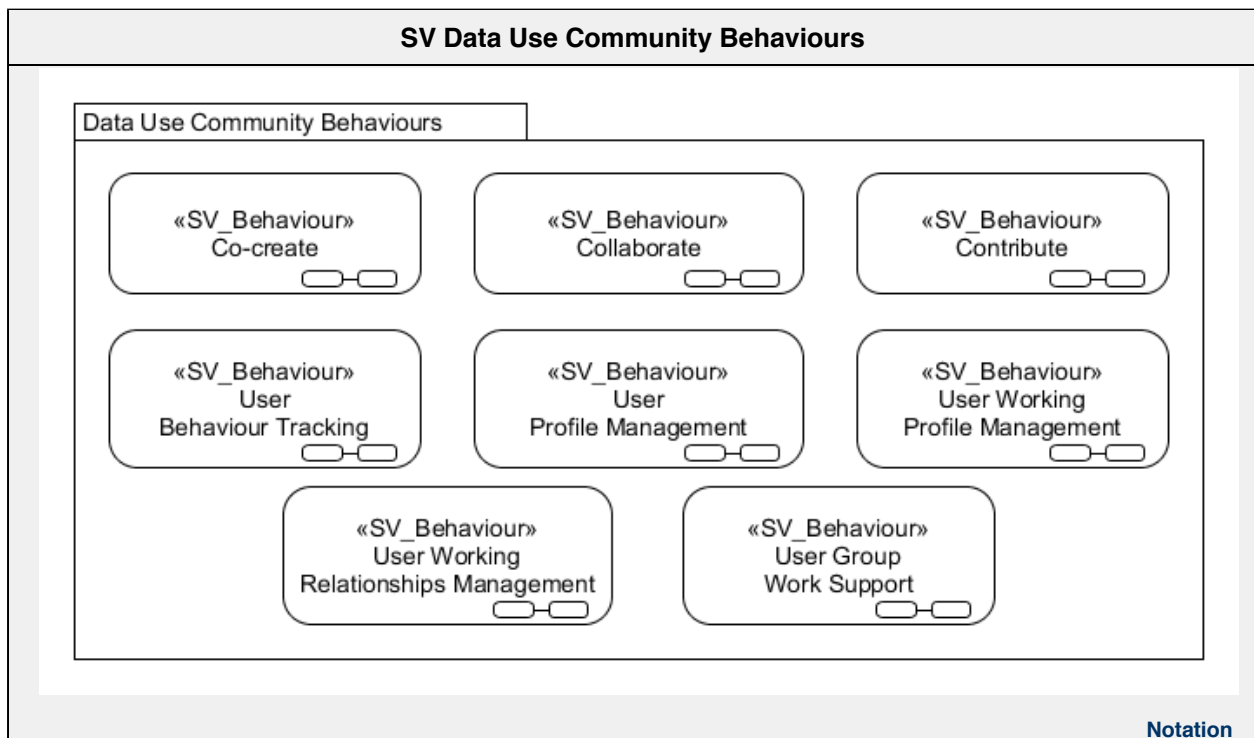
The roles of the data processing community are described at [Community Roles](#).

## Behaviours of the Data Use Community

The data use community can be divided in two main groups: (1) the behaviours performed by active roles (human actors) and (2) the behaviours performed by passive roles (computer resources). The first group encompasses the activities performed by human actors using the RI to interact with the different components of the RI. This can extend to all the actors in all the communities defined in the SV, in addition to the ones in the use community, for these reason these can also be called community support behaviours (or user support). The second group corresponds to the behaviours that enable the authorisation, authentication, and accounting of the activities of users, also known as AAI behaviours.

- **Co-create:** A behaviour performed by active roles which entails the design and planning of activities for the collection, preservation, analysis or publishing of research data in partnership with different communities.
- **Collaborate:** A behaviour performed by active roles which entails assisting/participating in some of the phases of the collection, preservation, analysis or publishing of research data.
- **Contribute:** A behaviour performed by active roles which entails directly collecting, preserving, analysing, or publishing research data held by the RI, according to a predefined protocol.
- **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.
- **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.



Notation

The roles of the data use community are described at [Community Roles](#).

## Information Viewpoint

The goal of the Information Viewpoint (IV) is to provide a common abstract model for the shared research data handled by the infrastructure. The focus lies on the data itself, without considering any platform-specific or implementation details. It is independent from the computational interfaces and functions that manipulate the data or the nature of technology used to store it. Similar to a high level ontology, the IV aims to provide a unique and consistent interpretation of the shared information objects of a particular domain.

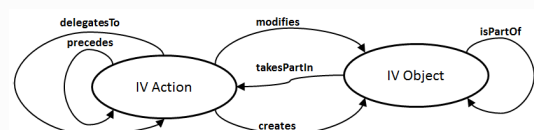
The IV specifies the types of the information objects and the relationships between those types. The main purpose of this viewpoint is to provide an abstract model of the lifecycles of the information objects handled by the RI. It also defines the constraints on information objects and the rules governing those lifecycles.

The models of the IV are grouped as follows.

- **Components:** collections of information objects and action types necessary to support the **minimal set of required functionalities**.
- **Information Objects Lifecycle:** descriptions of how information objects change as the infrastructure operates, illustrated using allowed state changes as the effects of the actions.
- **Information Management Constraints:** models of constraints that actions on information objects should implement to ensure the integrity and preservation of information objects.

The Information Viewpoint defines a set of IV objects and the set IV actions acting on those objects.

The diagram below shows the main elements of the IV and their relationships. Each ellipse contains a concept. The arrows connecting the concepts are directed and indicate the relationship between concepts. The label of the link indicates the type of relationship. From this, the diagram indicates that an IV object can be created by an IV action, as indicated by the **creates** relationship. Similarly, an IV object can be part of another IV object, as indicated by the **isPartOf** relationship. In this same way an action can be part of a chain of actions, this is indicated by two relationships **delegatesTo** and **precedes**.



Information Viewpoint components and their relationships

In the ENVRI RM research data and metadata are the main information objects managed by an RI. For this reason the IV is closely aligned with the **research data lifecycle model**.

## IV Components

The ENVRI RM information viewpoint defines a configuration of information objects, the behaviour of those objects, the actions that operate on those objects, and a set of constraints that should always hold for actions applied on objects. The presentation of IV components are organised as follows:

- **IV Information Objects:** definition of a collection of information objects manipulated by the system.



- **IV Information Object Instances**: definition of valid instances of information objects.
- **States**: detailed description of information object states and state transitions resulting from actions.
- **IV Action Types**: definition of events that cause state changes of information objects.

## IV Information Objects

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

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

- Data: research data processed by the RI, characterised as persisted data
  - Scientific data
  - Unique identifiers for the data identification
  - Backup (of data)
- Metadata: data typically related to the design of observation and measurement models, complements data by providing more precise details.
  - Design specification of the observation and measurement
  - Description of the measurement procedure
  - Quality Assurance (QA) annotations
  - Concepts from a conceptual model, e.g. an ontology
  - Mapping rules which are used for the model-to-model transformations
  - Provenance records
  - Management metadata(The data used to identify the states of data and metadata objects)
    - Data states
    - Metadata states
- Information Object Definitions
  - specification of investigation design
  - specification of measurements or observations
  - measurement result
  - concept
  - conceptual model
  - QA notation
  - metadata
  - metadata state
  - metadata catalogue
  - citation
  - persistent data
  - data state
  - unique identifier (UID)
  - backup
  - mapping rule
  - data provenance
  - service description
  - institution
  - person
  - project
  - community
  - role



**Note**

This specification can be included as metadata or as **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, qualitative, or cataloguing determinations of magnitude, dimension, and uncertainty to the outputs of observation instruments, sensors, sensor networks, human observers and observer networks.

**concept**

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

**conceptual model**

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

**QA notation**

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

In practice, this can be:

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

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

**metadata**

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

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

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

- HTML (Hyper-Text Markup Language): [www.w3.org/MarkUp/](http://www.w3.org/MarkUp/)
- XML (eXtensible Markup Language): [www.w3.org/XML/](http://www.w3.org/XML/)
- RDF (Resource Description Framework): [www.w3.org/RDF/](http://www.w3.org/RDF/)
- OWL (Web Ontology Language): [www.w3.org/2001/sw/](http://www.w3.org/2001/sw/)
- SGML (Standard Generalised Markup Language): [www.w3.org/MarkUp/SGML/](http://www.w3.org/MarkUp/SGML/)
- MARC (Machine Readable Cataloging): [www.loc.gov/marc/](http://www.loc.gov/marc/)
- MIME (Multipurpose Internet Mail Extensions): [www.ukoln.ac.uk/metadata/resources/mime/](http://www.ukoln.ac.uk/metadata/resources/mime/)
- DIME (Direct Internet Message Encapsulation): [xml.coverpages.org/draft-nielsen-dime-01.txt](http://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/](http://purl.oclc.org/metadata/dublin_core/)
- CERIF (Common European Research Information Format): [www.eurocris.org](http://www.eurocris.org)
- ISO 11179: [metadata-stds.org/11179/](http://metadata-stds.org/11179/)
- ISO 19115 (by iso-tc 211): [www.isotc211.org](http://www.isotc211.org)
- FGDC (The Federal Geographic Data Committee): [www.fgdc.gov/standards](http://www.fgdc.gov/standards)
- INSPIRE: [inspire.jrc.ec.europa.eu](http://inspire.jrc.ec.europa.eu)
- ISO 19115, Geographic information - metadata standard (metadata model closely related to INSPIRE) [www.iso.org](http://www.iso.org)
- DDI (Data Documentation Initiative): [www.ddialliance.org](http://www.ddialliance.org)
- TEI (The Text Encoding Initiative): [www.tei-c.org](http://www.tei-c.org)

- METS (Metadata Encoding and Transmission Standard): [www.loc.gov/standards/mets](http://www.loc.gov/standards/mets)
- MODS (Metadata Object Description Schema): [www.loc.gov/standards/mods/](http://www.loc.gov/standards/mods/)
- OAI (Reference Model for an Open Archival Information System)

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

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

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

#### metadata 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. Metadata registered within public catalogues.

#### metadata catalogue

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

#### citation

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

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

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

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

#### persistent data

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

#### data state

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, at a given instant in time (as defined in ODP, ISO/IEC 10746-2).

The data states and their changes as effects of actions are illustrated in [Data States](#).

In their lifecycle, data may have certain states:

- |                     |   |
|---------------------|---|
| • raw               | the primary results of observations or measurements   |
| • identified        | data which has been assigned a unique identifier  |
| • 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)                                  |

#### Note

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

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

#### unique identifier (UID)

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

There are 3 main generation strategies:

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

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

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

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

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

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

#### **backup**

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

#### **mapping rule**

Configuration directives used for model-to-model transformation.

Mapping rules can be transformation rules for:

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

#### **data provenance**

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

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

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

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

#### **service description**

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

In the Context of environmental RIs, an institution is any organisation participating in the RI within any of the communities which conform that RI.

person

Human actor member of an institution which may undertake one or more roles within a community

project

In the Context of environmental RIs, the project is collaborative enterprise planned to facilitate the acquisition, curation, publishing, processing and use of research data.

community

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

role

A role is a collection of IV actions that can be performed any number of times concurrently or successively.

## **IV Information Object Instances**

Information object instances are used to define valid instances of information objects. As explained earlier an information object can have several state transitions. An information object instance is a

model of an information object at a particular state.

The diagram on the right shows examples of different object instances. The main difference with information object is that the status of the information object instances is assigned a value from the list of allowed states.

Information objects instances are needed for two purposes:

1. to show the data state changes as effects of actions;
2. to show the relations between valid states of related data types, for instance that reaching the "published" requires a specific series of previous states which can be traced for QA and provenance validation.

The diagram also includes two types of conceptual models: "local conceptual model" and "global conceptual model".

#### Global conceptual model

A set of concepts accepted by a data sharing community.

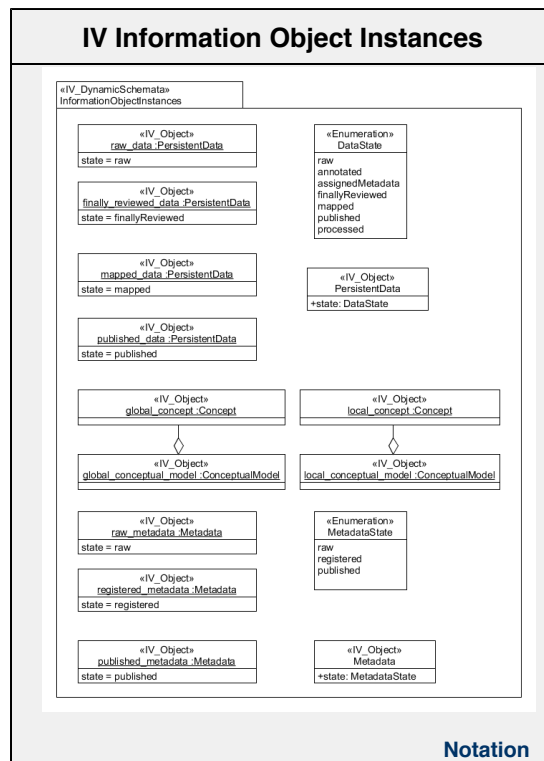
Global conceptual models contain global concepts. Examples of these types of models are global Thesauri like GEMET / EuroVoc / AGROVOC or global ontologies like Gene Ontology.

#### Local conceptual model

A set of concepts locally agreed by a limited user community, such as the members of a research institution.

A local conceptual model contains concepts which have a specific meaning according to the community using them for instance local definitions of person, institute, or data.

A conceptual model can be local or global depending on the size of the community which commits to it. Local conceptual models can contain concepts borrowed from global conceptual models. Alternatively mapping rules can be established to determine equivalences between global and local concepts.

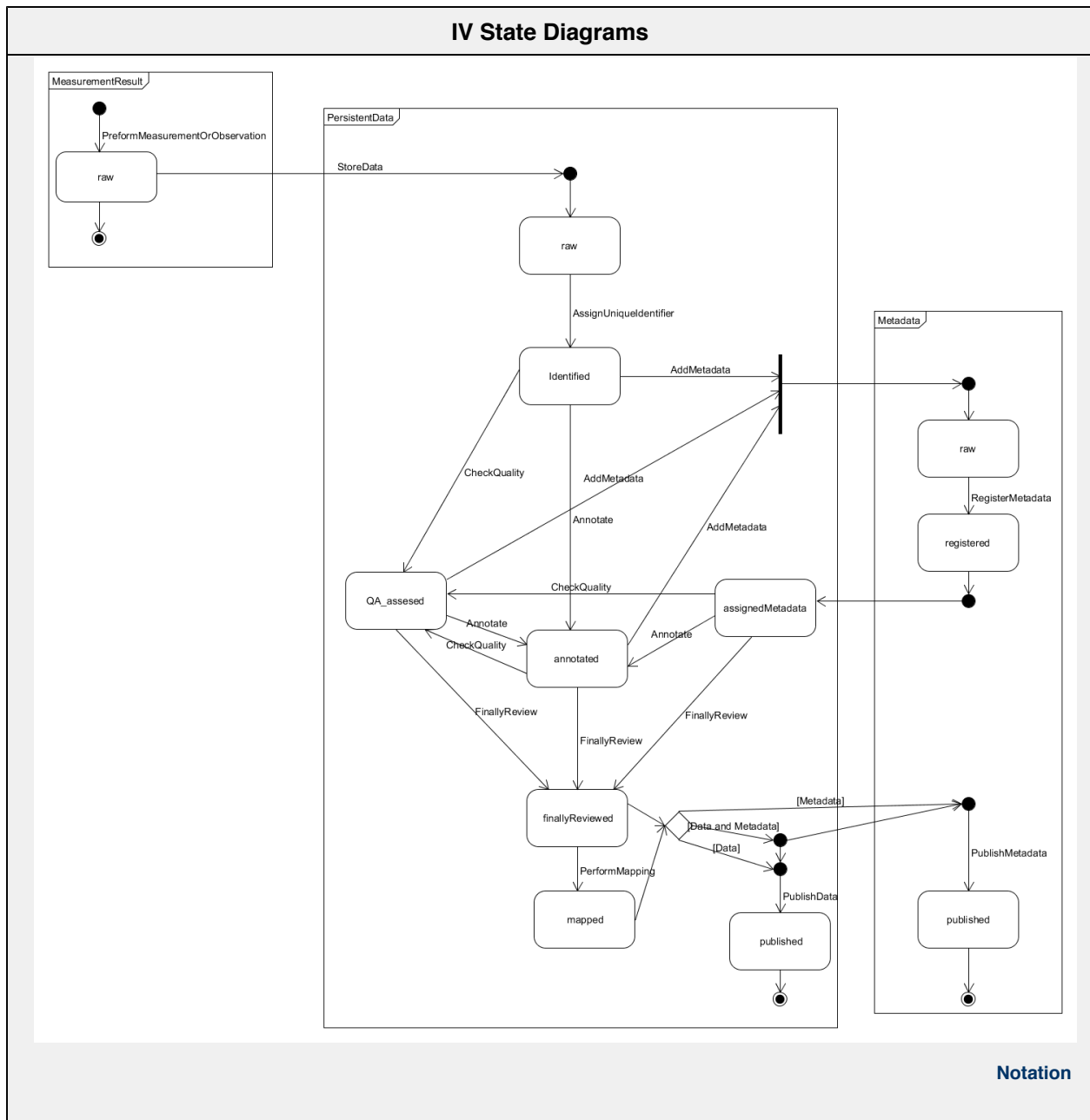


## IV States

The ENVRI RM IV defines **data states** and **metadata states** as the set of attributes which determine the actions that can be performed over a given information object.

The state changes, together with the **IV information actions** can be used to model the behaviour of data as it is managed by the RI.

The diagram below shows the states of three IV objects (Measurement Result, Persistent Data and Metadata) and their relationships.



This diagram shows all the possible states for each of the IV Objects.

The first IV object (left) is Measurement Result object. The object is created from a **PerformMeasurementOrObservation** action. The Measurement Result has only one state "raw".

The second IV object (middle) is PersistentData object. The **StoreData** action triggers the creation of the new Persistent Data object in the RI which has a raw state. This object can have six states. The transitions shown in the figure indicate possible transitions which can occur during the lifetime of the object.

The third IV object (right) is a "Metadata" object. The **AddMetadata** action triggers the creation of the new Metadata object in the RI. This object can have three states. However, the transitions from registered to published is not directly triggered. The diagram indicates that publish metadata must occur simultaneously with publish data

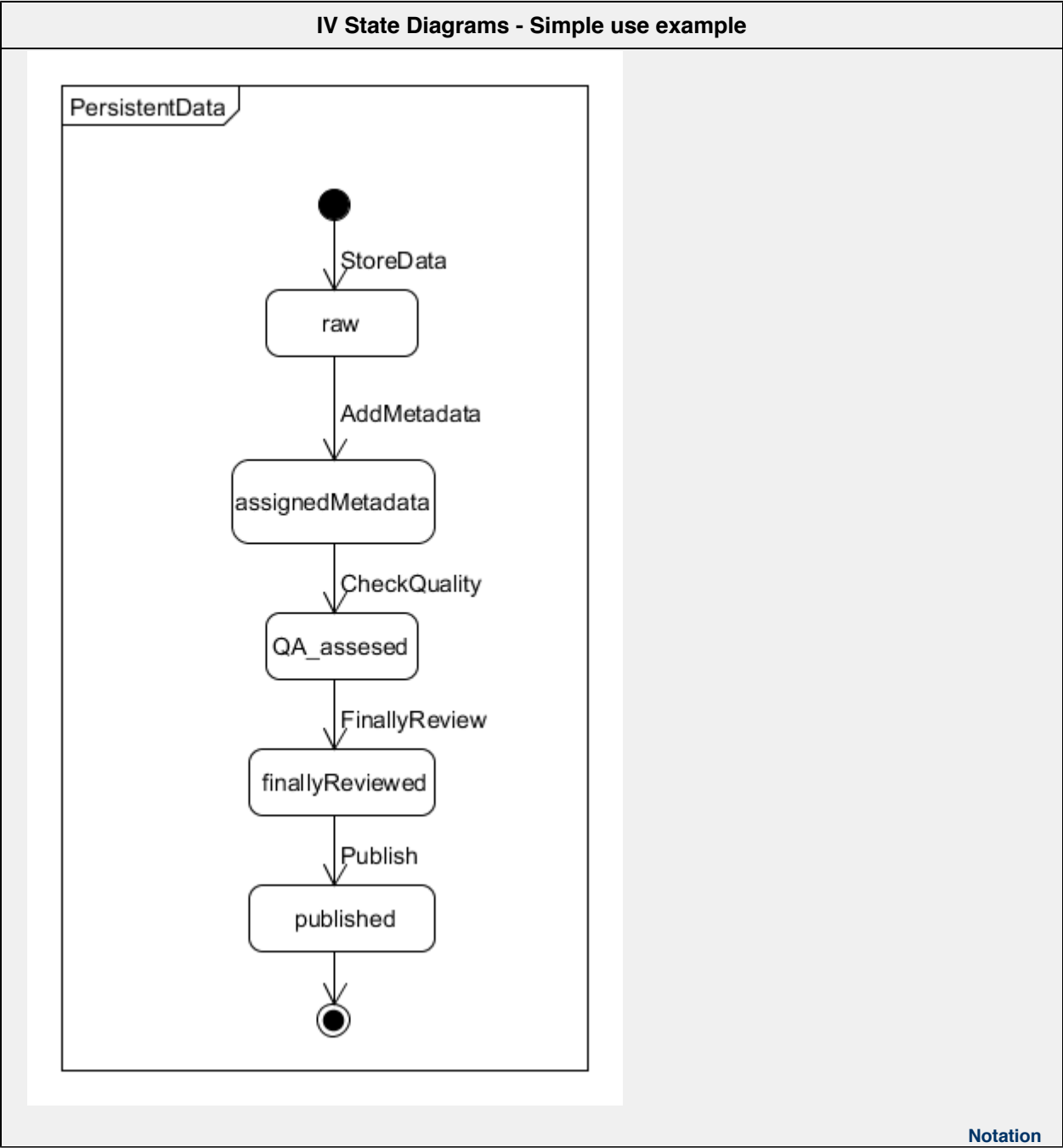
In the diagram the filled circle indicates a starting point or a junction. Used as starting point, a filled circle indicates a pseudo-state which represents the start of the lifetime of an object instance. Used as a junction, the filled circle indicates a pseudo-state where paths merge or split. The Rectangles with a label in the upper left corner are used to indicate the object whose states are represented. The rectangles with rounded corners are used to indicate states. Each arrow indicates a transition between states. The label of the arrow indicates the activity that triggers the transition. The bar figure in the diagram indicates a pseudo-state that can represent a fork or a merge of paths. The type of diagram presented is an UML state machine diagram [40].

#### Simple use example

The diagram shows the series of actions applied change the state of the IV object until it reaches a "published" state. In this diagram, only the persistent data object is shown. The diagram is linear and depicts contains only a subset of the allowed states and actions. This subset can vary from one RI to another. RIs can model their data lifecycles using state machine diagrams. The states described in the IV States diagram indicate a set of possible paths to follow when representing the data lifecycle. The instances developed by RIs may chose the states they



need to include to represent their corresponding data lifecycles. RIs can also include additional states which help them better represent their data processing flows.



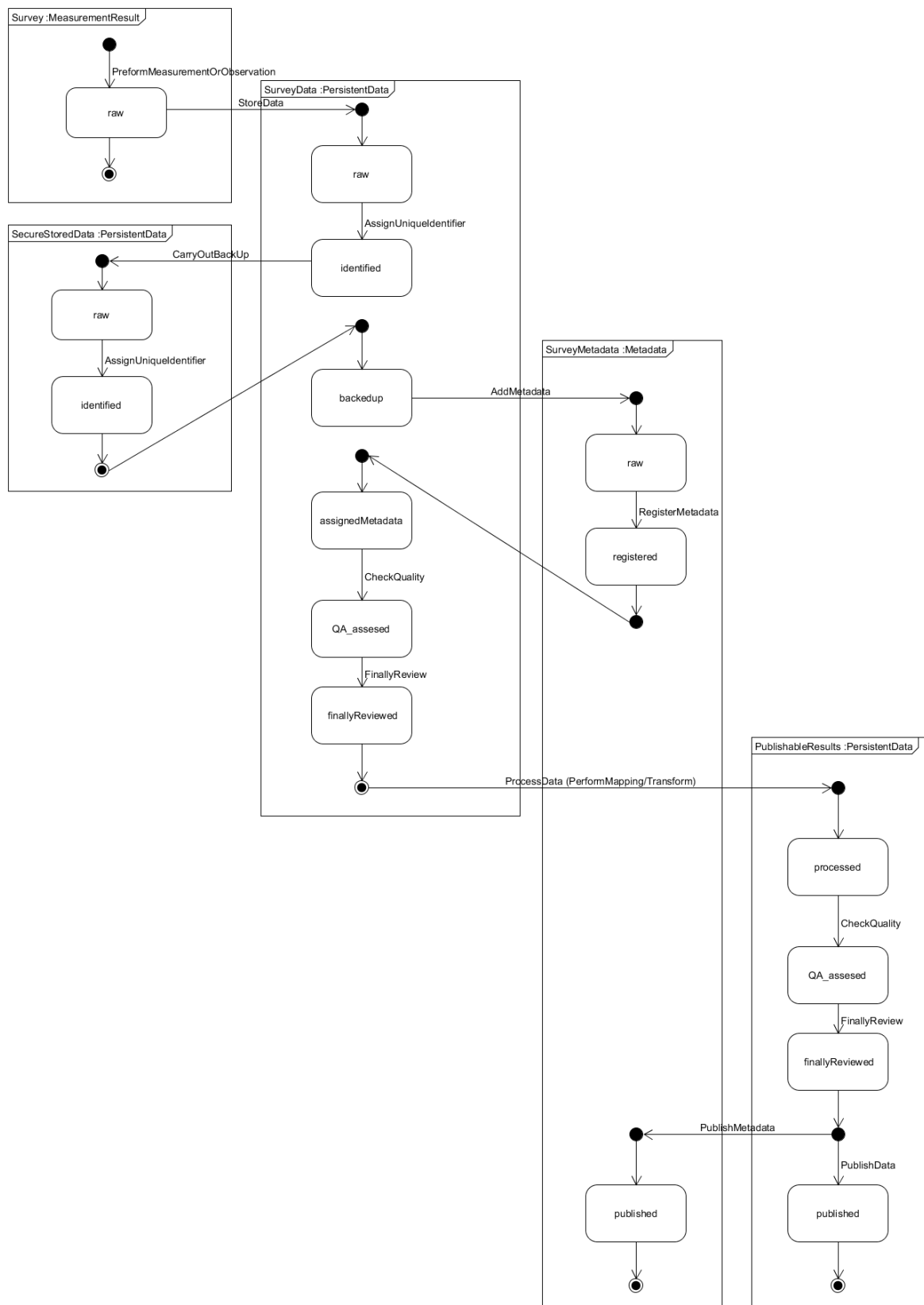
Advanced use example

The diagram shows the lifecycles of five IV objects and the way in which they relate to each other. The example is more complex but it is still linear and easy to follow.

Reading the diagram from the left to right and top to bottom it is possible to describe the lifecycles of the five IV objects. The first object is a called Survey. The diagram indicates that Survey is a type of MeasurementResult, when the Survey object is stored a new object is created SurveyData, a type of PersistentData object. The SurveyData object is then Identified, after identifying, a back up copy of the object, named SecuredStoredData, is created. Once the SecuredStoredData is stored and identified, the state of the SurveyData object changes to backedup. After backup, the RI adds metadata to the SecuredData object. The SurveyMetadata object is created and registered. This causes the change in the state of the SurveyData object to assignedMetadata. Subsequently the object is QA assessed and reviewed. Once the SurveyData is finally reviewed, the RI creates a new PersistentData object, PublishableResult, which results from the processing of the SurveyData object. The published result data object is then QA assesed and finally reviewed. The final pair of activities publish both the PublishableResult and the SurveyMetadata objects.

In this example, most of the states and actions are still subsets from the ones originally introduced on he IV State Diagram. The only exception is the backedup state. This is still valid, RIs can adapt the diagrams to their particular needs adding states and actions to better illustrate their data lifecycles as close to reality as possible.

## IV State Diagrams - Advanced use example



Notation

## IV Information Action Types

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

The figure shows a collection of action types specified in the information viewpoint.



### IV Actions

- specify investigation design
- specify measurement or observation
- perform measurement or observation
- store data
- carry out backup
- final review
- publish data
- add metadata
- annotate metadata
- register metadata
- publish metadata
- query metadata
- build conceptual models
- setup mapping rules
- annotate data
- annotate action
- resolve annotation
- perform mapping
- do data mining
- query data
- assign unique identifier
- check quality
- track provenance
- process data
- describe service

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.

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

perform measurement or 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 continued accessibility and usability.

carry out backup

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

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

add metadata

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

**annotate metadata**

Link metadata with meaning (concepts of predefined local or global conceptual models). This can be done by adding pointers from concepts within a conceptual model to the metadata. For instance, if concepts are terms of a SKOS thesaurus, identified by URIs and published as linked data, then annotation amounts to associating metadata with the terms' URIs.

**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.

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 conceptual model should be unambiguously defined;
- readability: the model should be readable by both human and machine. Ontologies, for instance, express the meaning of the concepts with the relations to other concepts while being 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 using a language that can be processed by software.

A minimal set of mapping rules should include the following data:

- 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.)

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**

Perform annotation of an information object

**resolve annotation**

Retrieve the reference to the specific set of objects that correspond to a set of annotation terms.

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).

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.

#### **query data**

Send a request to a data store to retrieve required data.

In practice, there are two types of data query:

- 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.

#### **assign unique identifier**

Obtain a unique identifier and associate it to the data.

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).

#### **track provenance**

Automatically generate and store metadata about the actions and the data state changes as provenance instances.

#### **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 instances.

#### **describe service**

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

## **IV Information Objects Lifecycle**

The specification of the lifecycles of the information objects is described combining IV object instances at different states and the sequences of allowed actions according to those states. The set of models used for describing this evolution are part of the dynamic schemata in ODP [3 7].

The specification of information objects lifecycle is presented in two parts:

- **Lifecycle overview**: overview of information objects state changes as effects of actions.
- **Lifecycle in detail**: detailed description of how information objects changes at each phase of the data lifecycle.

### **IV Lifecycle Overview**

This section describes the alignment between data processing in the RI systems and the data lifecycle using **information objects** and **information actions**. The description is framed against the phases of the **research data lifecycle model**.

The diagram shown on the right provides a high level view of the data lifecycle. The rounded rectangles represent IV actions on

data and the straight rectangles represent instances of IV objects at different states. The arrow lines link IV actions and IV objects as follows: arrows leaving an action connect to IV objects created by the action while arrows entering an action connect IV objects to actions applied on them. The black circle at the top of the diagram represents the starting point and the double circle at the bottom represents the end point. The types of diagrams used in this section are called activity diagrams (UML).

In the diagram each phase of the data lifecycle is represented as an action which produces a specific information object, in this case the main information object shown is persistent data. The diagram also adds a provenance tracking action. Provenance tracking is an action that can proceed in parallel during all phases of the data lifecycle. The overview of the data lifecycle phases is described as follows.

**Data Acquisition:** The data acquisition phase encompasses the actions defined for the observation/experimentation, storage, identification and storage of measurements/observations (raw data). In the diagram, the acquisition phase is represented by the "DataAcquisition" action which produces a measurement result data object with the state raw.

**Data Curation:** The data curation phase encompasses the actions that support the long term preservation and use of research data. The main product of this set of actions is persistent data in a stable state (curated data). In the diagram, the curation phase is represented by the "DataCuration" action which produces a persistent data object with the state curated.

**Note**  
Data curation includes preservation which may require data transformation, for example media migration to a digital form.

**Data Publishing:** The data publishing phase encompasses the actions that guaranty data access and discovery for entities (people and systems) outside the RI. In the diagram, the publishing phase is represented by the "DataPublishing" action which produces a persistent data object with the state published.

**Data Processing:** The data processing phase encompasses the actions that support making use of the RI published data. In the diagram, the processing phase is represented by the "DataProcessing" action which produces a persistent data object with the state processed.

**Data Use:** The data use phase is a bridge phase which sits between processing and acquisition. In this phase, the data is used and may produce new data (raw data) which can in turn be persisted by an RI. In the diagram the usage phase is represented by the "DataUse" action which produces a data product object with the state raw.

In the **detailed description** section, the actions in the diagram are expanded to present a more detailed view of the data lifecycle from the IV perspective.

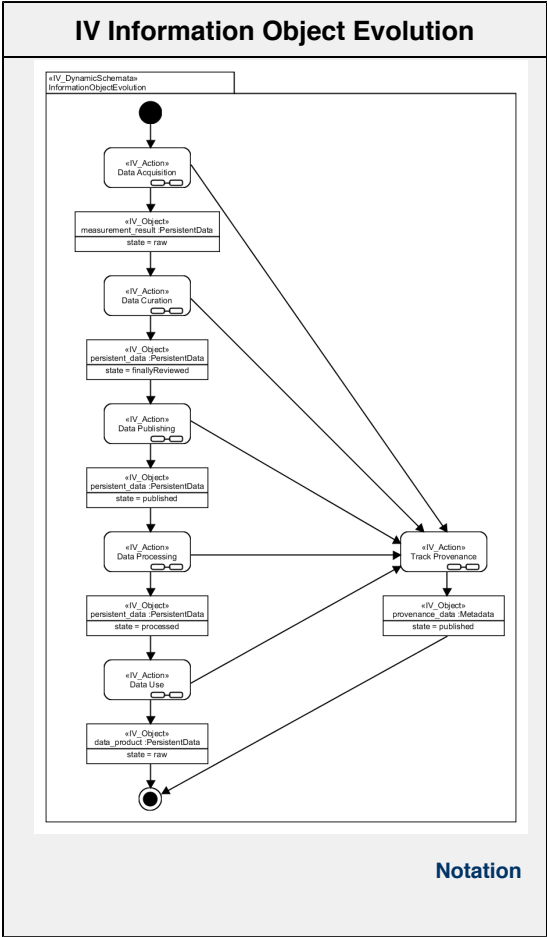
**Data Provenance Tracking**

It is important to track state changes of information objects during their lifecycle. As illustrated in diagram above, the ProvenanceTracking action takes place in parallel to the phases of the lifecycle that change the state of persistent data.

Some of the states changes of information objects as effects of actions are summarised in the following table. As shown in the diagram, the outputs of each transition in which a new stable state is reached can be used to produce provenance data. For example, a provenance tracking service may record information objects being processed, action types applied and resulting objects, the timestamps for the actions, and some additional data and store that as provenance data.

**Simplified example of some provenance tracking points**

Information Object	Applied Action Types	Resulting Information Objects
--------------------	----------------------	-------------------------------



	Data Acquisition	persistent data (raw)
persistent data (raw)	Data Curation	persistent data (finallyReviewed) metadata (registered)
persistent data (FinallyReviewed) metadata(registered)	Data Publishing	persistent data (published) metadata (published)
persistent data (published)	Data Processing	persistent data (processed)
persistent data (processed)	Data Use	data product (new form of persistent data (raw))

The citation of data referencing the actors of involved in production of the data is an example of the use of data provenance

Correct interpretation of the data can also depend on reviewing the provenance, for instance to ensure origin of the data matches its intended use.

## IV Lifecycle in Detail

This section expands the **overview** of the alignment between the information viewpoint and the data lifecycle. The descriptions uses the **information objects** and **information actions** to a greater extent providing a deeper insight into the processing of information objects by the RI.

The notation for the diagrams in this section is as follows. The rounded rectangles represent IV actions on data and the straight rectangles represent instances of information objects at different stages. The arrow lines link actions and objects as follows: arrows leaving an action connect to IV objects created by the action while arrows entering an action connect IV objects to actions using them.

- Data Acquisition
- Data Curation
- Data Publishing
- Data Processing
- Data Use

### Data Acquisition

The data acquisition phase encompasses the actions defined for the observation/experimentation, storage, identification and backup of measurements/observations (raw data).

The following paragraphs explain the detailed diagram of how the IV actions can be combined to support data acquisition.

#### Note

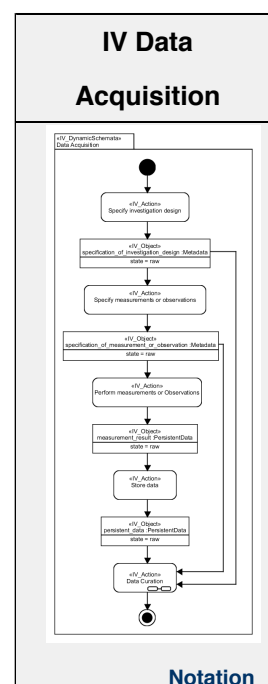
This example is provided for illustrative purposes. The example shows one of many alternatives for performing data acquisition. Other IV actions and IV objects can be introduced at this stage. Additional actions and objects not described in the IV of the ENVRI RM can also be incorporated.

**Specify investigation design:** 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, details about their processing, and the parameters defined have to be available (e.g. if a stratified selection of sites according to parameter A is done, the resulting value of parameter A can not be evaluated in the same way as other results).

**Specify measurement or observation:** 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. The details of the process and the parameters used have to be preserved to guarantee correct interpretation of the resulting data (e.g. when modelling 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).

**Perform measurement or observation:** After the measurement or observation method is defined, the experiment can be performed, producing measurement result(s) which is a form of persistent data in a raw state.

**Store data:** The measurement result data is stored. 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



Notation



a process done by humans. The storage process is the first step in the lifecycle of data that makes data accessible in digital form.

**Data curation:** Once data is stored, the next phase of the data lifecycle is data curation.

## Data Curation

The data curation phase encompasses the actions that support the the long term preservation and use of research data. The main product of this set of actions is persistent data in a stable state (annotated data). The following paragraphs explain the detailed diagram of how the IV actions can be combined to support data curation.

### Note

This example is provided for illustrative purposes. The example shows one of many alternatives for performing data curation. Other IV actions and IV objects can be introduced at this stage, for instance: Check quality, Register metadata, or Publish metadata. Actions and objects not described in the IV of the ENVRI RM can also be incorporated.

**Data Acquisition:** The first action is Data Acquisition, the phase of the data lifecycle that precedes data curation. This action produces three IV Objects: PersistentData, SpecificationOfMeasurementsOrObservations and SpecificationOfInvestigationDesign.

**Carry out backup:** As soon as data are available to the RI a backup can be made, independently of the state of the persisted data. This can be done locally or remotely, by the data owners or by dedicated data archiving centres.

**Assign Unique Identifier:** Data needs to be uniquely identified for correct retrieval and processing, the unique identifier can be local to the RI or global, to be used from outside the RI. As such it can be a simple numerical value assigned by the RI DBMS or a specific PID assigned following the standards of an external PID provider.

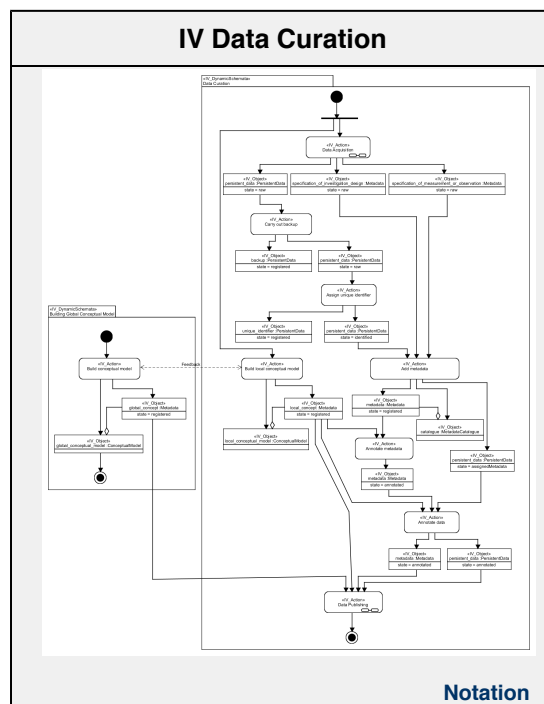
**Add metadata:** This action uses the specifications of investigation and measurements to facilitate the understanding of the associated persistent data object. In addition to this data the RI can add timestamps, and other identification data as metadata. Once the data is correctly stored and identified, and the corresponding metadata has been also created, persistent data can be linked to metadata.

**Annotate data:** Data is further enriched with additional metadata which can correspond to a specific ontology for the research field.

**Annotate metadata:** Metadata can also be further enriched with additional metadata which can correspond to a specific ontology for the research field.

**Build conceptual model:** The building of a **local conceptual model** mirrors the wider research community efforts to build a global conceptual model. In this set of activities concept are added to the local conceptual model of the RI. The conceptual model is made of the composition of concepts, which are used to help people know, understand, or simulate a subject the model represents. The pairing of data and metadata using semantic annotations creates a local concept (a new metadata object) and changes the state of the persistent data object to annotated.

**Global conceptual models** are ontologies, thesauri, dictionaries, or hierarchies built by a larger communities than a single RI, such as GEMET, DOLCE, SWEET. This action normally happens outside of the RI's main activities. Trough feedback mechanisms RIs participate in the creation of global conceptual models while developing their own models..



**Data Publishing:** Once data have been curated, the next phase of the data lifecycle is data publishing.

## Data Publishing

The data publishing phase encompasses the actions that make the data available for entities (people and systems) outside the RI. The following paragraphs explain the detailed diagram of how the IV actions can be combined to support data curation.

### Note

This example is provided for illustrative purposes. The example shows one of many alternatives for performing data curation. Other IV actions and IV objects can be introduced at this stage, for instance: QualityAssurance. Actions and objects not described in the IV of the ENVRI RM can also be incorporated.

**Data Curation:** The first action is Data Curation, the phase of the data lifecycle that precedes data Publishing. This action produces four IV Objects: PersistentData, LocalConceptualModel, LocalConcept, and Metadata.

**Finally Review Data:** Persistent data that is in the process of publishing needs to be reviewed before proceeding to publishing. It is important to clearly specify what the "finallyReviewed" state means. In some RIs it can mean, that those data will never change again, the optimum for the outside user. For other RIs 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" mean.

**Build Global Conceptual Model:** The construction of a global conceptual model makes sure that there is an appropriate fit between the persistent data to be published and their metadata (including the local conceptual model) with other models existing outside the RI. The GlobalConceptualModel is the representation of how that outside world looks to the RI.

**Semantic Harmonisation:** unifies data (and knowledge) models based on the consensus of collaborative domain experts to achieve better data (knowledge) reuse and semantic interoperability. This complex activity is performed in two stages: setup mapping rules and perform mapping, defined as follows.

**Setup Mapping Rule:** The Global model is used to generate a set of mapping rules to enable linking the RI data and metadata to global semantics. This may include simple conversions, such as conversions of units, but may also imply more sophisticated transformations like transformations of code lists, descriptions, measurement descriptions, and data provenance.

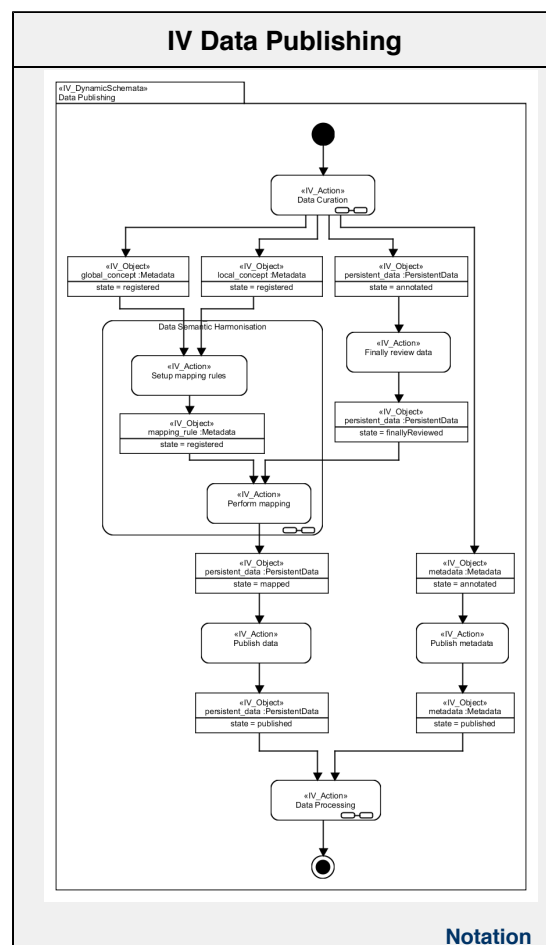
**Perform mapping:** This action carries out the linking of data and metadata to one or more global models.

**Publish data:** Mapped data is made available to the outside world. The PID is the main identifier of the data but the data can also be located by querying metadata.

**Publish metadata:** Metadata is also mapped and published to enable more sophisticated data querying.

**Data Processing:** Once data have been published, the next phase of the data lifecycle is data processing.

Data can be made directly accessible or indirectly. Direct access means, that a data request to a data server (query data) gets the data or an error message as answer. Indirect access means, initially accessing metadata (query metadata), searching for a fitting data set and then querying on the resulting data set. Those two steps can be extended further, when intermediate steps are involved. The multi-step approach is often used for data, which are not open, making metadata open but not data itself. For queries touching several data sets and/or filtering the data (like e.g. give me all NOx air measurement where O3 exceeds a level of Y ppb) the multi-step approach can be seen blocker.



## Data Processing

The data processing phase encompasses the actions that support making use of the RI published data. The following paragraphs explain the detailed diagram of how the IV actions can be combined to support data curation.

### Note

This example is provided for illustrative purposes. The example shows one of many alternatives for performing data processing. Other IV actions and IV objects can be introduced at this stage. Actions and objects not described in the IV of the ENVRI RM can also be incorporated.

**Data Publishing:** The first action is Data publishing, the phase of the data lifecycle that precedes Data Processing. This action produces three IV Objects: PersistentData, GlobalConceptualModel, and Metadata which are used in the data processing phase to access and process data.

**Provenance Tracking:** Is the action that keeps a log about the the actions and the data state changes as data evolves through the RI systems. The resulting provenance data is a form of metadata which may be of interest for referencing and citing the use of data within and outside the RI.

**Query data:** This action requests specific persisted data from the RI.

**Do data mining:** This action implies the execution of a sequence of metadata/data request/interpret/result/request which automatically produce or find patterns in the data being analysed. Usually this sequence helps to deepen the knowledge about the data.

**Resolve annotation:** This action implies finding a specific data set from a set of semantic annotation and constrains on those annotations. If the annotation is resolved the result should be a link or a set of links to specific data sets, if not the result is an empty set.

**Query metadata:** This action requests specific persisted data from the RI using metadata as additional parameters for narrowing down the search.

**Query provenance:** This action requests specific persisted data about the provenance of some data or metadata. This is usually done to determine the origin and validity of data but can also be helpful for citation and referencing.

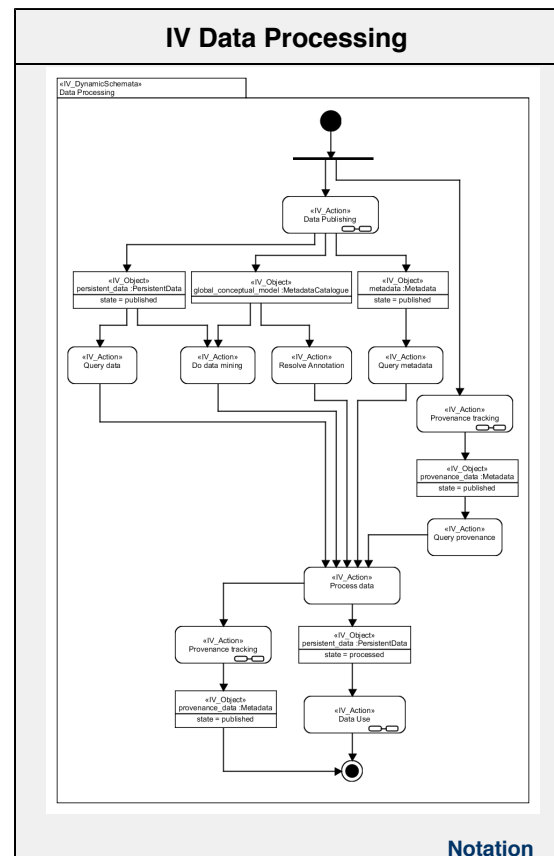
**Process data:** The performance of any of the five actions listed before, is automatically detected as a form of data processing by the RI system. This should result in changing the state of the data to "processed". The processed state can mean several things such as: the data has been consulted, the data has been referenced, the data has been downloaded, the data has been used as input for an external process, etc.

**Data use:** Once data have been processed, the next phase of the data lifecycle is data use, which to some extent overlaps with processing.

**Provenance Tracking:** As described in the overview, the provenance tracking action tracks all changes in the states of persistent data. This is an important action which has wide use inside and outside the RI.

## Data Use

The data use phase is a bridge phase which sits between processing and acquisition. In this phase, the data is used and may produce new data (raw data) which can in turn be persisted by an RI. The actions that act on data at this point can be



provided by same RI exposing the data or by external entities (RIs or other).

In the use phase, the RI system is open to the outside world. Users (persons or external systems) can use the services provided to produce new data products.

#### Note

This example is provided for illustrative purposes. The example shows one of many alternatives for performing data processing. Other IV actions and IV objects can be introduced at this stage. Actions and objects not described in the IV of the ENVRI RM can also be incorporated. In the diagram and associated descriptions below, cite data, convert data, produce model, and visualise data are some examples of these types of actions.

**Data Publishing:** The first action is Data publishing, the phase of the data lifecycle that precedes data Processing. This action produces three IV Objects: PersistentData, GlobalConceptualModel, and Metadata which are used in the Data Use phase to access and process data.

**Provenance Tracking:** Provenance tracking keeps a log about the the actions and the data state changes as data evolves through the RI systems. The resulting provenance data is a form of metadata which may be of interest for referencing and citing the use of data within and outside the RI.

**Data Processing:** Data Processing produces Persistent Data IV Objects.

**Query data:** This action requests specific persisted data from the RI.

**Do data mining:** This action implies the execution of a sequence of metadata/data request/interpret/result/request which automatically produce or find patterns in the data being analysed. Usually this sequence helps to deepen the knowledge about the data.

**Resolve annotation:** This action implies finding a specific data set from a set of semantic annotation and constrains on those annotations. If the annotation is resolved the result should be a link or a set of links to specific data sets, if not the result is an empty set.

**Query metadata:** This action requests specific persisted data from the RI using metadata as additional parameters for narrowing down the search.

**Query provenance:** This action requests specific persisted data about the provenance of some data or metadata. This is usually done to determine the origin and validity of data but can also be helpful for citation and referencing.

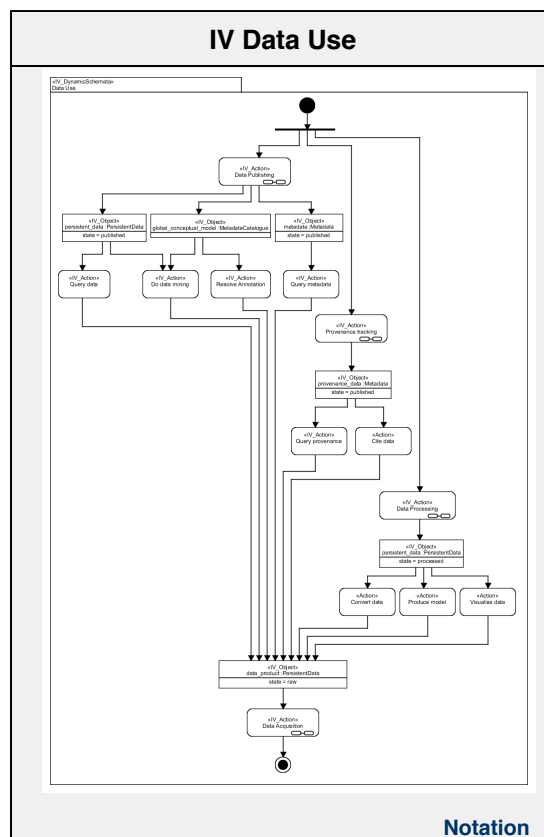
**Cite data:** Produce a reference to persistent data or metadata.

**Convert data:** converting and generating data products, for instance translating to a different format.

**Produce model:** creation of statistical models, simulation models or summaries with the data provided.

**Visualise data:** creating visual models which display data alpha-numerically, graphically, or geographically.

**Data Acquisition:** Use of data has the potential for creating data products which may need to be persisted, re-initiating the data lifecycle. For this reason, the the last action after Data Use actions is Data Acquisition.



## IV Information Management Constraints

The IV of the ENVRI RM provides the means for specifying constraints which describe the set of rules governing data management. The

set of models used for describing constraints are part of the static schemata in ODP [37]. In the ENVRI RM information management constraints establish mechanisms to:

1. avoid loss of data around measurements and observations.
2. provide information about the meaning of data.
3. make data and metadata available for external use.

The IV of the ENVRI RM provides three types of management constraints:

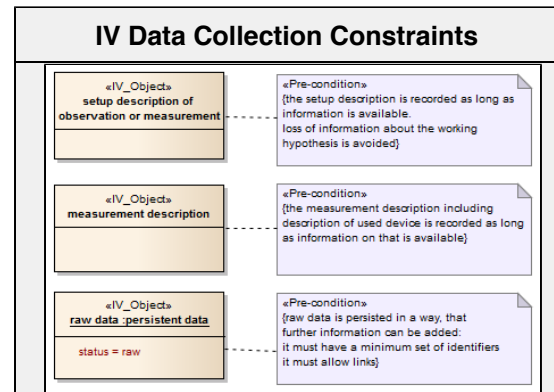
- [Data Collection Constraints](#)
- [Data Integration Constraints](#)
- [Data Publication Constraints](#)

## Data Collection Constraints

The constraints applied to data collection are illustrated in the figure below. The application of these constraints helps to avoid data loss or wrong interpretation.

Observing these three rules together ensures that data can be correctly interpreted and reduces the risk of data loss. This is because the rules guarantee that the original data can be retrieved and interpreted correctly through the lifetime of the information objects derived from them.

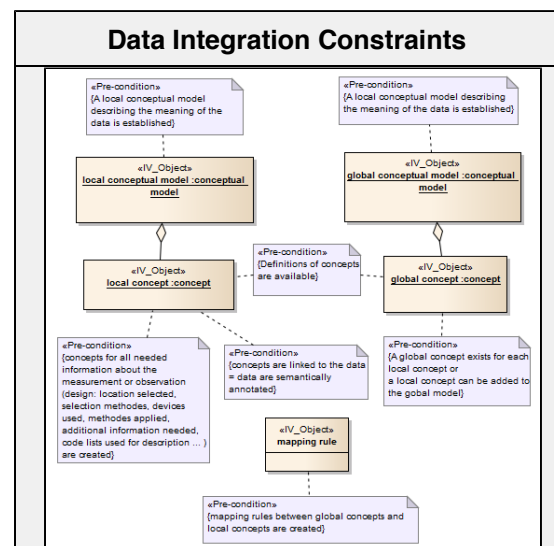
The rules guarantee the availability of the rationale for collecting (1st rule), the details about the how collection proceeded (2nd rule), and the original data collected.



## Data Integration Constraints

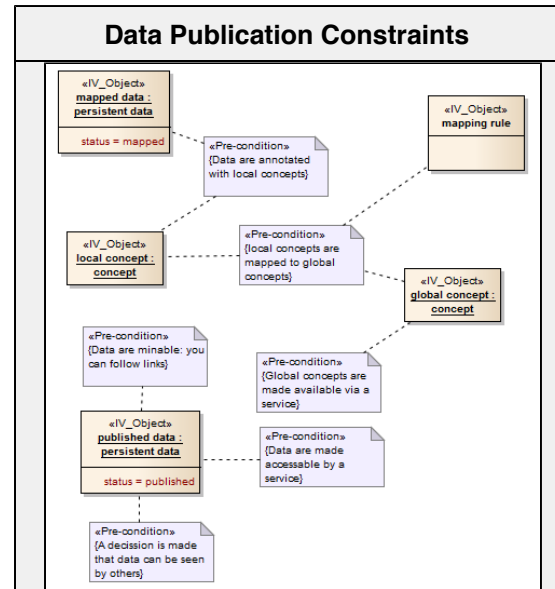
The constraints for data integration are illustrated in the following figure. Data integration constraints support the correct interpretation of data, helping external data users correctly interpret and map the semantics of data.

The observation of these rules makes possible integrating data within a RI and the outside world. This requires adding a special type of metadata, a local conceptual model and then mapping the data to a global conceptual model. Mapping data to global semantics may include simple tasks such as conversions of units, but can also need sophisticated transformations such as code lists cross-referencing or measurement descriptions, and data provenance.



## Data Publication Constraints

Constraints for data publication are illustrated in the following figure. The constraints specify conditions necessary for preparing the data to be publicly accessed.



## Computational Viewpoint

A research infrastructure (RI) provides a context in which investigators can interact with scientific data in a principled manner. To provide this context, an RI must support a portfolio of possible research interactions. These interactions can be realised by binding together different services via standard operational interfaces.

The Computational Viewpoint (CV) accounts for the major computational objects that can be found within an environmental research infrastructure, as well as the interfaces by which they can be invoked, and by which they can invoke other objects in the infrastructure. Each object encapsulates functionality that should be implemented by a service or other resource in a compliant RI. Binding of computational objects together via compatible interfaces creates a network of interactions that allows an RI to support the data related activities of its target research community.

The description of the CV is divided in three parts: **objects**, **support of data lifecycle**, and **integration points**.

- **Objects**: present computational objects according a generic architecture of the RIs.
- **Subsystems**: presents examples of how components are integrated for supporting the data lifecycle into five different subsystems.
- **Integration points**: defined to support the movement of research data between phases.

### Note

Before proceeding, the reader may wish to study the pages on [how to read the computational viewpoint](#) and [how to use the computational viewpoint](#).

The Computational Viewpoint defines computational objects (CV Object) and interfaces (CV Interfaces) which enable their interaction.

The diagram below shows the main elements of the CV and their relationships. Each ellipse contains a concept. The arrows connecting the concepts are directed and indicate the relationship between concepts. The label of the link indicates the type of relationship. From this, the diagram indicates that a CV object provides a CV interface, as indicated by the **provides** relationship. Similarly, a CV object can create another CV object, as indicated by the **canInstantiate** relationship. In this same way a CV interface can fit another CV interface, this is indicated the **fits** relationship.

Computation Viewpoint components and their relationships

## CV Objects

The archetype of a modern environmental research infrastructure has a brokered, service-oriented architecture. Core functionality is encapsulated within a number of key resources which can be accessed by means of externally-facing gateway services. Interaction by external agents with internal resources is overseen by one or more brokers (often closely integrated with the gateway) charged with validating requests and providing, where needed, an interoperability layer between otherwise heterogeneous components. The Computational Viewpoint of the ENVRI RM



provides a set of models which can help in the design, implementation, maintenance, and evolution of the systems and services that RIs provide for accessing data.

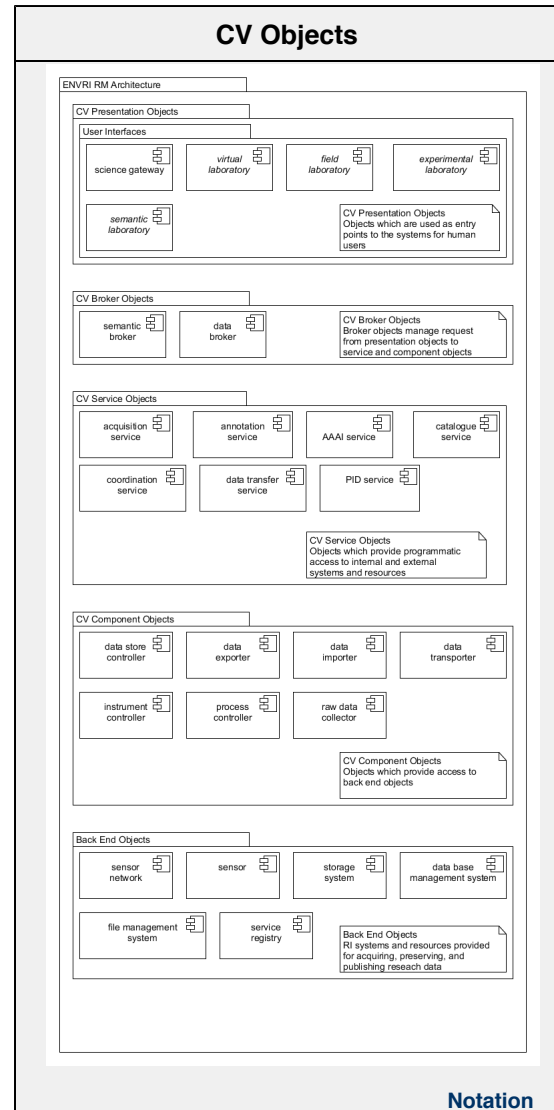
The CV prescribes a number of types of computational object for which there should be instances present in or around a research infrastructure in order to ensure that particular **key functions** are supported. The grouping of CV objects into sets is based on the software architecture that is expected to be implemented when providing access to data and other related resources during the research **data lifecycle** (evident in the RIs analysed as part of the **ENVRI and ENVRIplus**) projects. Consequently, the presentation of CV objects is arranged as five sets corresponding to each of the architectural layers of RI systems (note however, that this should not be read as prescriptive and that other groupings are possible):

- **Presentation Objects:** computational objects that facilitate access to RIs by human users.
- **Broker Objects:** computational objects that act as intermediaries for access to data held within the data store and facilitate performing semantic interpretation and routing of queries.
- **Service Objects:** computational objects that offer programmatic access to distributed systems and resources (internal and external).
- **Component Objects:** computational objects that provide access to back end objects.
- **Back End Objects:** computational objects that encompass the RI's systems and resources for accessing research data and derived data products.

The set of CV components included in the ENVRI RM comprises the computational objects that are common to many RIs. The set is not closed, so each RI can include the additional components they require to completely model their systems. The set does not contain compulsory items, so each RI can exclude objects and interfaces that are not relevant for them.

#### Note

Before proceeding, the reader may wish to study the pages on [how to read the computational viewpoint](#) and [how to use the computational viewpoint](#).



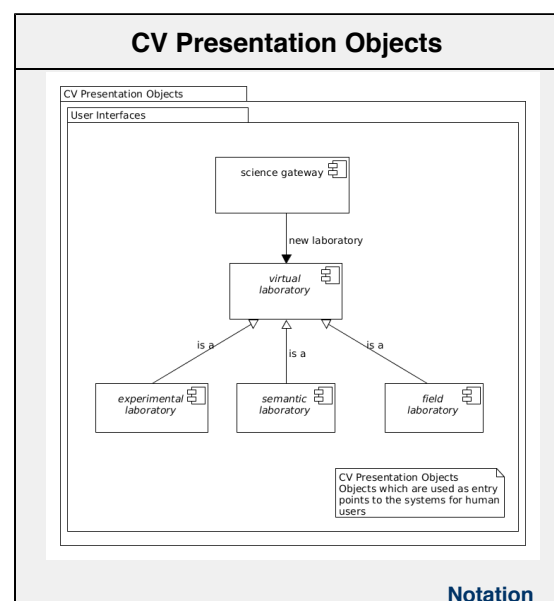
## CV Presentation Objects

CV Presentation objects are the entry points for human users to the systems and services provided to access research data and their derived products.

In the ENVRI RM, complex interactions between the components facilitating data use and other components are mediated by **virtual laboratories**; these objects are deployed by **science gateways** in order to provide a persistent context for such interactions between groups of users and components within the RI.

The Reference Model recognises the following specific sub-classes of laboratory:

- **Field laboratories** (so-named because they interact with raw data sources 'in the field') are used to interact with the **data acquisition** components, 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** from the data acquisition set.
- **Experiment 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.





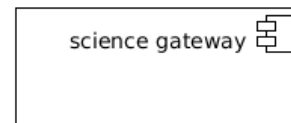
Regardless of provenance, all laboratories must interact with a **AA AI service** in order to authorise requests and authenticate users of the laboratory before they can proceed with any privileged activities.

## Science gateway

*Community portal for interacting with an infrastructure.*

A science gateway object encapsulates the functions required to interact with a research infrastructure from outside with the objects provided for 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 is also known as a Virtual Research Environment.

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



## Virtual laboratory

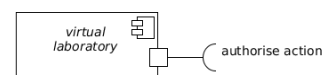
*Community proxy for interacting with RI components.*

A virtual laboratory 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 **AAAI 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 components.

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



## Field laboratory

*Community proxy for interacting with data acquisition instruments.*

A sub-class of **virtual laboratory** object encapsulating the functions required to access, calibrate, deploy or withhold instruments during the data acquisition phase.



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

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/or withdraw instruments used for data acquisition.

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

## Experiment laboratory

*Community proxy for conducting experiments within a research infrastructure.*

A sub-class of **virtual laboratory** 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 to allow researchers

interaction 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 laboratory

*Community proxy for interacting with semantic models.*

A sub-class of **virtual laboratory** 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.
- **semantic data request (client)** is used to make requests of the research infrastructure about metadata and annotations referring to data stored by the data set.

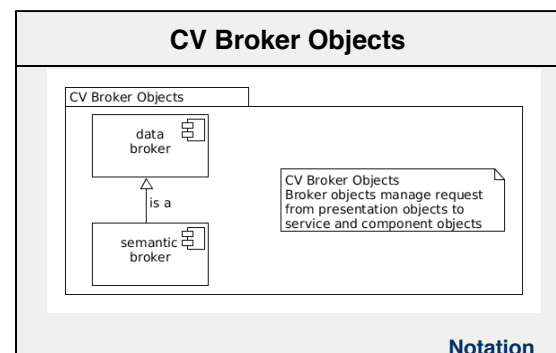


## CV Broker Objects

Broker objects act as intermediaries for access to data held within the data store and facilitate performing semantic interpretation and routing of queries. For this brokers keep registries of **service objects** to which actions are routed. Whenever possible, advanced brokers which make use of metadata should be preferred to hard coded brokers.

- **data broker** objects act as intermediaries for access to data held within the data store.
- **semantic broker** objects perform semantic interpretation.

Brokers are responsible for verifying the agents making access requests and for validating those requests. 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).



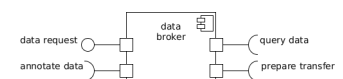
**Notation**

## Data broker

*Broker for facilitating data access/upload requests.*

A data broker object intercedes between the data publishing objects and the data curation objects, 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 objects, as this is handled more efficiently by an acquisition service.

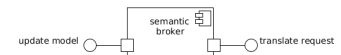


A data broker should provide four operational interfaces:

- **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 objects of a research infrastructure.
- **prepare data transfer (client)** is used to negotiate data transfers with the data curation

- objects of a research infrastructure.
- **query data (client)** is used to forward queries onto the data curation objects of a research infrastructure and receive the results.

## Semantic broker



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

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 provides the functionalities required to update the semantic models used by an infrastructure to describe data held within.

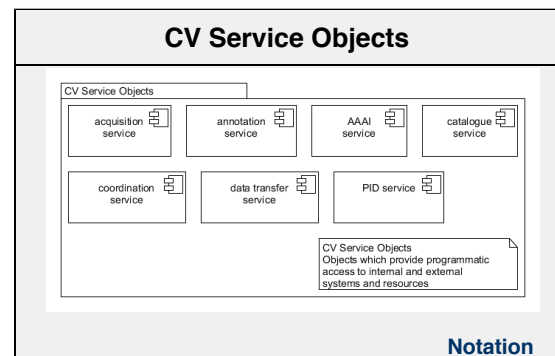
A semantic broker should provide two operational interfaces:

- **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.

## CV Service Objects

CV service objects offer programmatic access to distributed systems and resources (internal and external). This allows building RIs using both internal and external sourced components. The service layer includes the main services that enable data access, processing and transformation used in different phases of the research data lifecycle.

- The **acquisition services**, responsible for ensuring that any data is delivered into the infrastructure in accordance with current policies.
- The **annotation service**, concerned with the updating of records (such as datasets) and catalogues in response to user annotation requests.
- The **AAAI service** handles authorisation requests and authentication of users before they can proceed with any privileged activities.
- The **catalogue service**, concerned with the cataloguing of metadata and other characteristic data associated with datasets stored within the infrastructure.
- The **coordination service** delegates all processing tasks sent to particular execution resources, coordinates multi-stage workflows and initiates execution.
- The **data transfer service**, concerned with the movement of data into and out of the infrastructure.
- The **PID service** provides globally-readable persistent identifiers (PIDs) to infrastructure entities, mainly datasets, that may be cited by the community.



**Notation**

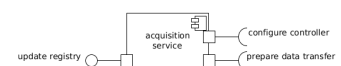
## Acquisition service

*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 phase.
- **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 to data curation objects.



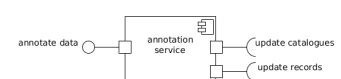
## Annotation service

*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 (server)** 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 update annotation records of existing datasets curated within one or more data stores.

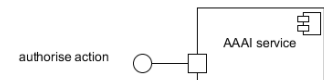
## AAAI service

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

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

An AAAI 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



## Catalogue service

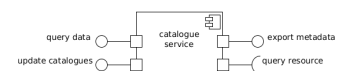
*Oversight service for cataloguing curated datasets.*

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.

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 the data curation store objects (data stores).
- **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.



## 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 could require executing 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 store objects in order to register new results or in preparation for the generation of such results.



## Data transfer service

*Oversight service for the transfer of data into and out of the data store objects.*

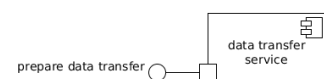
A data transfer service object encapsulates the functions required to integrate new data into the RI and export that integrated data on demand. The data transfer service is responsible for setting up data transfers, including any repackaging of datasets necessary prior to delivery.

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 stores of a RI.

## 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.

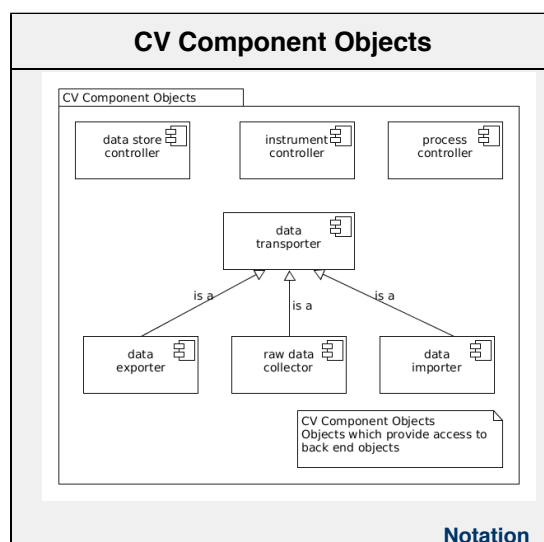
A PID service should provide at least two operational interfaces:

- **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).

## CV Component Objects

CV component objects offer programmatic access to the actual RI's systems and resources, the back end objects. This allows providing intermediate façades for systems and resources which may be inter-changed or replaced as needed.

- **Data store controllers** provide access to data stores that may have their own internal data management regimes.
- **Instrument controllers** encapsulate the accessible functionalities of instruments and other raw data sources out in the field.
- **Process controllers** represent the computational functionality of registered execution resources.
- **Data transporters** are provided for managing the movement of data from one part of a research infrastructure to another.
- **Raw data collectors** manage the movement of data from one or more data acquisition objects to one or more data store objects.
- **Data importers** manage the movement of data from external sources (such as user-originated datasets and derived datasets from data processing) to one or more data stores objects.
- **Data exporters** manage the movement of data from one or more data store objects to external destinations (such as a user machine or downstream service gathering data from the research infrastructure).



## Data store controller

*A data store supporting data preservation.*

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 a data store of the RI, 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.

## 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.

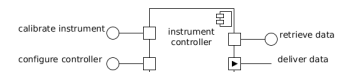
'Instrument' is a logical entity, and may to multiple physical entities deployed in the real world should they act in tandem sufficiently closely to justify being treated as one data source. Any instrument represented by an instrument controller should however be considered independently configurable 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.
- **retrieve 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.



## Process controller

*Part of the execution platform that controls the deployment of processing components and the assignment of processing tasks.*

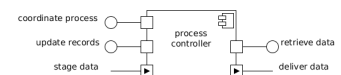
A process controller object encapsulates the functions required for using an execution resource (generically, any computing platform that can host some process) as part of any infrastructure workflow.

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 stream interfaces:

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

- **update records (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 acquisition objects.

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 (data acquisition objects).
- **import data for curation (producer)** is used to deliver (repackaged) raw data to data store objects.

## Data importer

*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 RI.

A data importer 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 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 RI.
- **import data for curation (producer)** is used to deliver (repackaged) datasets to one or more data stores within the RI.

## Data exporter

*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 objects 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.
- **deliver dataset (producer)** is used to deliver (repackaged) curated data to a designated external data store outside of the RI.

## CV Back End Objects

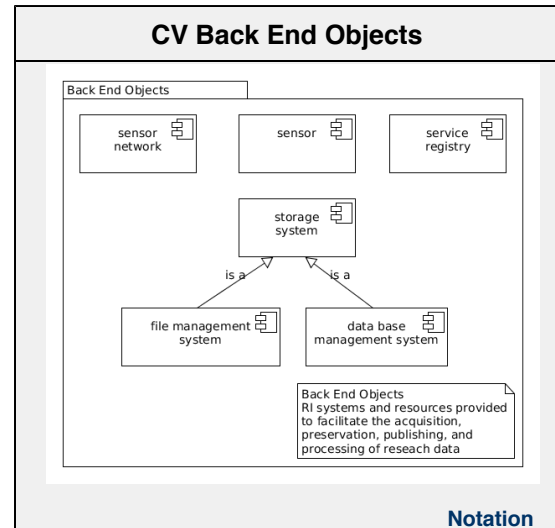
Back End Objects are computational objects which encompass





the RI's systems and resources provided for acquiring, preserving, publishing, and processing research data and derived data products.

- **Sensor network:** is a network consisting of distributed sensors which monitor physical or environmental conditions.
- **Sensor:** 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.
- **Storage System:** is a systems that manages the storage and retrieval of data and metadata.
- **File Management System:** is a storage systems that manages the storage and retrieval of data as files in a computer system.
- **Database Management System:** is a storage systems that manages the storage and retrieval of data and metadata into logically structured repositories.
- **Service Registry:** is an information system for registering services.



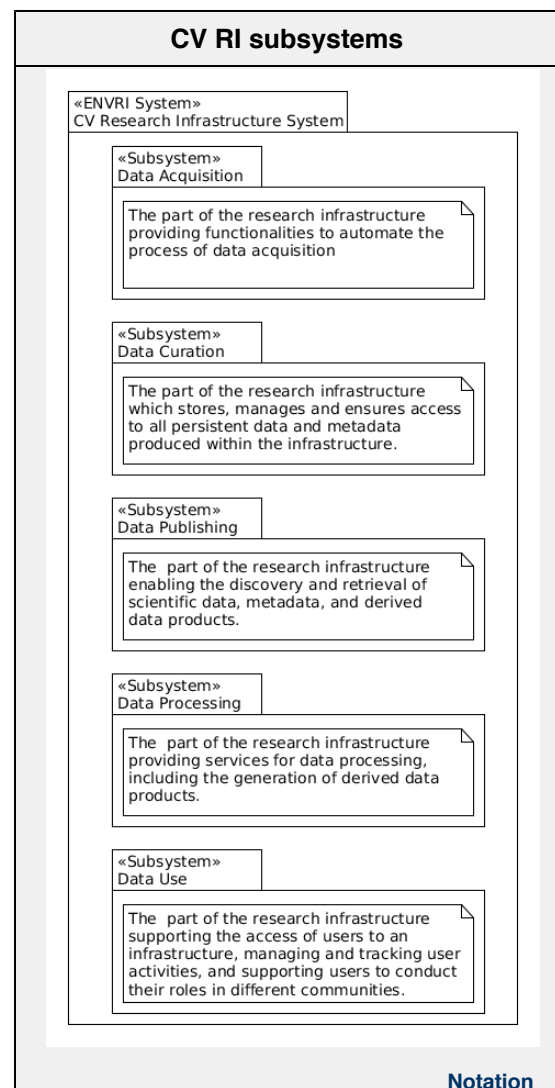
## CV Objects and Subsystems

The **science viewpoint roles** include five subsystem roles which support each of the phases of the data lifecycle. In this section, the models of those subsystems are developed further using computational viewpoint components. The five subsystems defined are:

- **Data acquisition**
- **Data curation**
- **Data publishing**
- **Data processing**
- **Data use**

### Note

Before proceeding, the reader may wish to study the pages on [how to read the computational viewpoint](#) and [how to use the computational viewpoint](#).



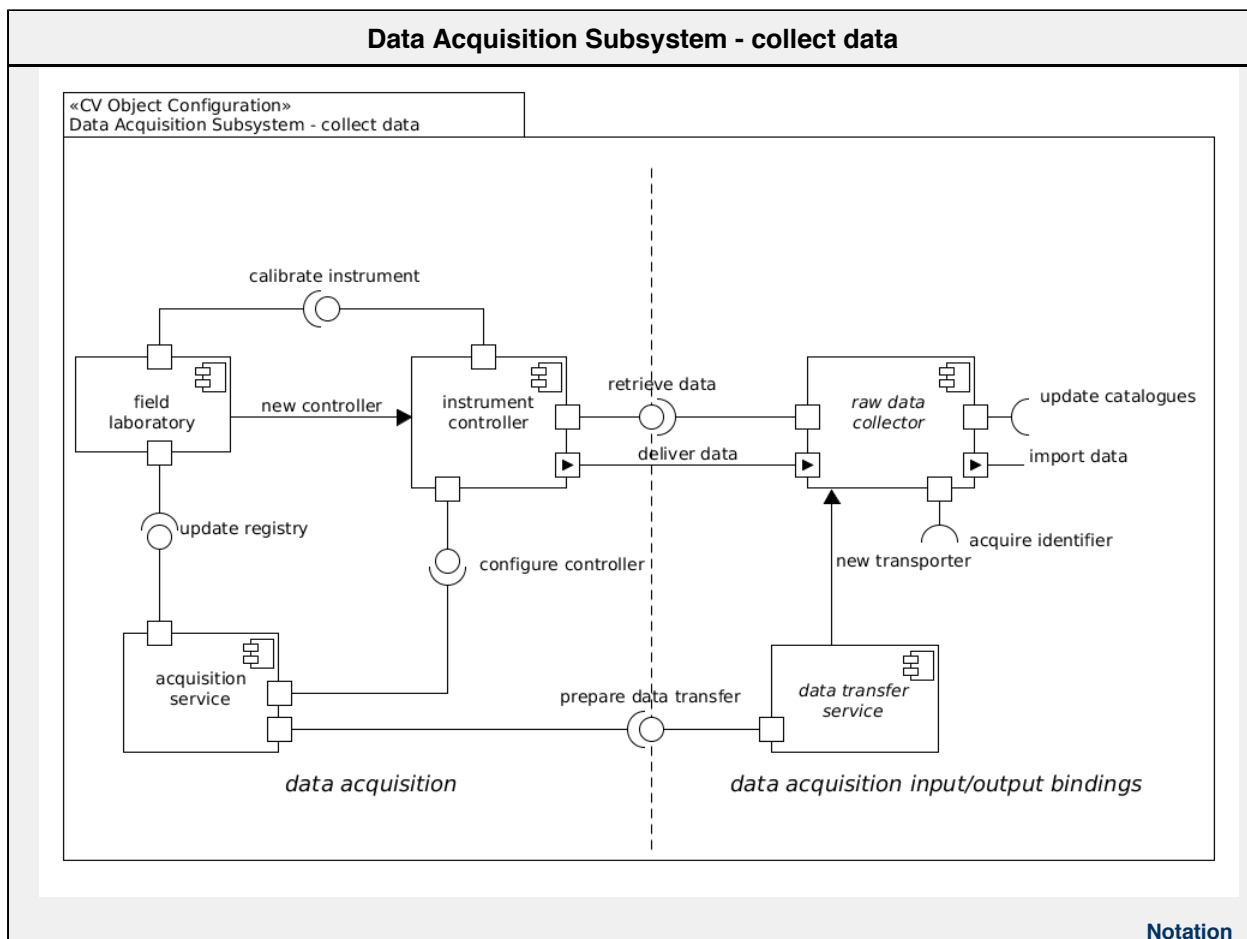
## CV Data Acquisition

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. The following examples present the acquisition of data from instruments and from external data sources.

### Data acquisition from sensors

The diagram shows the organisation of five CV objects as part of an RI which are used for collecting data from an instrument. The instrument controller could be a simple device collecting data from a single sensor or a complex device managing the collection of data for a sensor network.



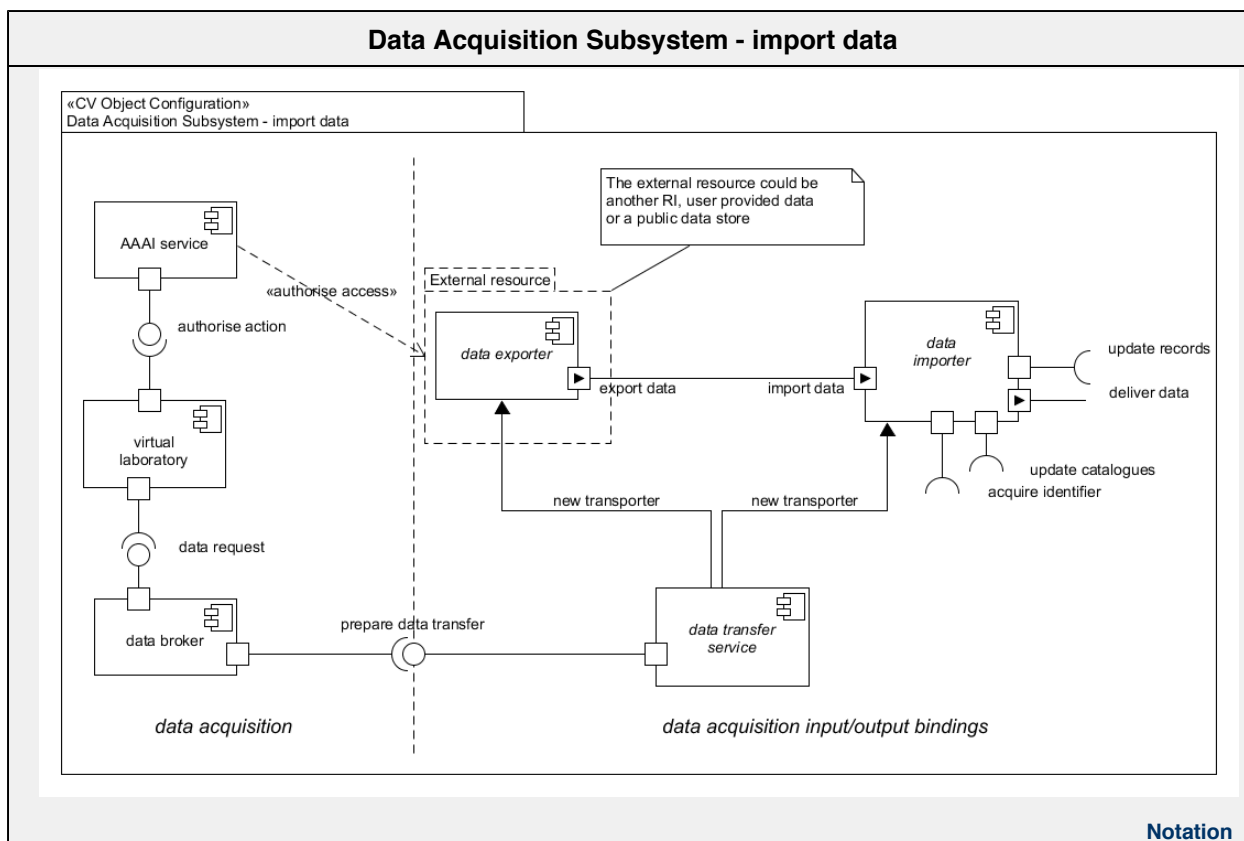
Acquisition is manipulated via **field laboratories**, community proxies by which authorised agents can add and remove instruments from the network (by registering and de-registering instrument controllers) as well as calibrate instrument readings where applicable in accordance with current community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

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 services** invoke **data transfer services** which instantiate the appropriate **raw data collector** which retrieves data from the instrument controller. The four unlinked interfaces of the raw data collector will be linked to appropriate objects of the **data curation** subsystem.

### Data acquisition from external resources

The diagram shows the organisation of six CV objects which are used for collecting data from an external resource. The external resource could be another RI, user uploaded data, or a public data store. The external resource could also be an interface for user observations provided by the RI, for instance for citizen observers



The six components used to model data acquisition from external resources. The external resource is a data source, not necessarily integrated into the infrastructure, providing data to data stores.

Acquisition is manipulated via a **virtual laboratory**, a community proxy, by which authorised agents can submit data to the RI. The **virtual laboratory** invokes a **AAAI service** to retrieve the appropriate credentials for accessing the external resource's **data exporter** and the internal **data importer**. After obtaining the credentials, the **virtual laboratory** invokes a **data broker** which in turn contacts a **data transfer services** which instantiates the appropriate **data exporter** and **data importer** objects and coordinates the transfer of data.

The four unlinked interfaces of the **data importer** will be linked to appropriate objects of the **data curation** subsystem.

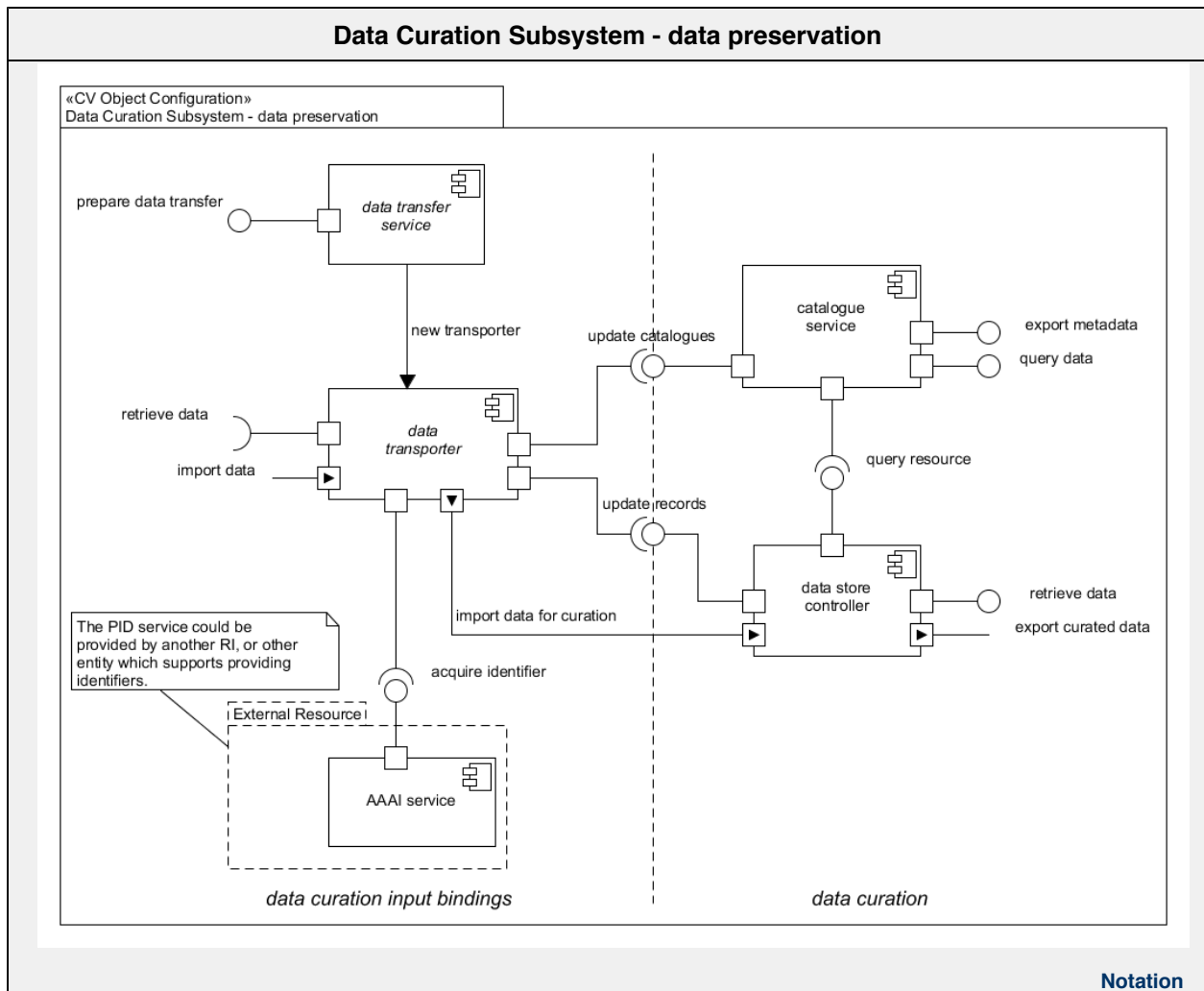
## CV Data Curation

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 phase of the data lifecycle, 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), publishing of persistent identifiers (to aid discovery) and cataloguing (aiding discovery and allowing more sophisticated requests to be made over the entirety of curated data). The following examples present two of the main functionalities of the data curation subsystem: data preservation and data annotation.

### Data Preservation

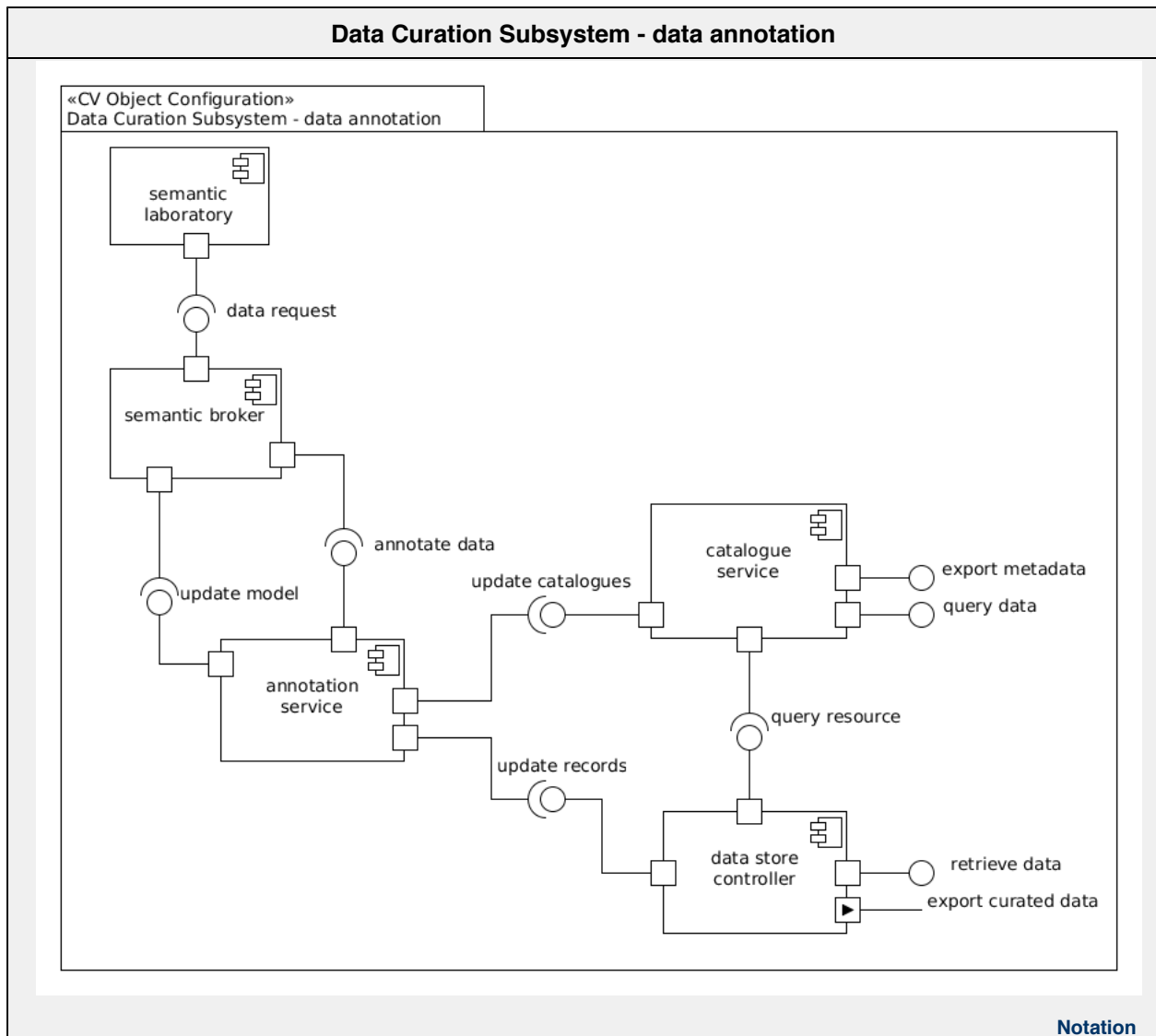
The diagram shows the organisation of five CV objects which participate in the preservation of research data. The **data transporter** in the diagram could be replaced by a **raw data collector** or a **data importer** object, and the change would not affect the integrity of the system. Consequently, this configuration supports both types of data acquisition described in the data acquisition subsystem section. In the example there is no **presentation object** which implies that the data preservation process is automated.

The **data transporter** modelled in the diagram is a complex device which at the same time invokes the **PID service**, the **catalogue service**, and the **data store controller**. The **PID service** is invoked to acquire a unique identifier for the incoming data set. The **catalogue service** is invoked to store the metadata associated with the incoming data set. The **data store controller** is invoked to store the incoming data set, along with its persistent identifier and linked to its associated metadata.



### Data annotation

The diagram shows the organisation of five CV objects which participate in the annotation of research data. This task is carried with the oversight of a user or on request from a user, this is why the presentation object **semantic laboratory** is included. The **semantic laboratory** invokes a **semantic broker** which in turn invokes the **annotation service**. The **annotation service** provides two functionalities annotation and updating of the conceptual model, both the annotation and conceptual model are special types of metadata which are stored in the RI's catalogues and linked to a specific dataset, for this the annotation service invokes the **catalogue service** and the **data store controller**.



## CV Data Publishing

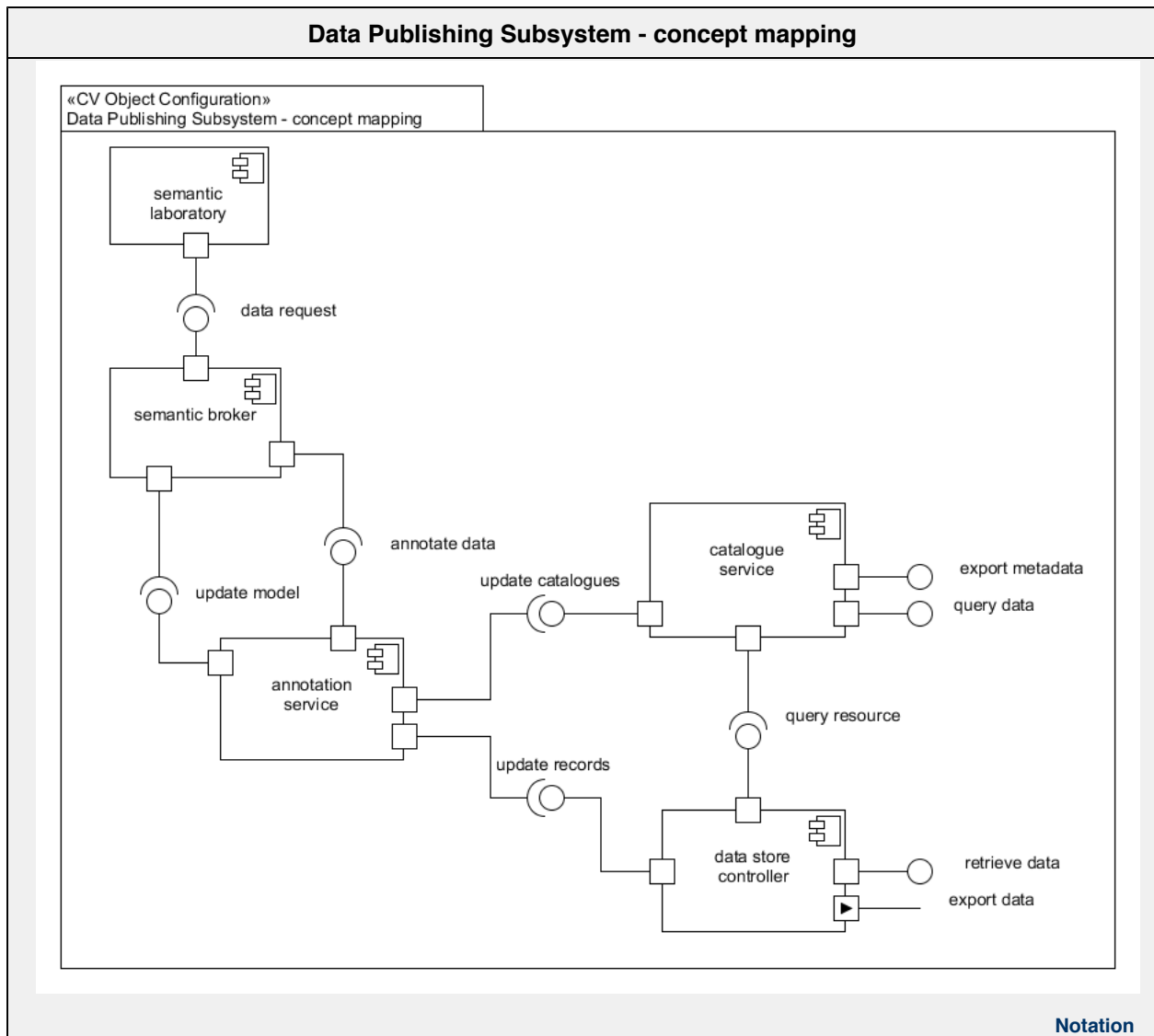
Aside from the curation of scientific data, a research infrastructure must provide means to access that data. Access can be provided in a number of ways, including the export of curated datasets and the querying of data catalogues. Beyond the actual mechanism of access however are the issues of discovery 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). Additionally, a functionality to allow identifying the location of specific datasets in data stores should exist. 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.

The data publishing objects provide **broker objects** which mediate between data stores and catalogues and presentation objects (**virtual laboratories**). **Data brokers** act as intermediaries for access to data held within the data store objects supporting **data curation**. **Semantic brokers** enable semantic interpretation. 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.

The following examples present two important groups of functionalities provided by the data publishing subsystem: concept mapping and data publishing.

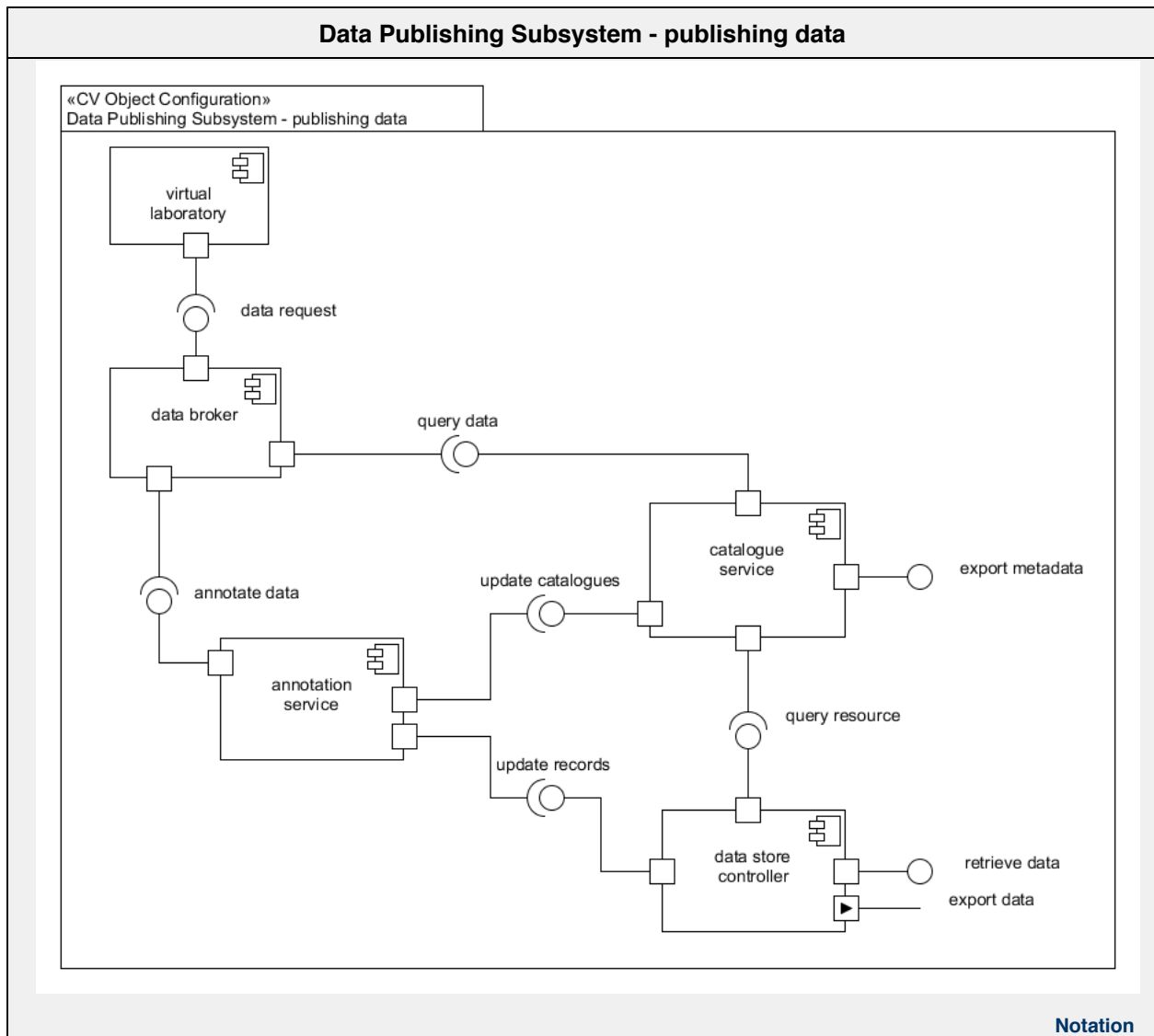
### Concept Mapping

The **semantic laboratory** facilitates three activities which support linking data and metadata to one or more global models: (1) Build Global Conceptual Model, (2) Setup Mapping Rule, and (3) Perform Mapping. The **semantic broker** will facilitate updating the data and internal concept model to preserve the mappings by invoking the **catalogue service** and the **data store controller**.



### **Publishing Data and Metadata**

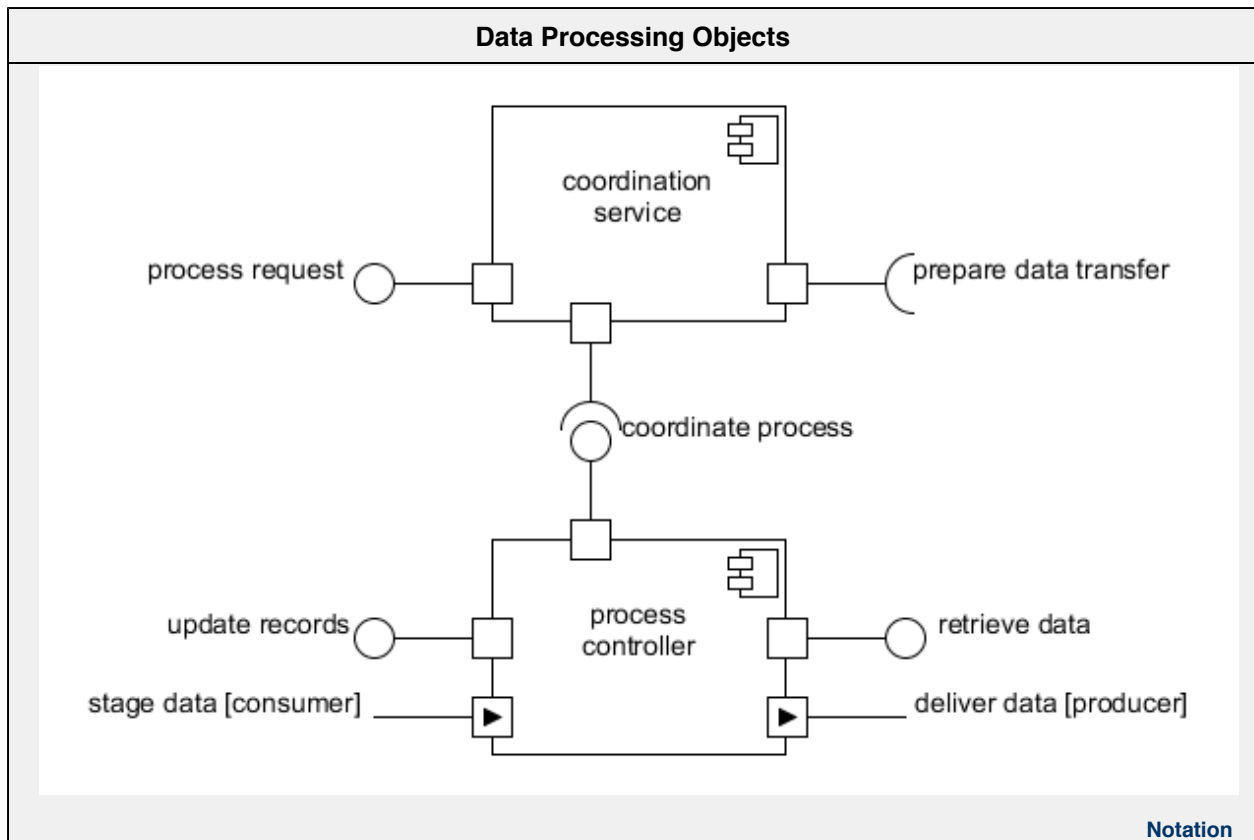
The **virtual laboratory** facilitates the actives which support publishing data and metadata: (1) reviewing the data to be published, (2) publishing data, and (3) publishing metadata. The **data broker** will facilitate updating the data and metadata catalogues to stablish that the data has been reviewed, as well as storing the new links that make data and metadata publicly accessible. These actions are performed by invoking the appropriate **catalogue services** and the **data store controller**.



## CV 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. In general, the more complicated processing tasks will require the use of separated services. The provision of dedicated processing services becomes significantly more important when large quantities of data are being curated within a research infrastructure. Scientific data is an example which is often subject to extensive post-processing and analysis in order to extract new results. The data processing objects of an infrastructure encapsulate the dedicated processing services made available to that infrastructure, either within the infrastructure itself or delegated to a client infrastructure.





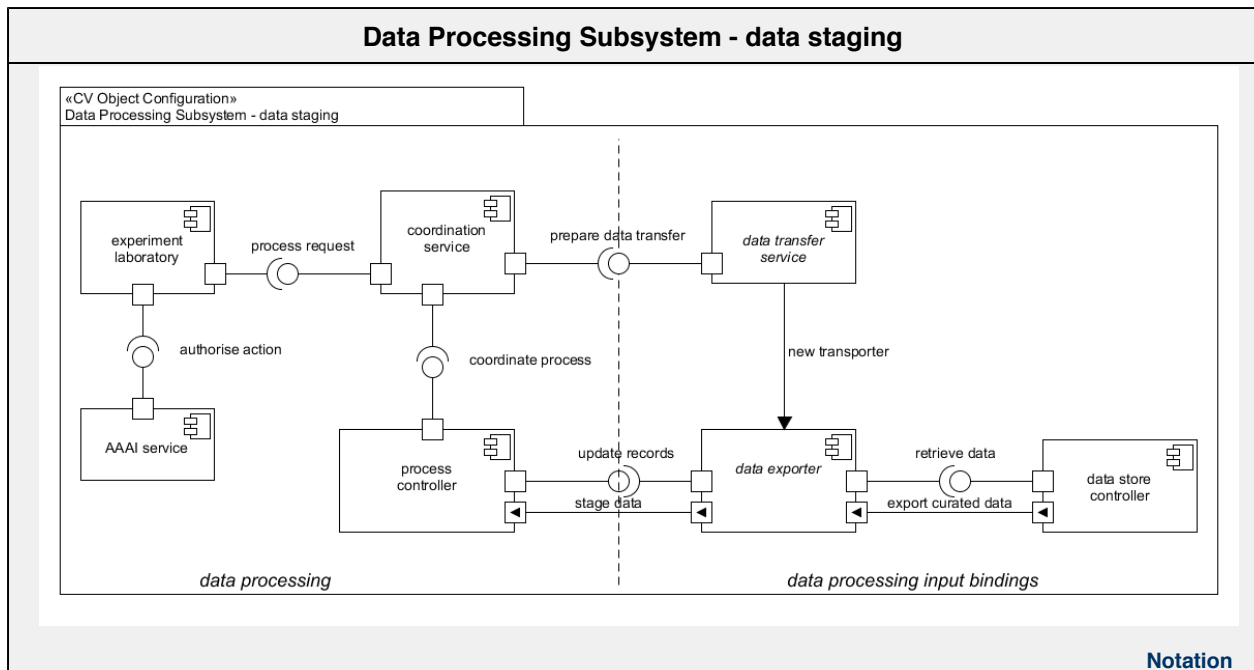
**CV data processing** objects are described as a set of **process controllers** (representing the computational functionality of registered execution resources) monitored and managed by a **coordination service**. The coordination service delegates all processing tasks sent to particular execution resources, coordinates multi-stage workflows and initiates execution. Data may need to be **staged** onto individual execution resources and results **persisted** for future use; data channels can be established with resources via their process controllers. The following diagrams shows the staging and persistence of data.

### Data Staging

The internal staging of data within an infrastructure for processing requires coordination between data processing components (which handle the actual processing workflow) and data curation components (which hold data within the infrastructure). The diagram below displays these two groups of objects which integrate part of the processing subsystem.

Data processing requests generally originate from **experiment laboratories** which validate requests by invoking an **AAAI service**. The **experiment laboratory** will send a process request to a **coordination service**, which interprets the request and starts a processing workflow by invoking the required **process controller**. Data will be retrieved from the data store and passed to the execution platform, the **coordination service** will request that a **data transfer service** to prepare a data transfer.

Data will be retrieved from the data store and passed to the execution platform, the **coordination service** will request that a **data transfer service** to prepare a data transfer. The **data transfer service** will then configure and deploy a **data exporter** which will handle the transfer of data between the storage and execution platforms, i.e. performing data staging. A data-flow is established between all required **data store controllers** and **process controllers** via the **data exporter**. After the data-flow is established, processing starts. Processing can include a host of activities such as summarising, mining, charting, mapping, amongst many others. The details are left open to allow the modelling of any processing procedure. The expected output of the processing activities is a derived data product, which in turn will need to be persisted into the RIs data stores.

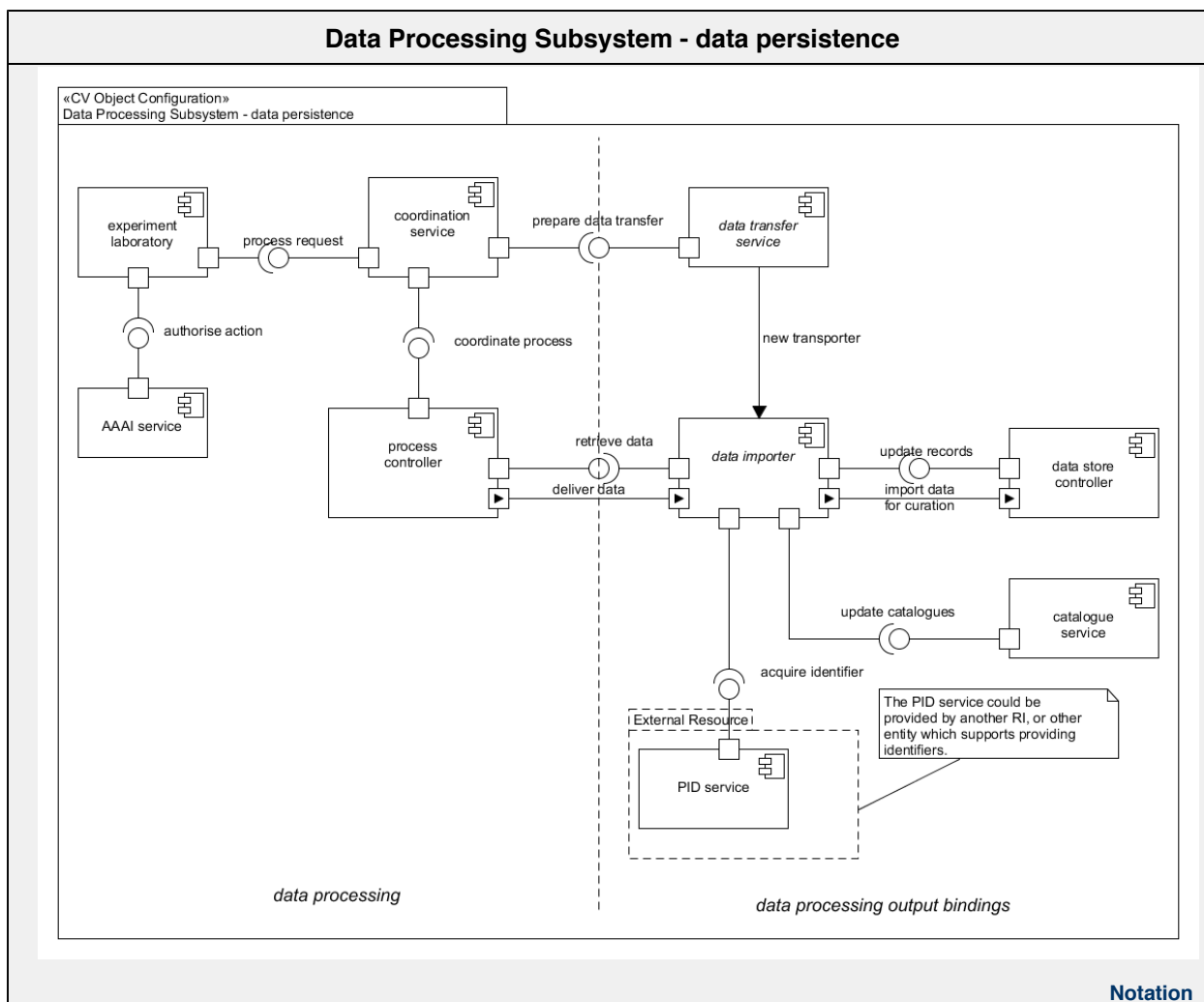


### Data Persistence

The persistence of derived data products produced after processing of data within an infrastructure also requires coordination between data processing components (which handle the actual processing workflow) and data curation components (which hold data within the infrastructure). The diagram below displays these two groups of objects which integrate part of the processing subsystem.

Data processing requests generally originate from **experiment laboratories** which validate requests by invoking an **AAAI service**. The **experiment laboratory** can present results and ask the user if the results need to be stored, alternatively the user may configure the service to automatically store the resulting data. In either case, after processing, the **experiment laboratory** will send a process request to the **coordination service**, which interprets the request and invokes the **process controller** which will get the result data ready for transfer.

The **data transfer service** will then configure and deploy a **data importer** which will handle the transfer of data between the execution and storage platforms. A data-flow is established between **process controller** and **data store controller** via the **data importer**. After the data-flow is established, the data transfer starts. The persistence of data will trigger various curation activities including data storage, backup, updating of catalogues, requiring identifiers and updating records. These activities can occur automatically or just as signals sent out to warn human users that an action is expected.



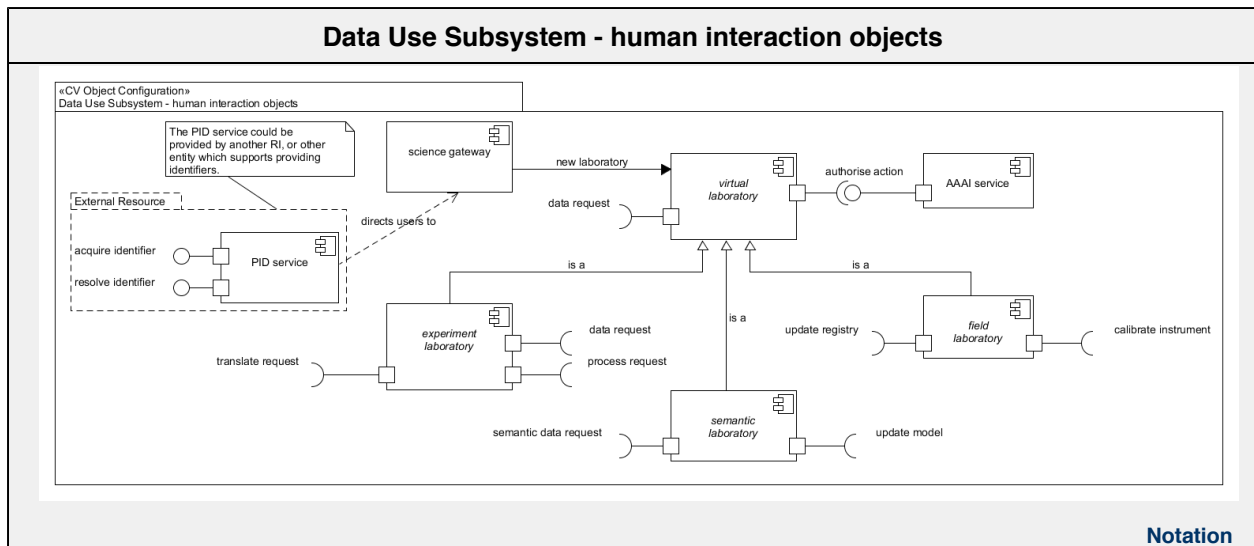
## CV Data Use

A research infrastructure is not an isolated entity, a research infrastructure aims to interact with the broader scientific community. In the ENVRI RM, a **science gateway** (Also known as *virtual research environment*) is assumed to be the main interaction platform for end users (in essence a scientific community portal). The **science gateway** is usually web-based and provides a number of services 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).

The data use components are part of the **presentation** and **service** layers. The **presentation layer** includes different types of human interfaces aimed at providing access to the internal RI resources and services. The **service layer** encapsulates services provided for outside entities that require programmatic interaction with the RI.

In this sense, the data use subsystem can be subdivided in two object categories: **human interaction objects** and **service objects**.

### Human Interaction Objects



In the ENVRI RM, more complex interactions between the components facilitating data use and other components 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 components within the RI. The Reference Model recognises the following specific sub-classes of laboratory:

- **Field laboratories** (so-named because they interact with raw data sources 'in the field') are used to interact with the **data acquisition** components, 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** from the data acquisition set.
- **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.

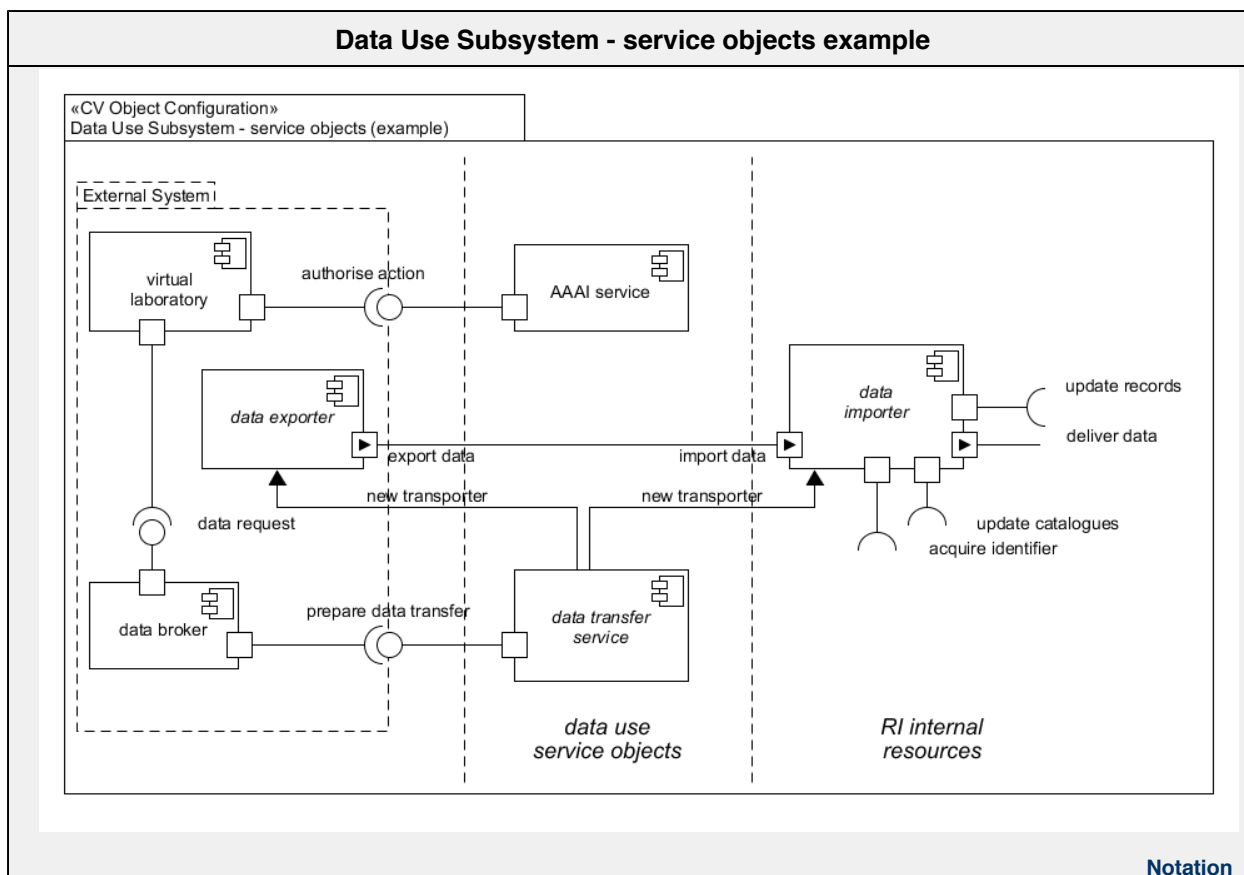
Regardless of provenance, all laboratories must interact with an **AAA service** in order to authorise requests and authenticate users of the laboratory before they can proceed with any privileged activities.

A **PID service** provides globally-readable persistent identifiers (PIDs) to infrastructure entities, mainly datasets, that may be cited by the community. PIDs can also be assigned to processes, services and data sources. This service is assumed 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.

### Service Objects

A constantly increasing portion of the interactions with an RIs are expected to be carried out by external systems interacting with data and other resources. In this case, the **service layer** becomes relevant, services are meant to provide access to external systems. In this case, external systems can include other RIs, universities, government agencies, industry applications, or other research groups which need to exploit the RIs data resources using client programs and the internet as a means to get to those data resources. In this form of integration, external systems are expected to implement **presentation** and **broker** objects which communicate with the RI services using public interfaces.

The following diagram shows an example of the use of service objects to connect an external system which will supply data to an RI. The components of the diagram are the same of those used internally for **data acquisition**, the difference is that the **virtual laboratory**, **data broker**, and **data exporter** objects are all part of an external system. These components interact with the **AAA** and **data transfer** services. The **AAA service** will authorise the requested action and provide the required credentials. The **data transfer service** will establish the data interchange channel between the external **data exporter** and the internal **data importer** objects.



## CV Integration points

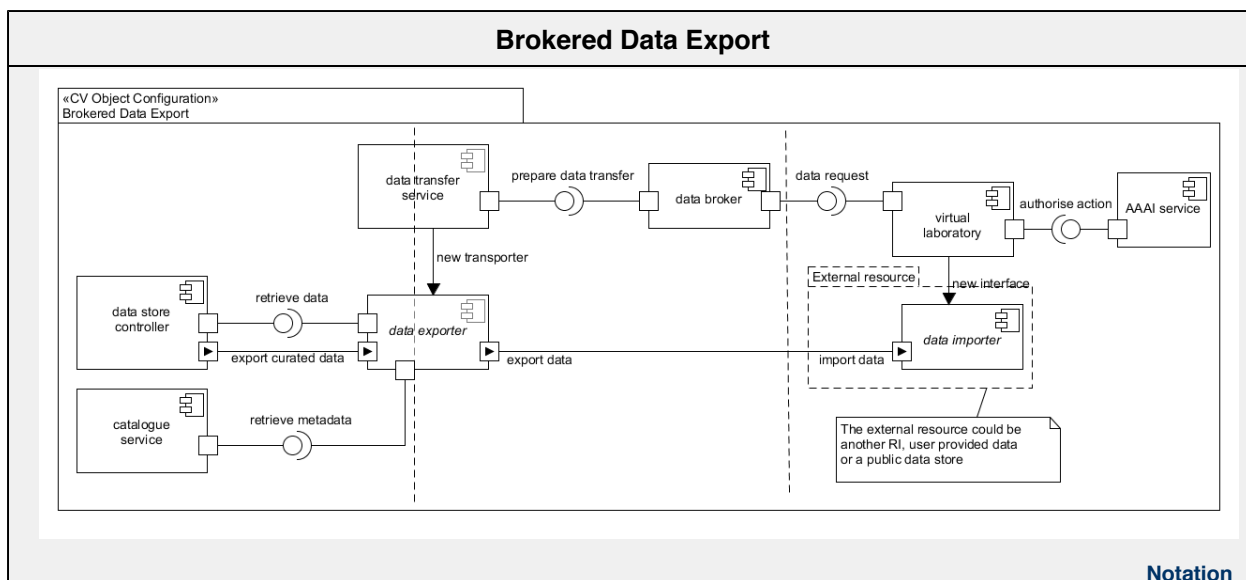
The CV defines the interfaces that support mutual invocation of CV objects functionality, allowing the composition objects to support complex interactions. Examination of these interfaces permits a set of possible bindings to be derived; for each of these bindings, the interaction between the bound objects can be specified in order to define the objects' behaviour when such a binding occurs. This then serves as a basis by which to synthesise the computational behaviour of the entire RI under different use-cases. The CV describes these use cases in detail by providing six integration models. These interactions can occur between lifecycle phases provided by a single RI, but also allow integration of components provided by third parties. The interactions define compound bindings between objects that allow the movement of scientific dataset between different parts of a research infrastructure.

- [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)
- [Raw data collection](#) (the acquisition of raw data from integrated data sources)

The aggregation of these core interactions form a minimal computational model for environmental science research infrastructures that can be used as a starting point for modelling real infrastructures.

## CV 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 data use and data publishing subsystems.

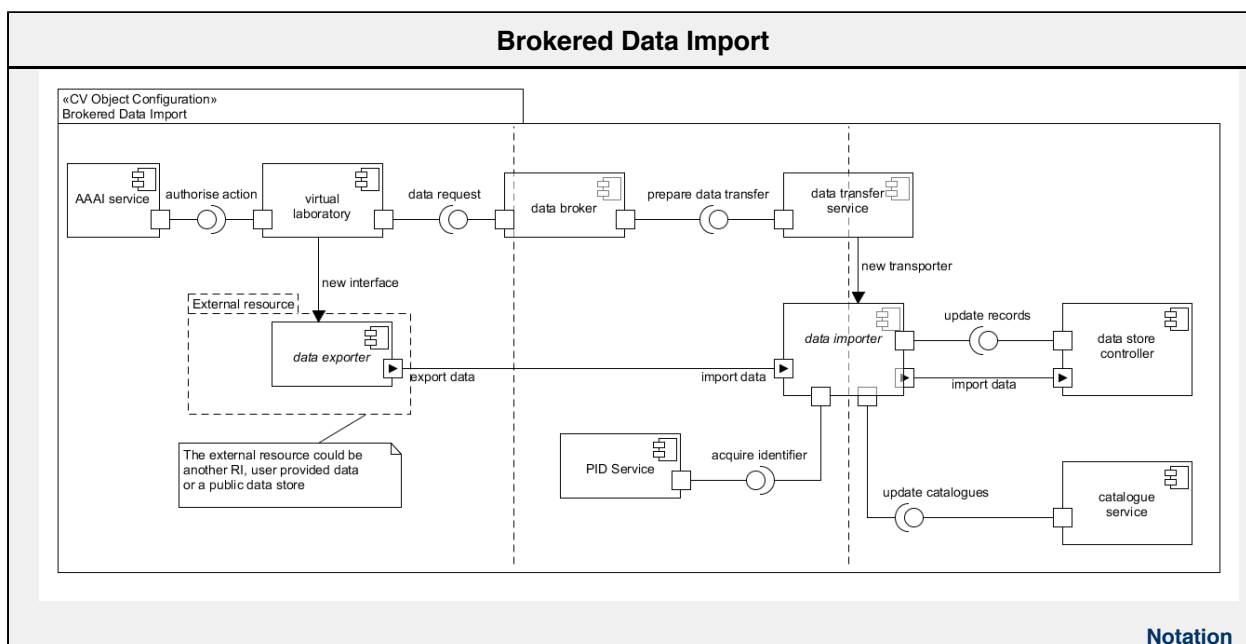


Generally requests for data to be exported to an external resource originate from a **virtual laboratory**. All requests are validated by the **AAAI service** via its *authorise action* interface. The laboratory provides an interface to an external resource (this might take the form of a URI and a preferred data transfer protocol) and submits a request to a data broker in the data publishing 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.

## CV Brokered Data Import

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



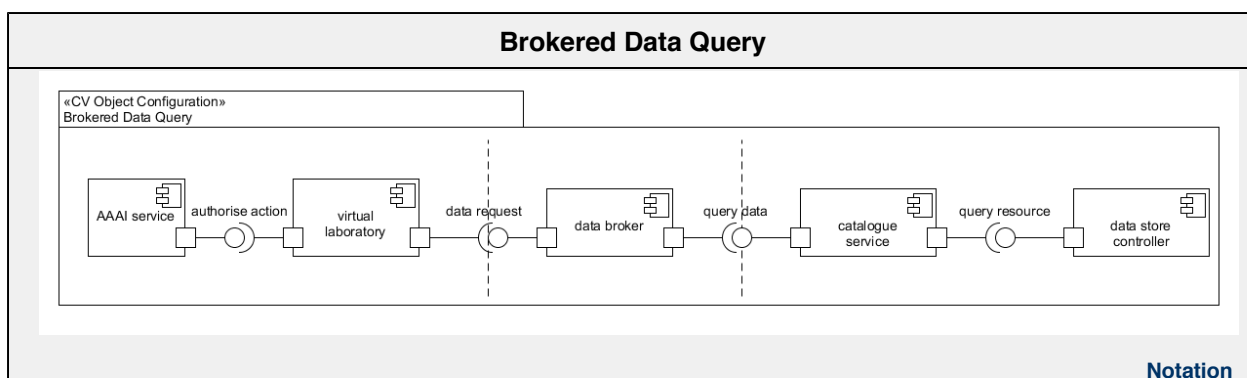
A **virtual laboratory** can be used by researchers to upload new data into a research infrastructure. All requests are validated by the **AAAI service** via its *authorise action* interface. The laboratory provides an interface to an external resource (this might take the form of a URI and

a preferred data transfer protocol) and submits a request to a **data broker** in the data publishing 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** with in the data curation subsystem.

The data transfer service will configure and deploy a **data importer**, the 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*.

## CV Brokered Data Query

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

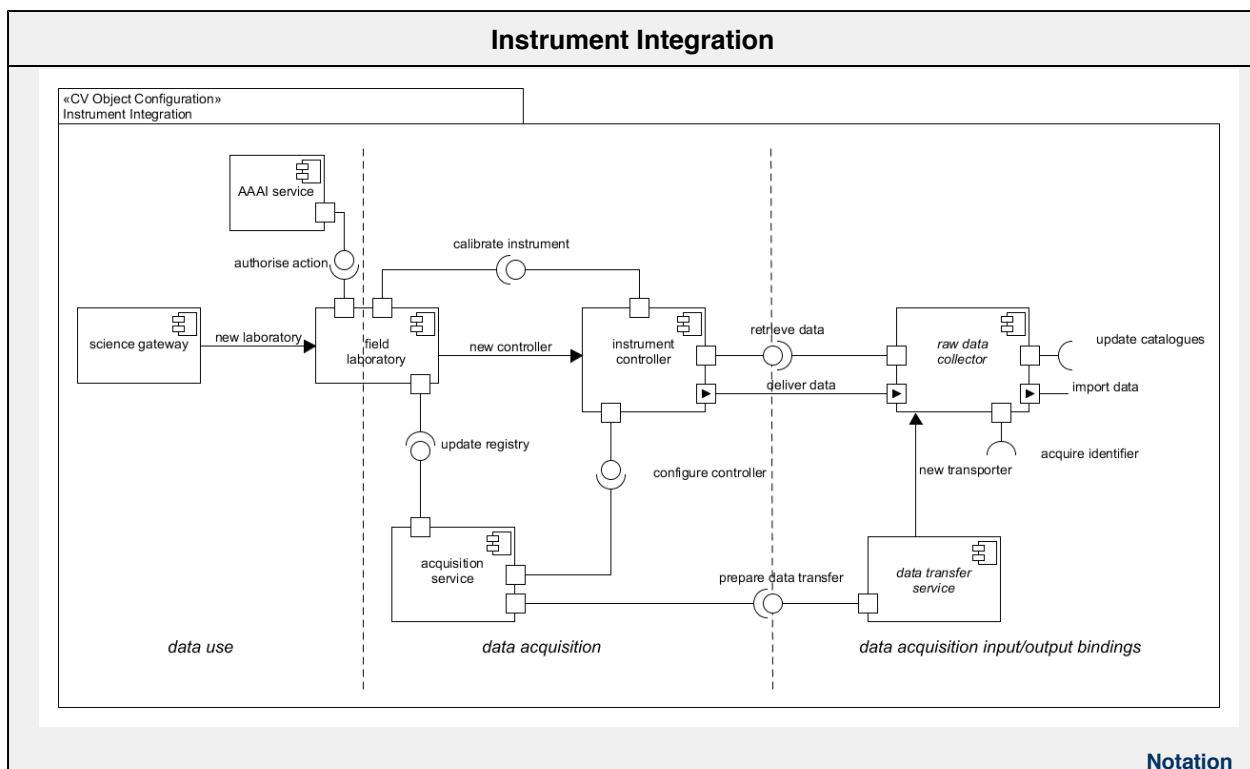


Any kind of **virtual laboratory** is able to query the data held within a research infrastructure subject to access privileges governed by the **AAAI service** (invoked via its *authorise action* interface). Data requests are forwarded to a **data broker** within the data publishing subsystem, which 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**.

## CV 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.





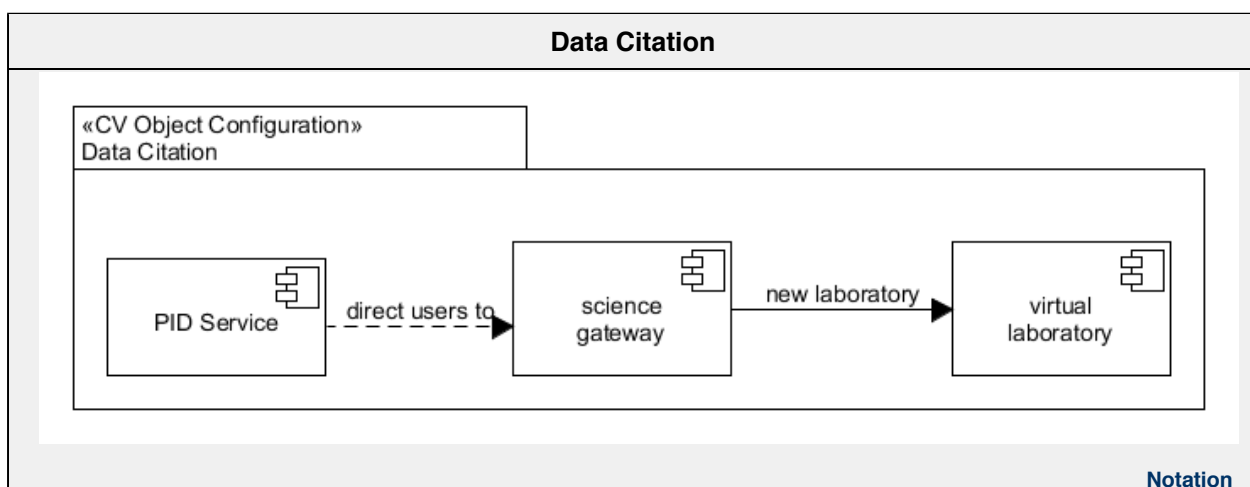
Instruments can be added to and removed from a data acquisition network by a **field laboratory** provided by a **science gateway**. The field laboratory must be able to provide an **instrument controller** for any new instrument added in order to allow the data acquisition subsystem to interact with the instrument. Deployment, un-deployment or re-calibration of instruments requires authorisation - this can only be provided a valid **AAAI 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.

### CV 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.

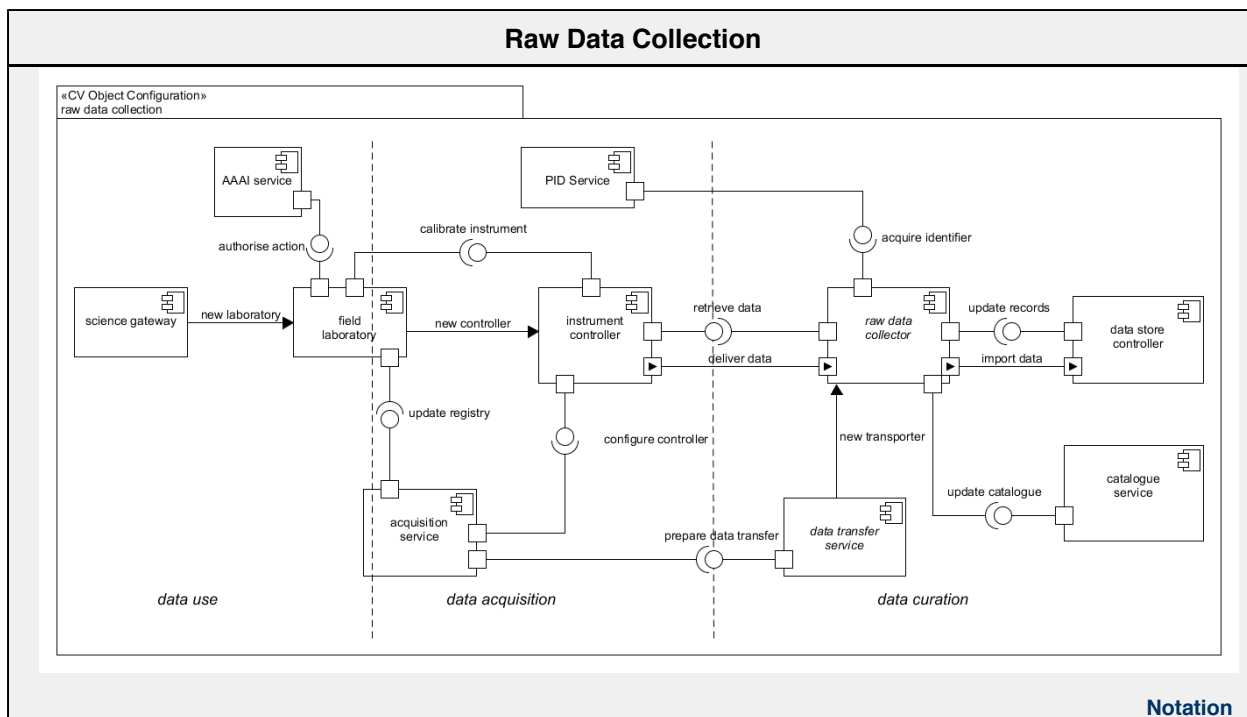


A user or external service tries to *resolve an identifier* (found in a citation) with the global **PID service** used by the research infrastructure. By

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.

## CV Raw Data Collection

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



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 an **AAAI service** authorisation, via the **AAAI 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.

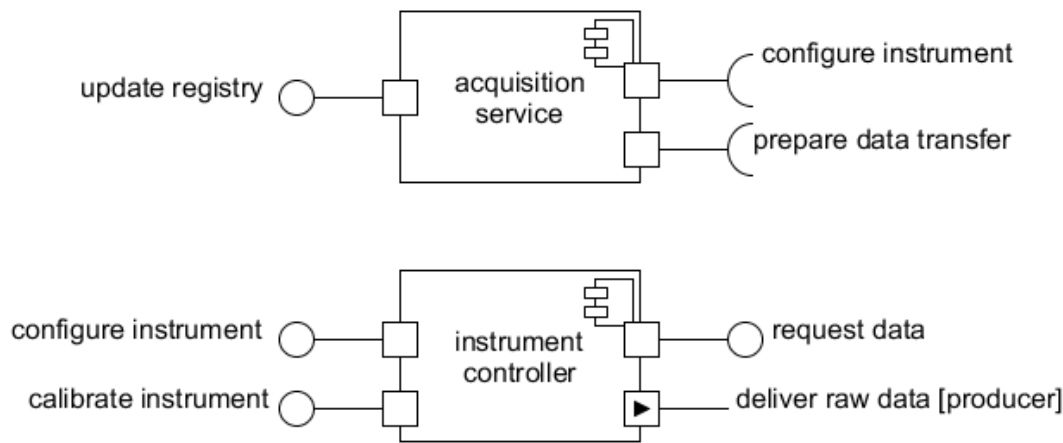
The raw data collector is considered responsible for packaging any raw data obtained into a format suitable for curation - this may entail chunking data streams, assigning persistent identifiers and associating metadata to the resulting datasets. To assist in this, a raw data collector may acquire identifiers 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**.

## How to read the Model (Computational Viewpoint)

The computational viewpoint (CV) is concerned with the modelling of computational objects and the interactions between their interfaces, according to the ODP specification [37]. The ENVRI RM uses a lightweight subset of the full ODP specification to model the abstract computational requirements of an archetypical environmental science research infrastructure.

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 CV is the *computational object*:



In diagrams, each a computational object is represented using a rectangle with a decoration on the upper right corner.. The text within the object indicates the name of the object. The decoration on the upper right corner is standard UML notation for component.

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:

- **Operational 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* interface (providing access to functions that can be invoked by other objects) or a *client* interface (providing a means by which an object operations can be invoked on other objects).

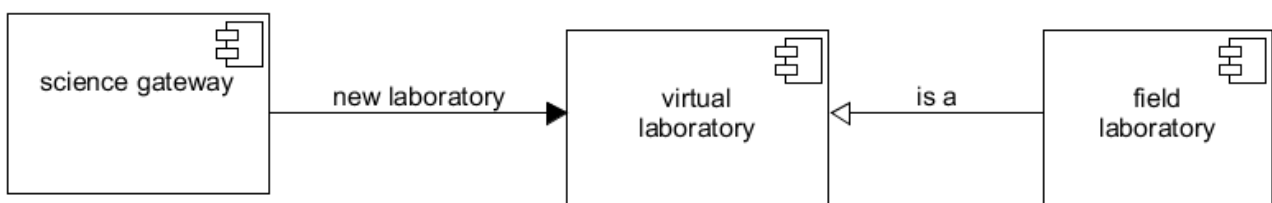
In diagrams, client and server interfaces are linked using 'ball and socket' notation: clients expose sockets (half-circles) whilst servers expose balls (complete circles).

- **Stream interfaces** are used to deliver datasets 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 all required consumers are available to receive that data (whether one, all or some of the consumers must be available depends on the circumstances of the data transfer). Data channels are typically established by operations invoked via operational interfaces (which typically negotiate the terms of the transfer), but can persist independently of them (which is useful for long-term continuous transfers such as from sensor networks to data stores).

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

The decoration on the port boxes is not standard UML but is used to distinguish streaming interfaces.

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.



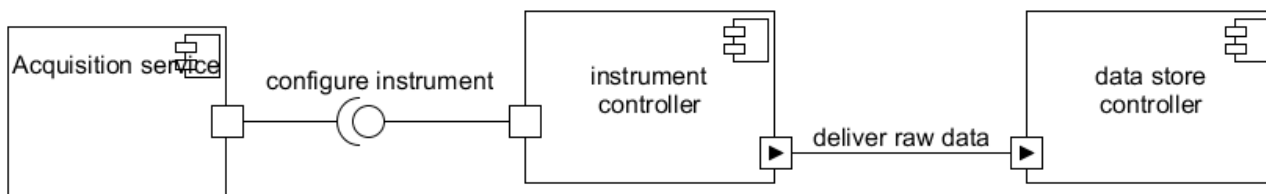
Some objects extend the functionality of other objects; these objects possess all the interfaces of the parent (usually in addition to some of their own) and can be created by the same source object if the capability exists.

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'.

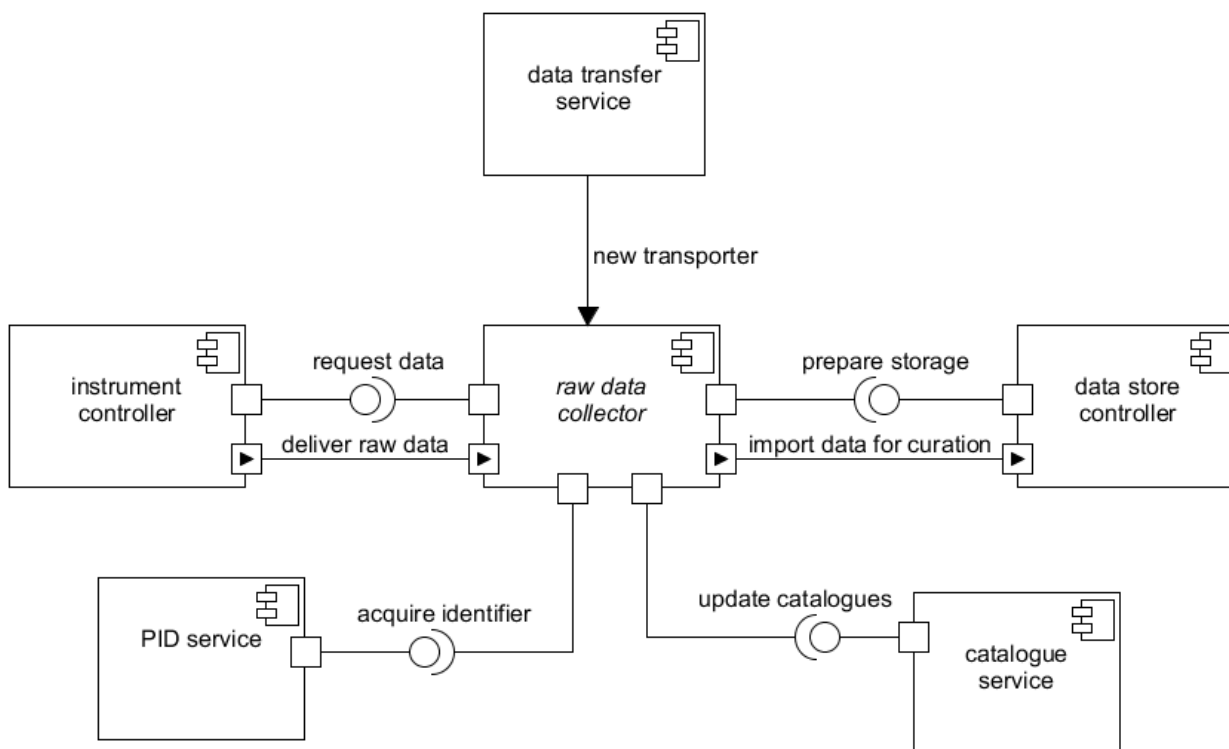
Each interface on a computational object supports a certain type of interaction between objects, which determine the bindings that can be made between interfaces. A *binding* is simply an established connection between two or more interfaces in order to support a specific interaction between two or more computational objects. A client operational interface can be bound to any server operational interface that provides access to the functions that the client requires. Likewise a producer stream interface can be bound to any consumer stream interface that can consume the data produced by the former.

For simplicity, client and server interfaces designed to work together in the Model share the same name; thus a client interface *x* can bind to any server interface *x* and a producer interface *y* can bind to any consumer interface *y*. When a binding is explicitly shown in a diagram, 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 Model is targeted, given that the specific distribution of control between interacting objects is often idiosyncratic to different infrastructure architectures.

The names of binding objects are typically *italicised* in diagrams to better distinguish them from 'basic' computational objects.

## A note about implementation

In principle, all computational objects and their interfaces can be implemented as services or agents within a service-oriented architecture –

this is not required however. Certain objects may be implemented by working groups or even individuals within the infrastructure organisation, bindings between their interfaces implemented by physical interactions, or otherwise human-oriented processes (such as sending data via email).

For example, in the Model, a *field laboratory* has the ability to calibrate instruments (represented by *instrument controllers*) via a binding of their common *calibrate instrument* interfaces. Potentially, the field laboratory could be implemented by a virtual research environment within which authorised users can interact online with instruments deployed in the field, modifying how they acquire data. In practice, the 'field laboratory' may simply abstractly represent the activities of field agents (scientists and technicians) who actually travel to sites where instruments are deployed and manually make adjustments.

This possibility of this kind of 'human-driven' implementation of interactions between computational objects should be accounted for when considering the 'computational' viewpoint of a research infrastructure.

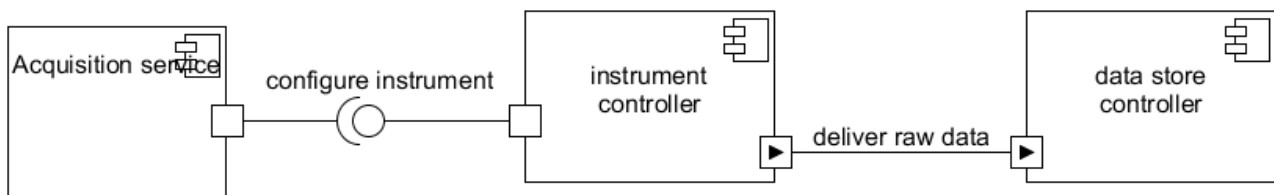
## How to use the Model (Computational Viewpoint)

The computational viewpoint of the Model identifies a standard set of components and interfaces from which can be derived a standard set of interactions that a research infrastructure design should address. The Model does *not* specify how those interactions should be implemented – indeed, over the course of the lifetime of a research infrastructure, implementations may change. Nevertheless, the set of the most important interactions should remain constant regardless of implementation changes.

Someone trying to apply the Computational Viewpoint of the Model to their existing or planned research infrastructure should conduct two primary activities: mapping agents and services to computational objects, and defining the interactions that should occur when two or more interfaces are bound together.

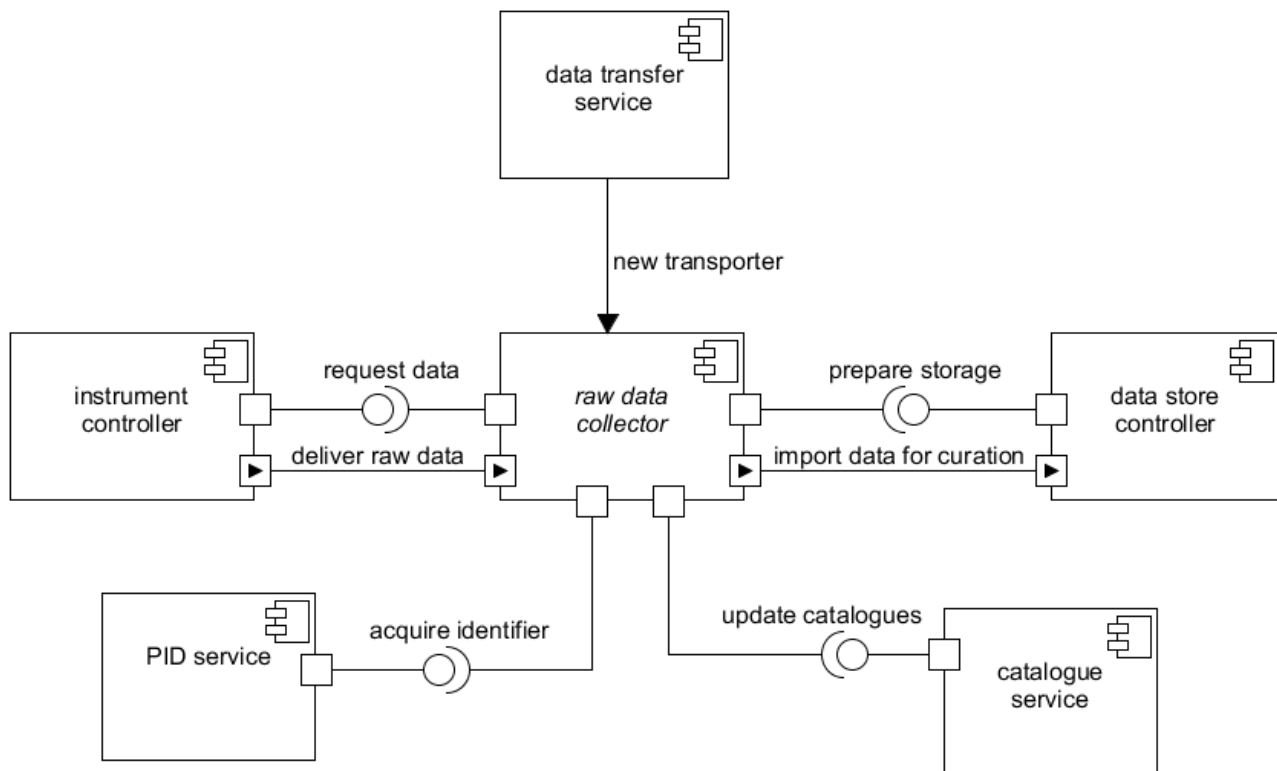
For each computational object in the Model, there should be at least one component or service (or group thereof) provided by the infrastructure that can provide the functions described – depending on the architecture of the infrastructure, there may be multiple candidate, particularly for federated infrastructures. Every such candidate could provide an instantiation of the given object. If no candidates exist, then either (a) the infrastructure does not provide the service embodied by the computational object (and it should be clearly understood that this is indeed the case) or (b) the infrastructure is missing functionality that should be implemented to bring it in compliance with the Model.

For each compatible pair of interfaces (operational or stream), there exists an interaction that should occur given a binding between those two interfaces. The Model does *not* prescribe these interactions, instead simply providing the means to identify them. A compliant research infrastructure should in principle have a well-defined description for *every* possible binding between interfaces on objects that it provides an implementation for.



In the above diagram, an (operational) primitive binding has been established between the *configure instrument* interfaces of an *acquisition service* object and an *instrument controller* object, as well as a (stream) primitive binding between the *deliver raw data* interfaces of the *acquisition service* and a *data store controller* (see [how to read the Model](#) to understand the above notation and terms). Thus, assuming a Model-compliant research infrastructure that provides at least one acquisition service and instrument controller, there should be a specification of what happens when a 'configure instrument' binding occurs between an acquisition service and instrument controller. Likewise, there should be a specification of how raw data is delivered from an instrument (represented by the instrument controller) to a data store (represented by its own controller).

Many *primitive* (two-interface) bindings are linked in that the establishment of one binding will necessarily lead to the establishment of other bindings, implying a unified interaction description. This is particularly true for *compound* bindings where a particular binding object is created to establish pairwise primitive bindings with multiple computational objects that must all contribute to the given interaction. A compliant research infrastructure must therefore identify all such compound bindings and should define how any binding objects created to coordinate interactions are instantiated (generally as either an oversight service or as 'abstractly' as a distributed process involving agents / services participating in the resulting interaction).



In the above diagram, there exist multiple primitive bindings to a central binding object (the *raw data collector*) that nonetheless all relate to a single compound interaction (describing how the transfer of data from an instrument to a data store is configured and managed). It is very important to properly describe the relationship between the individual bindings and how the compound interaction between the various computational objects involved is produced if constructions like in the diagram above are to be properly understood. In the reference material for the Model, a number of 'core' reference interactions have been described informally to provide a [starting point](#) for Model implementors.

Interaction specifications (whether for primitive or compound interface bindings) can take any form deemed suitable by the developers of the infrastructure – for example, UML diagrams such as activity or sequence diagrams may be appropriate, as might be a formal logic model or BPEL workflow, or even natural language if the interaction is simple enough.

## Conclusions 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 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. The RM couples the semantic description of architectures and provides semantic interoperability between model descriptions. 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 model is structured to support the semantic interoperability between data (data objects, metadata and annotations) which is provided by semantic mediation (or mapping, or translation) between descriptions of data (units, parameter, methods and others) and by semantic

mediation of nominal and ordinal values, and/or taxonomies.

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).

## Appendix A Common Requirements of Environmental Research Infrastructures

The following tables describe the common requirements environmental research infrastructures. The requirements are divided in five sets that correspond to the five stages of the datalifecycle. The requirements highlighted on each table are the minimal model.

### Data Acquisition (A)

	Functions	Definitions
A.1	Instrument Integration	Functionality that creates, edits and deletes a sensor.
A.2	Instrument Configuration	Functionality that sets-up a sensor or a sensor network.
A.3	Instrument Calibration	Functionality that controls and records the process of aligning or testing a sensor against dependable standards or specified verification processes.
A.4	Instrument Access	Functionality that reads and/or updates the state of a sensor.
A.5	Configuration Logging	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	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	Functionality that outputs the values of parameters and measured variables a display device.
A.8	<i>(Real-Time) (Parameter/Data) Visualisation</i>	<i>Specialisation of (Parameter) Visualisation which is subject to a real-time constraint.</i>
A.9	Process Control	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.
A.10	Data Collection	Functionality that obtains digital values from a sensor instrument, associating consistent timestamps and necessary metadata.
A.11	<i>(Real-Time) Data Collection</i>	<i>Specialisation of Data Collection which is subject to a real-time constraint.</i>
A.12	Data Sampling	Functionality that selects a subset of individuals from within a statistical population to estimate characteristics of the whole population.
A.13	Noise Reduction	Functionality that removes noise from scientific data.
A.14	Data Transmission	Functionality that transfers data over communication channel using specified network protocols.
A.15	<i>(Real-Time) Data Transmission</i>	<i>Specialisation of Data Transmission which handles data streams using specified real-time transport protocols.</i>
A.16	Data Transmission Monitoring	Functionality that checks and reports the status of data transferring process against specified performance criteria.

### Data Curation (B)



	Functions	Definitions
B.1	Data Quality Checking	Functionality that detects and corrects (or removes) corrupt, inconsistent or inaccurate records from data sets.
B.2	Data Quality Verification	Functionality that supports manual quality checking.
B.3	Data Identification	Functionality that assigns (global) permanent unique identifiers to data products.
B.4	Data Cataloguing	Functionality that associates a data object with one or more metadata objects which contain data descriptions.
B.5	Data Product Generation	Functionality that processes data against requirement specifications and standardised formats and descriptions. (optional/may be null)
B.6	Data Versioning	Functionality that assigns a new version to each state change of data, allows to add and update some metadata descriptions for each version, and allows to select, access or delete a version of data.
B.7	Workflow Enactment	Functionality that 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	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.
B.9	Data Replication	Functionality that creates, deletes and maintains the consistency of copies of a data set on multiple storage devices.
B.10	Replica Synchronisation	Functionality that exports a packet of data from on replica, transports it to one or more other replicas and imports and applies the changes in the packet to an existing replica.

#### Data Publishing (C)

	Functions	Definitions
C.1	Access Control	Functionality that approves or disapproves of access requests based on specified access policies.
C.2	Resources Annotation.	Functionality that creates, changes or deletes a note that reading any form of text, and associates them with a computational object.
C.3	(Data) Annotation	<i>Specialisation of Resource Annotation which allows to associate an annotation to a data object.</i>
C.4	Metadata Harvesting	Functionality that (regularly) collects metadata (in agreed formats) from different sources.
C.5	Resource Registration	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.
C.6	(Metadata) Registration	<i>Specialisation of Resource Registration, which registers a metadata object in a metadata registry.</i>
C.7	(Identifier) Registration	<i>Specialisation of Resource Registration, which registers an identifier object in an identifier registry.</i>
C.8	(Sensor) Registration	<i>Specialisation of Resource Registration which registers a sensor object to a sensor registry.</i>
C.9	Data Conversion	Functionality that converts data from one format to another format.
C.10	Data Compression	Functionality that encodes information using reduced bits by identifying and eliminating statistical redundancy.
C.11	Data Publication	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 publicly accessible or to those who agree to certain conditions of use, and to individuals who meet certain professional criteria.
C.12	Data Citation	Functionality that assigns an accurate, consistent and standardised reference to a data object, which can be cited in scientific publications.
C.13	Semantic Harmonisation	Functionality that unifies 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	Functionality that retrieves requested data from a data resource by using suitable search technology.
C.15	Data Visualisation	Functionality that displays visual representations of data.

#### Data Processing (D)

	Functions	Definitions
D.1	Data Assimilation	Functionality that combines observational data with outputs from a numerical model to produce an optimal estimate of the evolving state of the system.
D.2	Data Analysis	Functionality that inspects, cleans, and transforms data, providing data models which highlight useful information, suggest conclusions, and support decision making.
D.3	Data Mining	Functionality that supports the discovery of patterns in large data sets.
D.4	Data Extraction	Functionality that retrieves 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	Functionality that supports the generation of abstract, conceptual, graphical or mathematical models, and to run an instances of those models.
D.6	<i>(Scientific) Workflow Enactment</i>	<i>Functionality provided as a specialisation of Workflow Enactment supporting the composition and execution of computational or data manipulation steps in a scientific application. Important processing results should be recorded for provenance purposes.</i>
D.7	<i>(Scientific) Visualisation</i>	Functionality that graphically illustrates scientific data to enable scientists to understand, illustrate and gain insight from their data. (optional or may be null)
D.8	Service Naming	Functionality that encapsulates the implemented name policy for service instances in a service network.
D.9	Data Processing Control	Functionality that initiates calculations and manages the outputs to be returned to the client.
D.10	Data Processing Monitoring	Functionality that checks the states of a running service instance.

#### Data Use (E)

	Functions	Definitions
E.1	Authentication	Functionality that verifies a credential of a user.
E.2	Authorisation	Functionality that specifies access rights to resources.
E.3	Accounting	Functionality that measures the resources a user consumes during access for the purpose of capacity and trend analysis, and cost allocation.
E.4	<i>(User) Registration</i>	<i>Specialisation of Resource Registration which registers a user to a user registry.</i>
E.5	Instant Messaging	Functionality for quick transmission of text-based messages from sender to receiver.
E.6	<i>(Interactive) Visualisation</i>	Functionality that enables users to control of some aspects of the visual representations of information.
E.7	Event Notification	Functionality that delivers message triggered by predefined events.

## Appendix B Terminology and Glossary

- [Acronyms and Abbreviations](#)
- [Terminology](#)

### Acronyms and Abbreviations

CCSDS Consultative Committee for Space Data Systems

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

## 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.

**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.

**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 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 Publishing Subsystem:** A subsystem that enables discovery and retrieval of data housed in data resources.

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

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

**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.

**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 **knowledge base** enabled by the research infrastructure.

**Instrument Controller:** An integrated raw data source.

**Knowledge Base:** (1) A store of information or data that is available to draw on. (2) The underlying set of facts, assumptions, and rules which a computer system has available to solve a problem.

**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 persist 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.

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

**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.

**Semantic Laboratory**: Community proxy for interacting with semantic models.

**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.

**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.

**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.

## Appendix C Notation

The notation used for the diagrams of the ENVRI RM is based on the UML notation suggested for ODP, the **UML4ODP** notation. The notation sections include a set of tables that describe the UML elements used to produce the diagrams presenting the different viewpoint models.

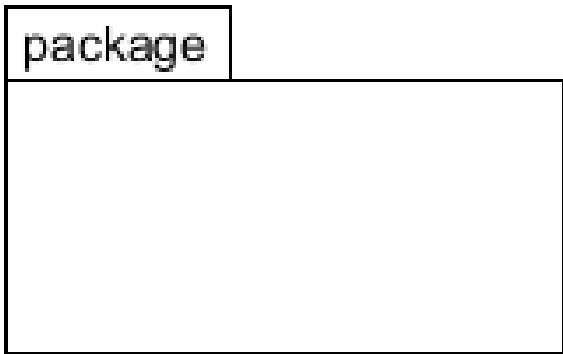
- [Science viewpoint models](#)
- [Information viewpoint models](#)
- [Computational viewpoint models](#)




## Notation of Science Viewpoint Models

### Communities

SV communities are modelled using an object diagram. The following table describes the elements used in that diagram.

Table 1 Notation for community diagrams

Figure	Description
	<p>A Package, in UML notation, is a grouping element. Package is used "to group elements, and to provide a namespace for the grouped elements".</p> <p>A package may contain other packages, thus providing for a hierarchical organization of packages.</p> <p>Classes, objects, use cases, components, nodes, node instances etc. can all be organized as packages, enabling a manageable organization of the elements of UML models.</p>

	<p>Objects are used to represent communities in the RM.</p> <p>The name refers to the represented entity</p> <p>The stereotype indicates the namespace where the object is grouped. Sometimes the stereotype can be an image. The image can be used in place of the figure. For ODP, the stereotype for community is a group of people:</p> 
	<p>A note is used to provide additional information about a diagram.</p> <p>If the note refers to a specific element in the diagram, then it is connected to that object with a simple arc.</p>

In the following example diagram the package represents an Environmental research infrastructure. The infrastructure contains five objects which are all communities. Notes are used to describe the objectives of each of the communities.

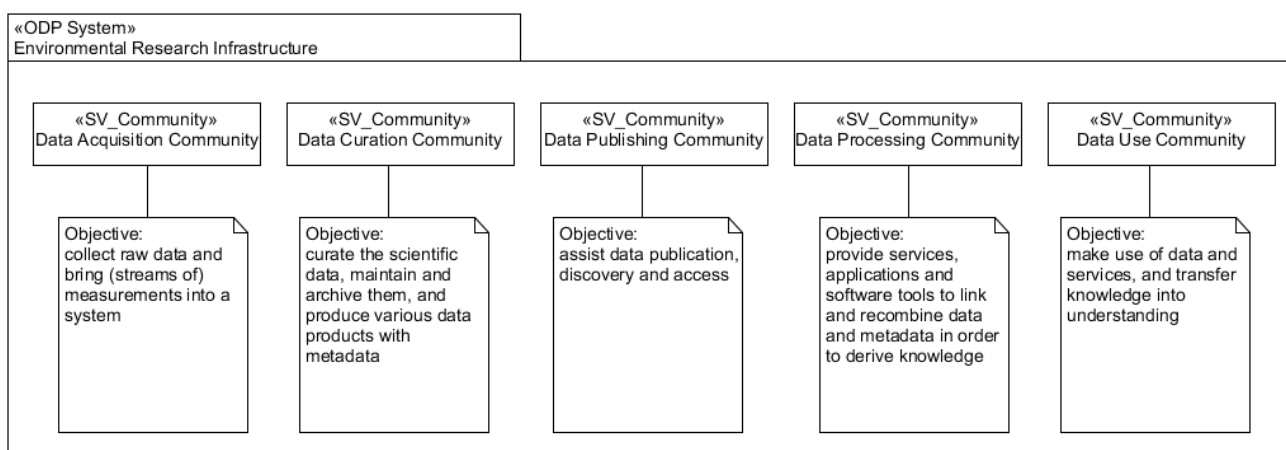


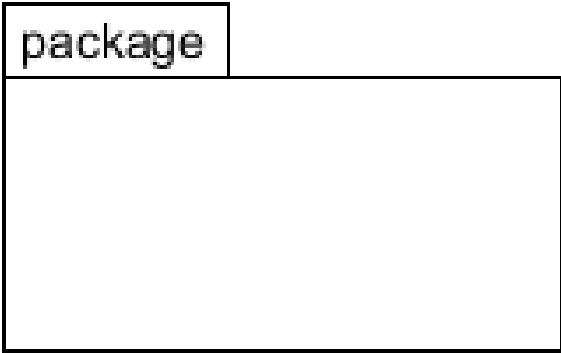
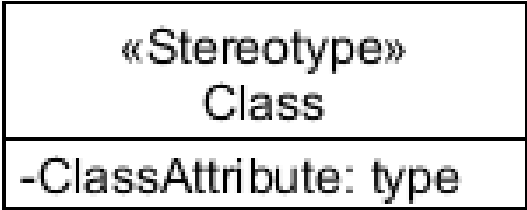
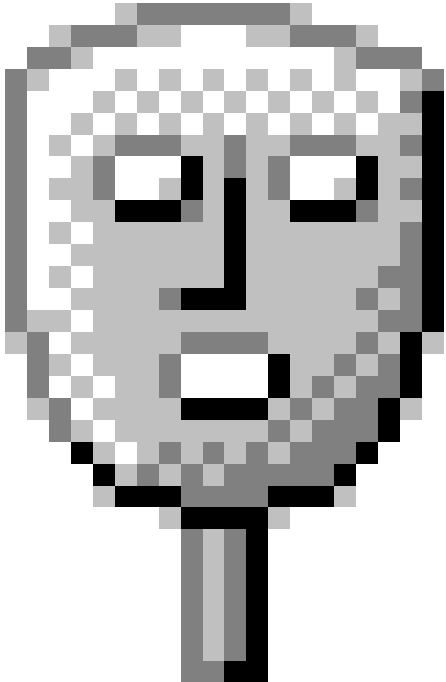
Figure 1 Example of a community diagram

## Community Roles

SV Roles are represented using a class diagram with packages and classes

Table 2 Notation for role diagrams

Figure	Description
--------	-------------

	<p>A Package, in UML notation, is a grouping element. Package i group elements, and to provide a namespace for the grouped</p> <p>A package may contain other packages, thus providing for a h organization of packages.</p> <p>Classes, objects, use cases, components, nodes, node instan all be organized as packages, enabling a manageable organiz elements of UML models.</p>
	<p>Classes are used to represent roles in the RM.</p> <p>Classes can have additional compartments to express proper attributes) and behaviours (called methods). Omitting the com means that the behaviour and attributes are undefined at the t building the diagram.</p> <p>Name tag indicates the name of the class. Typically, classes a using no spaces and starting each word that makes up the nar camelcase.</p> <p>The stereotype indicates the namespace where the class is gr Sometimes the stereotype can be an image. The image can b place of the figure. For ODP, the stereotype for role is a mask:</p> 

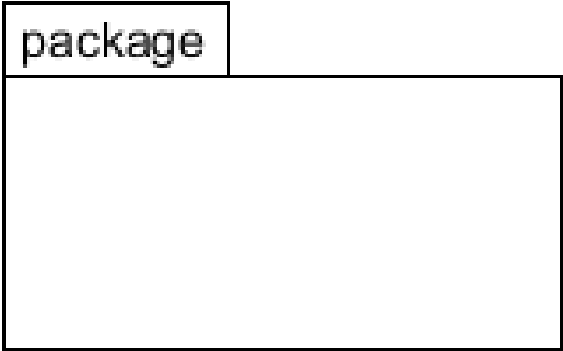
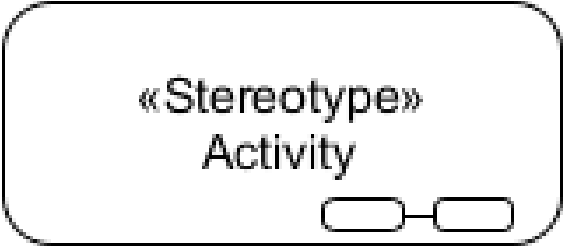
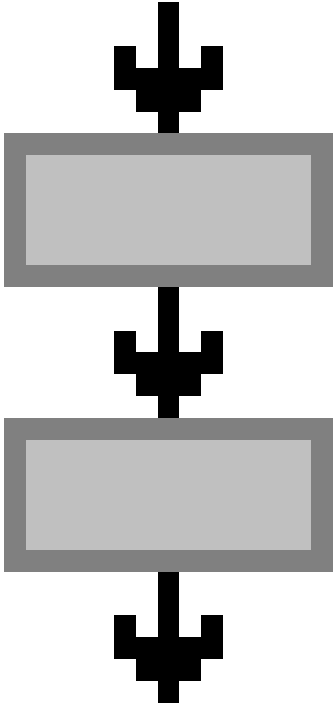
In the example diagram the package represents the data curation community. The community contains eight role classes. The ENVRI RM provides a detailed description of each role in text.

Figure 2 Example of a SV Role diagram

### Community Behaviours

SV Behaviours are represented using an activity diagram with packages and activities

Table 3 Notation for behaviour diagrams

Figure	Description
	<p>A Package, in UML notation, is a grouping element. Package i group elements, and to provide a namespace for the grouped</p> <p>A package may contain other packages, thus providing for a h organization of packages.</p> <p>Classes, objects, use cases, components, nodes, node instan all be organized as packages, enabling a manageable organiz elements of UML models.</p>
	<p>Activities are used to represent behaviours in the RM.</p> <p>Name tag indicates the name of the behaviour. Behaviours are using a short phrase that describes the event or action being r</p> <p>The small decoration in the activity indicates that the activity is and can be subdivided into smaller tasks.</p> <p>A stereotype can be used to indicate the namespace where th grouped. Sometimes the stereotype can be an image. The ste image can be used in place of the figure. For ODP, the stereot behaviour is process icon:</p> 

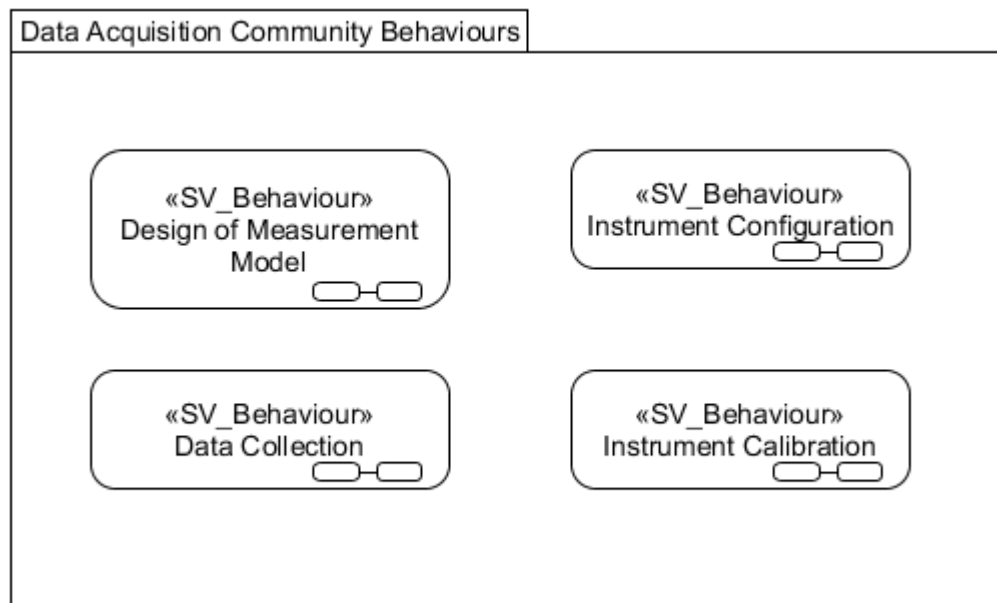


Figure 3 Example of a SV Behaviour diagram

In the example diagram the package represents a community, data acquisition. The community implements four basic behaviours. The RM also provides a detailed description of each behaviour in text.

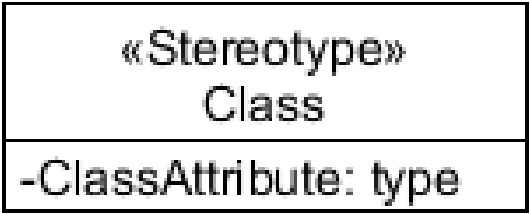
## Notation of Information Viewpoint Models

### Information Objects

IV Objects are represented using a class diagram.

Table 4 Notation for information object diagrams

Figure	Description
	<p>A package, in UML notation, is a grouping element. Package is used to group elements, and to provide a namespace for the grouped elements.</p> <p>A package may contain other packages, thus providing for a hierarchical organization of packages.</p> <p>Classes, objects, use cases, components, nodes, node instances, and relationships are all organized as packages, enabling a manageable organization of UML models.</p>

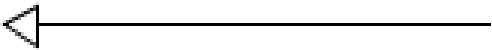


Classes are used to represent information objects in the RM.

Classes can have additional compartments to express properties (attributes) and behaviours (called methods). Leaving the compartments blank means that the behaviour and attributes are undefined at the time of creating the diagram.

Name tag indicates the name of the class. Typically, classes are named using no spaces in camelcase.

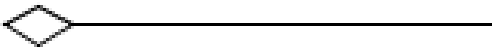
The stereotype indicates the namespace where the class is defined. Sometimes the stereotype can be an image. The image can be placed instead of the figure. For ODP, the stereotype for information object is an icon with a tag on top:



Generalisation relationship indicates that one of the two related classes (the subclass) is considered to be a specialized form of the other class).

Generalisation is represented with an arc with a blank triangle. The blank triangle points to the super class and the undecorated end of the arc is connected to the subclass.

The generalization relationship is also known as the inheritance relationship.



Aggregation relationship indicates an association that represents a part-whole or part-of relationship.

Aggregation is represented with an arc with a blank rhombus. The blank rhombus shape indicates the composite and the undecorated end of the arc is the component.

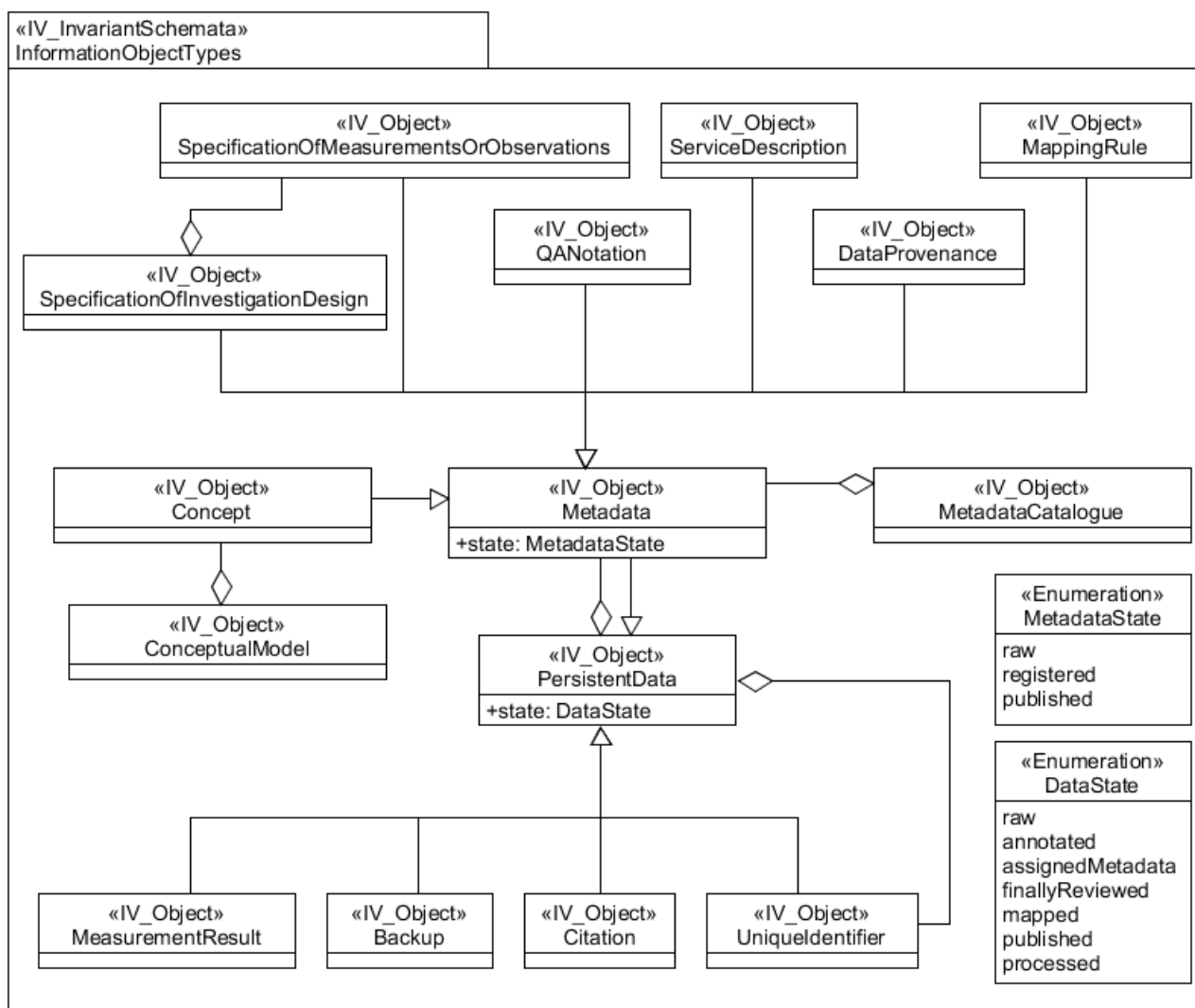


Figure 4 Example of an IV Object diagram

In the example diagram the package represents the collection of all information objects described by the ENVRI RM. The stereotype for the package is invariant schemata, which indicates that these are the parts of the model that are stable. The main objects are persistent data and metadata. The RM also provides a detailed description of each object in the text.

## Information Actions

IV Actions are represented using an activity diagram with packages and activities

Table 5 Notation for action type diagrams

Figure	Description
--------	-------------



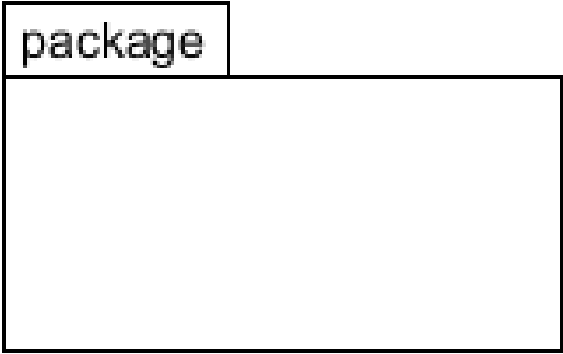
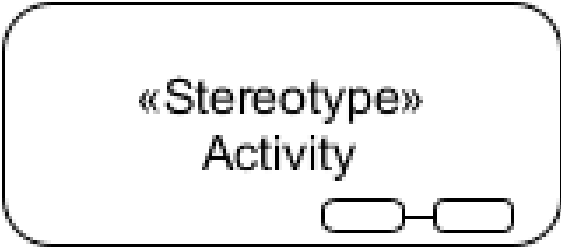

	<p>A package, in UML notation, is a grouping element. Package is used to group elements, and to provide a namespace for the grouped elements.</p> <p>A package may contain other packages, thus providing for a hierarchical organization of packages.</p> <p>Classes, objects, use cases, components, nodes, node instances, and relationships can all be organized as packages, enabling a manageable organization of UML models.</p>
	<p>Activities are used to represent actions in the RM.</p> <p>Name tag indicates the name of the action. Actions are named by a short phrase that describes the event or action being represented.</p> <p>The small decoration in the box indicates that the action is complex and can be subdivided into smaller tasks.</p> <p>A stereotype can be used to indicate the namespace where the action is grouped. Sometimes the stereotype can be an image. The stereotype image can be used in place of the figure. For ODP, the stereotype for information action is an arrow icon with a lowercase "i":</p> 

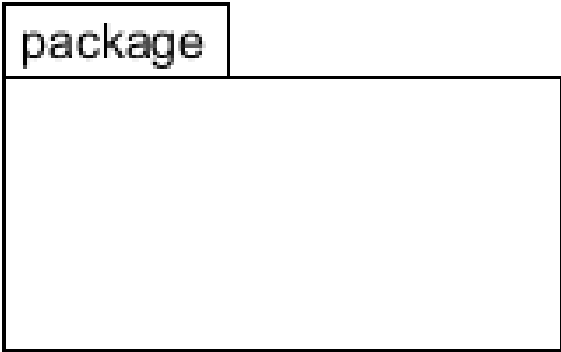
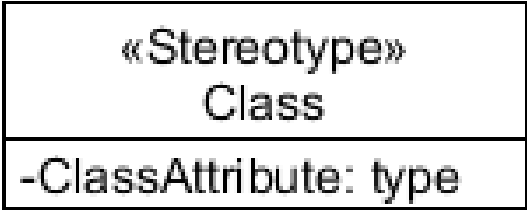

Figure 5 Example of an IV Action Types diagram

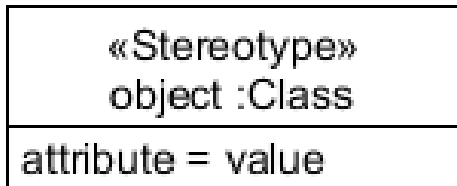
In the example diagram the package represents the information action types described by the ENVRI RM. The stereotype for the package is invariant schemata, which indicates that these are the parts of the model that are stable. The RM also provides a detailed description of each action in text.

## Information Object Instances

IV Objects instances are represented using an object diagram. The type of diagram is similar to the class diagram with the difference that the entities represented are objects not classes. Object instances have a specific state and this can change depending on the moment when the object is observed. Object instances are useful for representing the dynamic nature of the systems.

Table 6 Notation for information object instances diagrams

Figure	Description
	<p>A package, in UML notation, is a grouping element. Package is used to group elements, and to provide a namespace for the grouped elements.</p> <p>A package may contain other packages, thus providing for a hierarchical organization of packages.</p> <p>Classes, objects, use cases, components, nodes, node instances, and associations all be organized as packages, enabling a manageable organization of UML models.</p>
	<p>Classes are used to represent information objects in the RM.</p> <p>Classes can have additional compartments to express properties (called attributes) and behaviours (called methods). Leaving the compartments blank means that the behaviour and attributes are undefined at the time of creating the diagram.</p> <p>Name tag indicates the name of the class. Classes are named using spaces and capitalising the first letter of each word that makes up the name, camelcase.</p> <p>The stereotype indicates the namespace where the class is grouped. Sometimes the stereotype can be an image. The image can be placed in place of the figure. For ODP, the stereotype for information object is a circle with a tag on top:</p> 

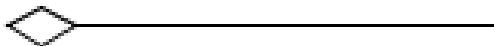
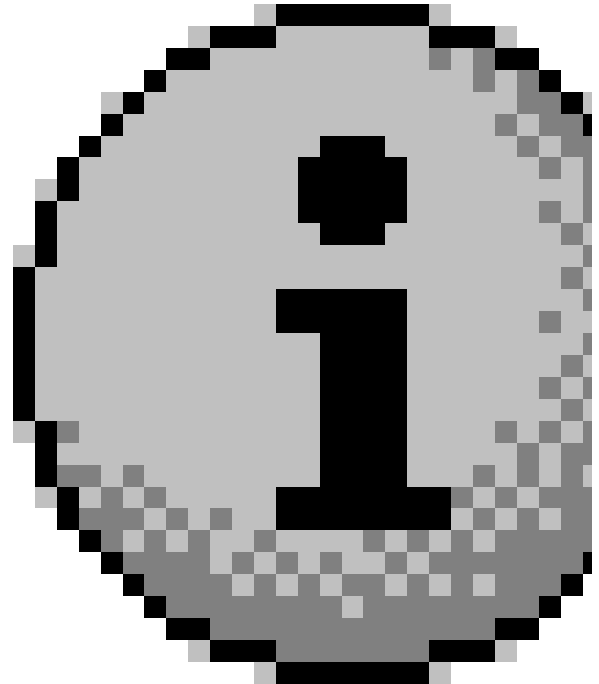


Objects are used to represent object instances in the RM.

Name tag indicates of the entity

The set of attributes with a value assigned characterises the s  
object.

The stereotype indicates the namespace where the object is g  
Sometimes the stereotype can be an image. The image can be  
place of the figure. For ODP, the stereotype for information ob  
is an "i" icon:



Aggregation indicates an association that represents a part-wh  
relationship.

Aggregation is represented with an arc with a blank rhombus c  
The end with the blank rhombus indicates the composite and t  
connects to the component.

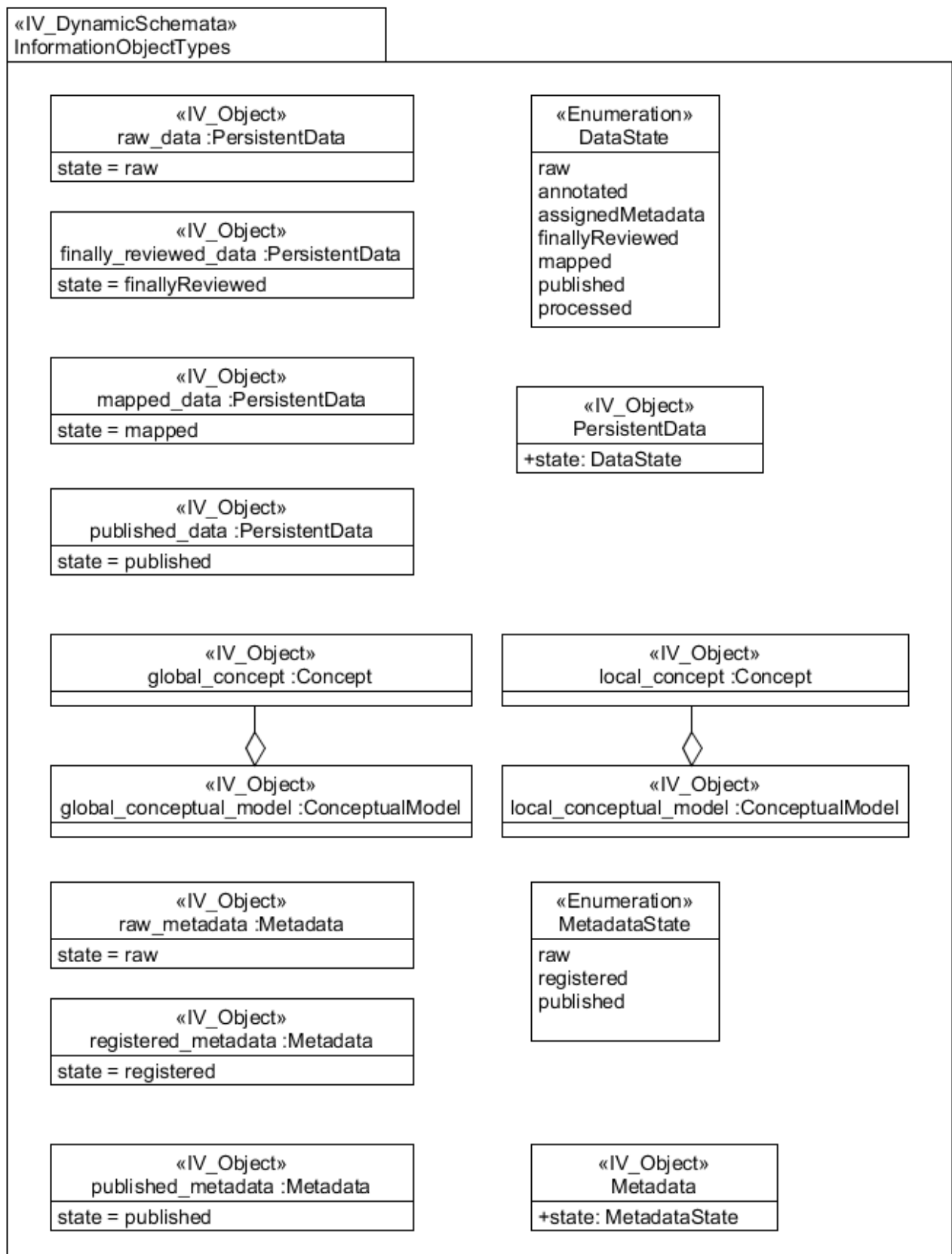


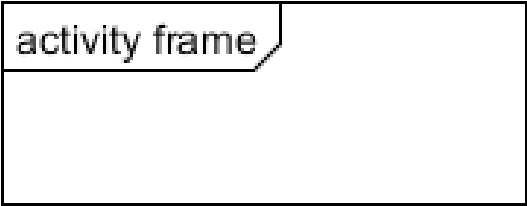
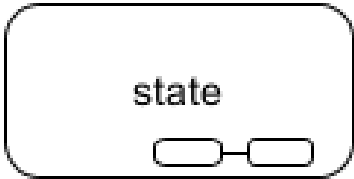



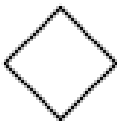
Figure 6 Example of an IV Object diagram

In the example diagram the package represents a collection of some information object instances. The stereotype for the package is dynamic schemata, which indicates that these are the parts of the model that can change depending on when the system is observed. The diagram presents four sample instances of persistent data objects and three examples of metadata objects. The diagram also includes the class definitions of persistent data and metadata objects for reference

## State Diagrams

IV Object instances can have different states during their lifespan. The basic information objects persistent data and metadata have specific sets of states associated to them. The state changes, together with the IV Activities can be used to model the behaviour of data as it is managed by the RI. For this we use a state machine diagram. The main components of state machine diagrams are activity frames, states, activities, and pseudo-states

Table 7 Notation for information object instances diagrams

Figure	Description
	<p>Frames are used to indicate the information object instance being represented.</p> <p>The name indicates the information object instance being modelled</p>
	<p>States are used to represent the state of an information object resulting from the effects of an IV action</p> <p>The name tag indicates the state reached by the information object</p> <p>The small decoration in the box can be included to indicate that the state is complex and can be subdivided into sub-states</p>
	<p>The arcs connecting states represent information actions applied to objects at a given state. The arrow end indicates the resulting state, the undecorated end indicates the initial state</p>
	<p>A filled circle is a pseudo-state, it can be used to model a start state or an intermediate connecting state</p>
	<p>A circle with a smaller filled circle in the middle is a pseudo-state to model an end state</p>
	<p>Decision pseudo-state, is used to model an exclusive fork in the execution of activities. It can also be used to model exclusive joins after forks.</p>



Fork/merge pseudo-state, is used to model a forks and joints in the execution of activities.

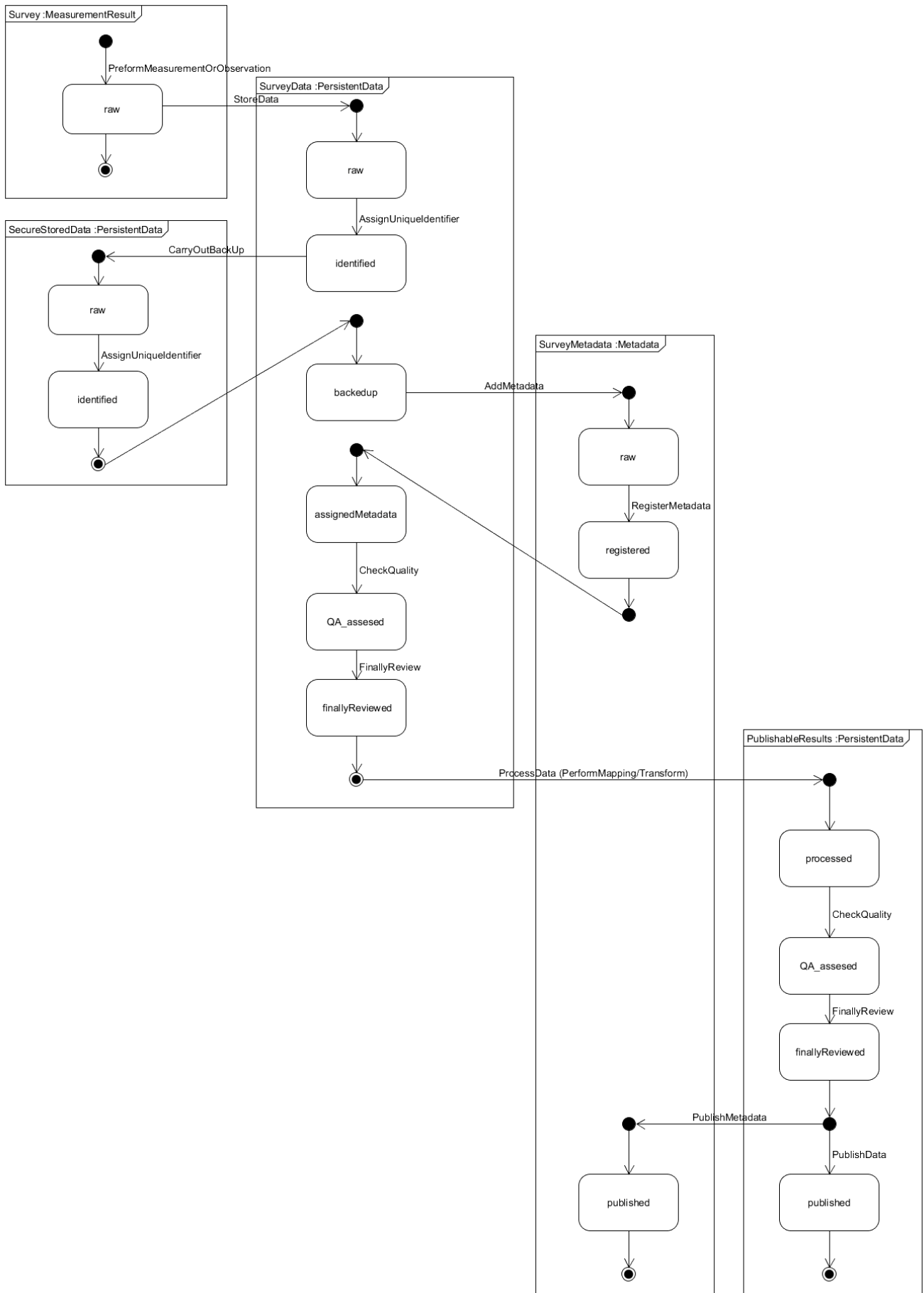


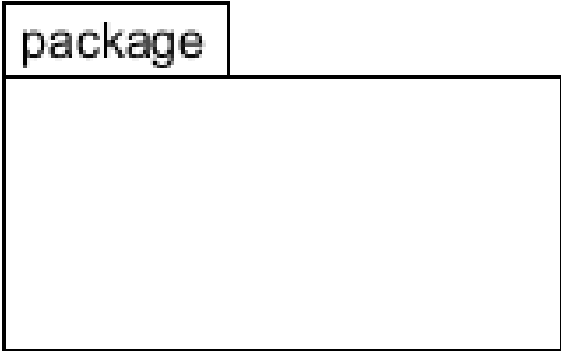
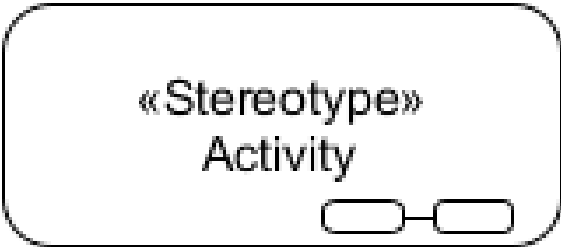
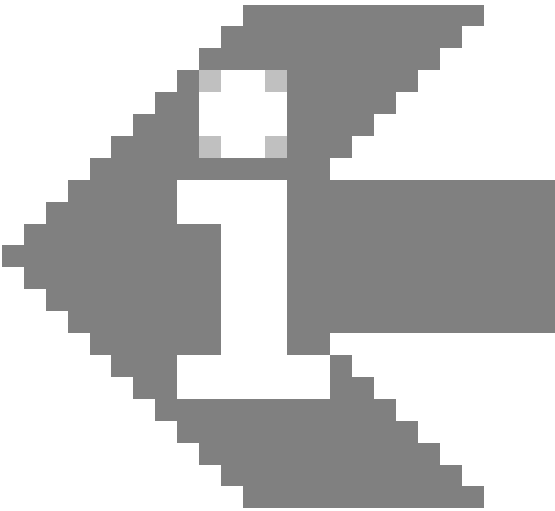
Figure 7 Example of an IV Information Object Evolution diagram

In the example diagram, five information object instances are presented. The possible transitions between states are indicated with arcs labelled using the names of IV actions.

Evolution of information objects

The evolution of information objects can also be represented using activity diagrams. Activity diagrams combine IV Object Instances and IV actions can also be combined into

Table 8 Notation for information object evolution with activity diagrams

Figure	Description
	<p>A package, in UML notation, is a grouping element. Package is used to group elements, and to provide a namespace for the grouped elements.</p> <p>A package may contain other packages, thus providing for a hierarchical organization of packages.</p> <p>Classes, objects, use cases, components, nodes, node instances, and relationships all be organized as packages, enabling a manageable organization of UML models.</p>
	<p>Activities are used to represent action in the RM.</p> <p>Name tag indicates the name of the action. Actions are named by a short phrase that describes the event or action being represented.</p> <p>The small decoration in the box indicates that the action is composite and can be subdivided into smaller tasks.</p> <p>A stereotype can be used to indicate the namespace where the action is grouped. Sometimes the stereotype can be an image. The stereotype image can be used in place of the figure. For ODP, the stereotype information action is an arrow icon with a lowercase "i":</p> 



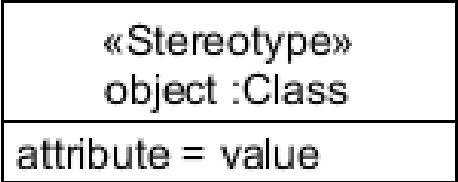

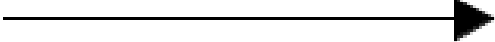


	<p>Objects are used to represent object instances in the RM.</p> <p>Name tag indicates of the entity</p> <p>The set of attributes with a value assigned characterises the s object.</p> <p>The stereotype indicates the namespace where the object is g Sometimes the stereotype can be an image. The image can b place of the figure. For ODP, the stereotype for information ob is an "i" icon:</p> 
	<p>The arcs connecting states represent transitions between infor actions. Arcs can connect activities to information object instar indicating the result of performing an action. When linking an action, the arc indicates that the object is part of the input used that action.</p>
	<p>A filled circle is used to model the start of a set of actions</p>
	<p>A circle with a smaller filled circle in the middle is used to model state</p>

Figure 8 Example of an IV Information Object Evolution using an activity diagram

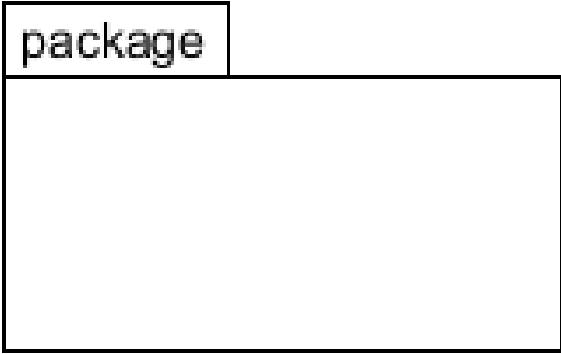
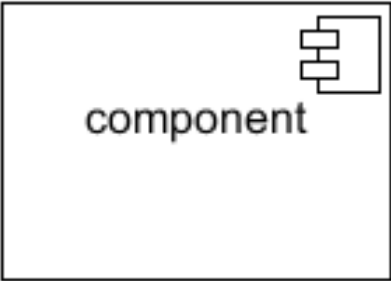

In the example diagram, an overview of the evolution of data in a RI is presented

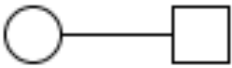
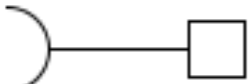

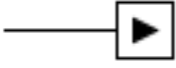


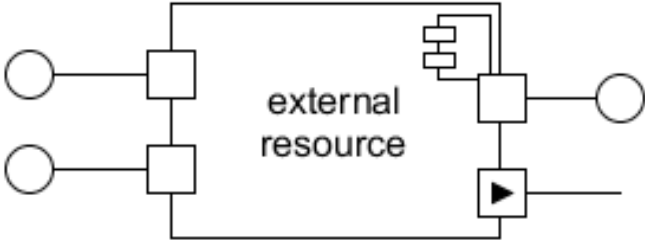
### Notation of Computational Viewpoint Models

#### Computational Objects

In the ENVRI RM, component diagrams are used for the representation of computational objects and interfaces.

Table 9 Notation for information object instances diagrams

Figure	Description
	<p>A package, in UML notation, is a grouping element. Package is used to group elements, and to provide a namespace for the grouped elements.</p> <p>A package may contain other packages, thus providing for a hierarchical organization of packages.</p> <p>Classes, objects, use cases, components, nodes, node instances, and relationships can all be organized as packages, enabling a manageable organization of UML models.</p>
	<p>Components are used to represent computational objects. The component notation contains the name of the computational object and a decorator that indicates it is a component (UML standard).</p> <p>Components can also have a stereotype, and an image associated with that stereotype. In ODP the stereotype image for computational objects is the icon of a box with a class tag in front of it:</p> 

 <p>server interface and port</p>  <p>client interface and port</p>  <p>Streaming interface producer</p>  <p>Streaming interface consumer</p>	<p>Ports and interfaces are used to represent the means of communication between objects. A small box in the border of an object is used as a port.</p> <p>A blank circle connected by an arc to a port represents a server.</p> <p>A semicircle with an arc connected to a port represents a client.</p> <p>A port with an arc and an arrow pointing away from the object represents a producer streaming interface.</p> <p>A port with an arc and an arrow pointing towards the object represents a consumer streaming interface.</p>
	<p>Generalisation is used to indicate if one object extends another, illustrated using an unfilled arrow from the sub-object to the parent object with the annotation 'is a'.</p>
	<p>The ability to create objects is noted by a single filled arrow from the creating object to the object being created, with the annotation 'new &lt;object&gt;'.</p>
	<p>Example of a computational object with four ports and four interfaces.</p>

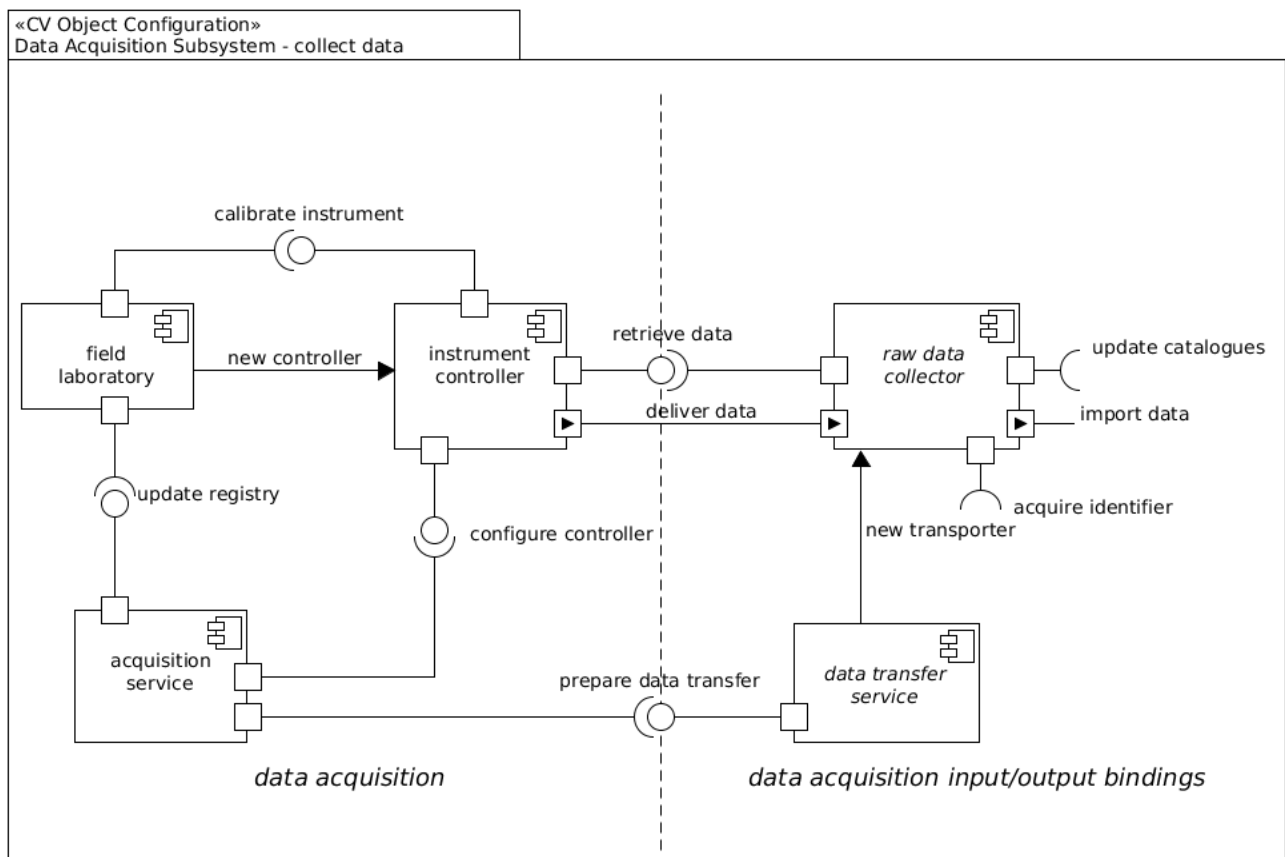
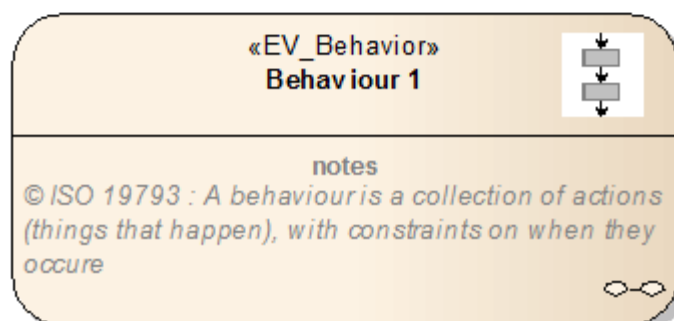
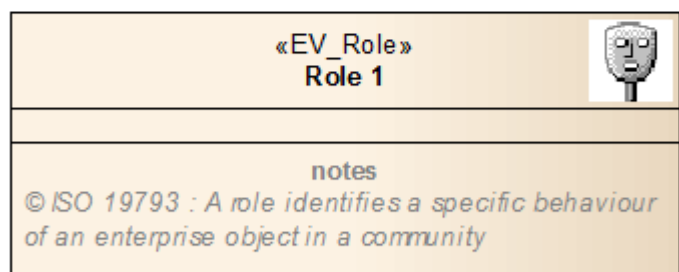
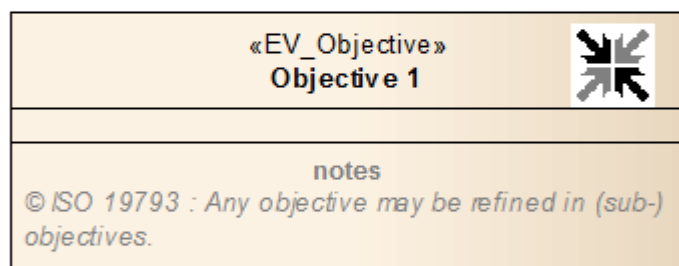
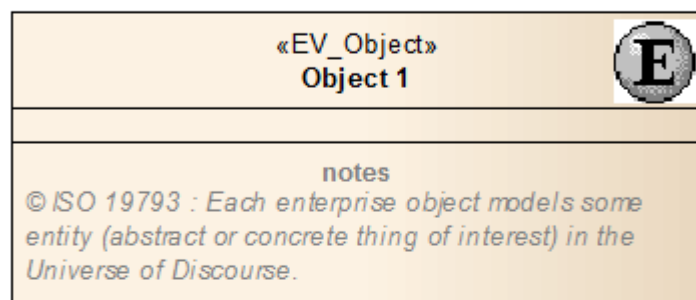
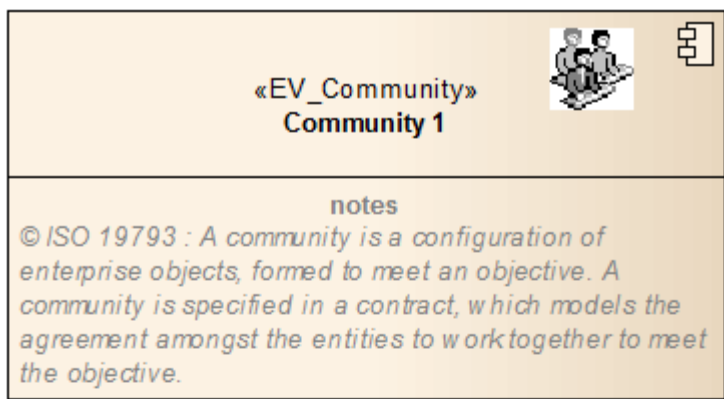


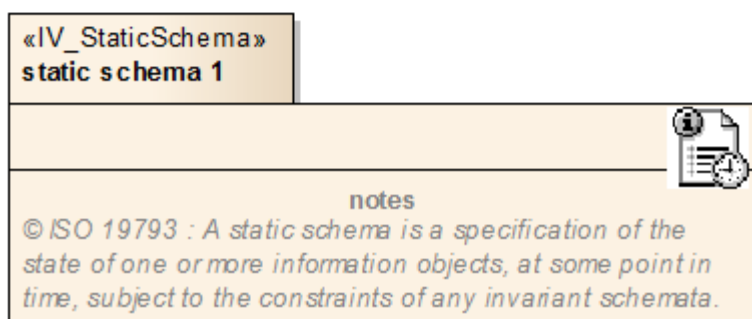
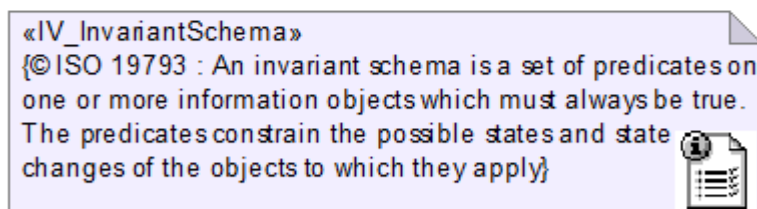
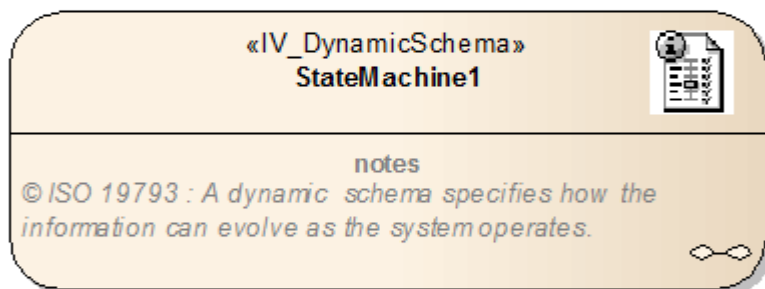
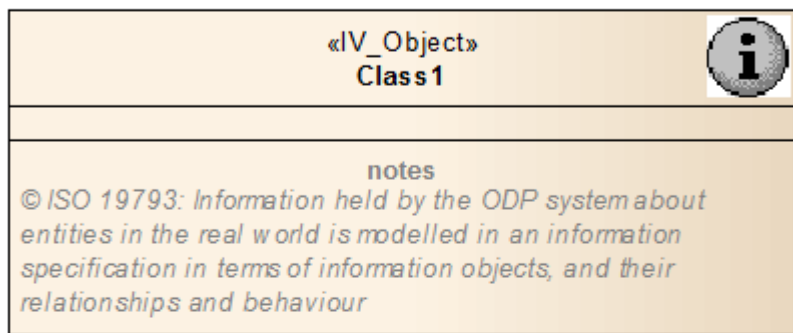
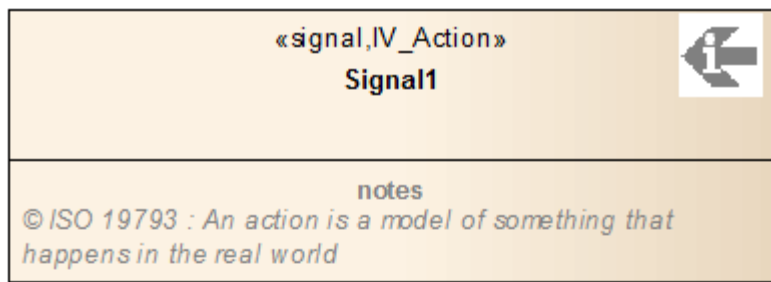
Figure 9 Example of the CV Objects for data acquisition

In the example diagram, three computational objects are presented. Balls and sockets are matched and the names of the client/server interfaces are supposed to be the same. In the example, the field laboratory client interface “calibrate instrument” is connected to the instrument controller server interface “calibrate instrument”

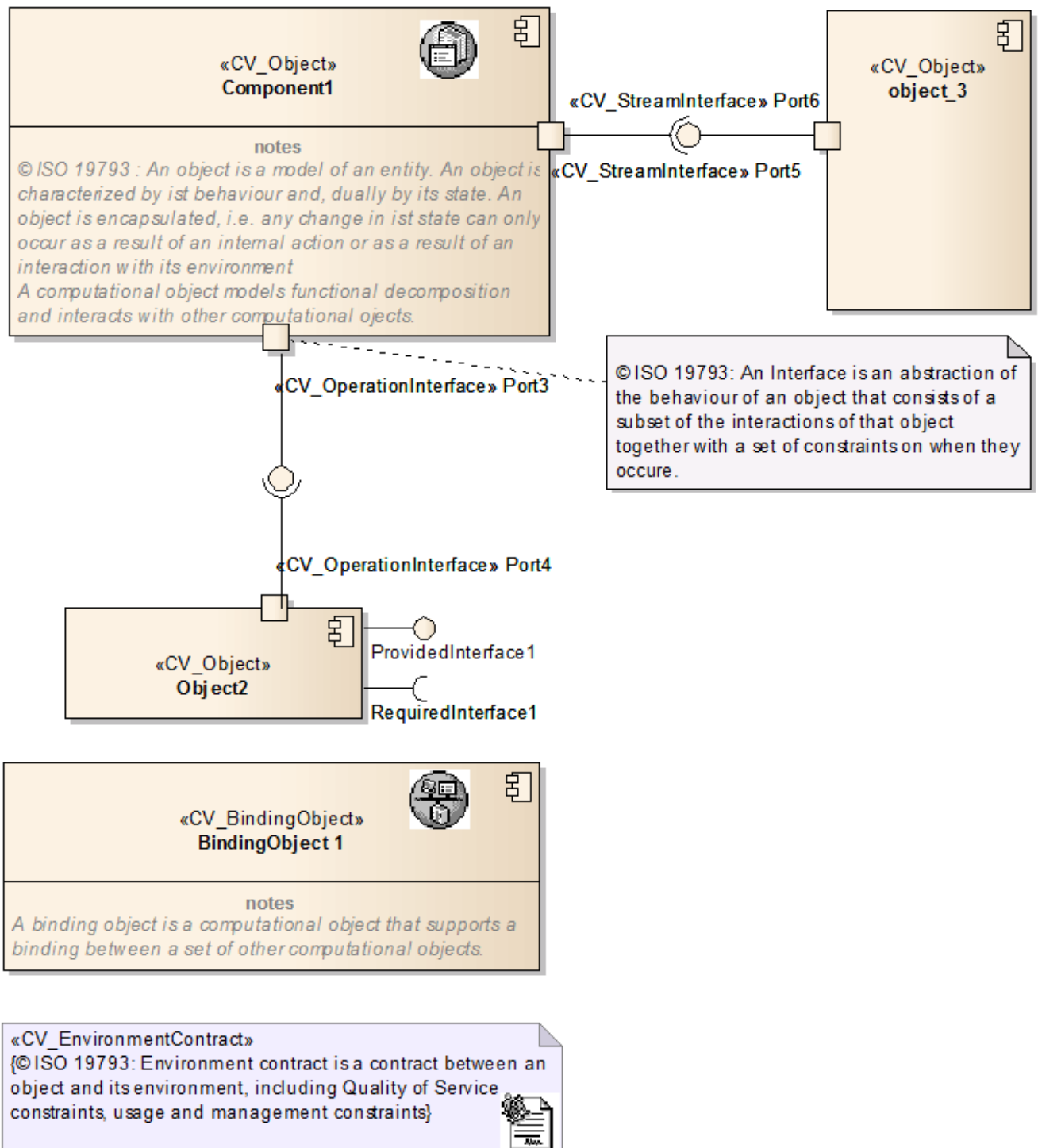
## UML4ODP Graphical Notation

### Enterprise viewpoint





computational viewpoint



## Bibliography

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, G. Athanasopoulos, D. Castelli, K. El Raheb, P. Innocenti, Y. Ioannidis, A. Katifori, A. Nika, G. Vullo, and S. Ross. "The Digital Library Reference Model", *DL.org*, 2011. <http://referencemodel.dlorg.eu/>
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)".
22. I. 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. I. 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. A. 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. A. 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.
38. Oracle, *Oracle Information Architecture: An Architect's Guide to Big Data*, An Oracle White Paper in Enterprise Architecture, August 2012.
39. Tarasova, T., Argenti M., and Marx M., *Semantically-Enabled Environmental Data Discovery and Integration: demonstration using the Iceland Volcano Use Case*, To appear in proc. of the 4th Conference on Knowledge Engineering and Semantic Web (KESW), Saint-Petersburg, Russia, 2013.
40. OMG, "Unified Modeling LanguageTM (OMG UML), Superstructure Version 2.2" , *OMG Standard*, 2009

## Guidelines for using the Reference Model

### Introduction

The development of the ENVRI Reference Model provides the ESFRI Environmental Research Infrastructures with a common ontological framework for description and characterisation of computational and storage infrastructures, and provides them a community standard to help achieve greater levels of interoperability between their heterogeneous resources.

The Reference Model defines a conceptual model that captures computational requirements and state-of-the-art design experiences. In a sense, the model reveals a snapshot of the existing landscape of the ESFRI environmental science research infrastructures at a high level of abstraction.

In order to help Reference Model users map the abstraction to concretions, so as to better apply the knowledge in their daily practices, we prepare this guideline that introduces our own experiments with the Reference Model, and in doing so reveal the principles of usage. These principles are neither bound nor enforced. They are not mandatory for users to follow. The intention is to provide users with a way of thinking, which may lead to exploration of the model itself and inspire the discovery of various way of using the model.

Rather than going through each model term and explaining the meaning of it, we use a set of practical examples, each of them illustrating some aspects of the usage of the reference model as well as introducing a number of model concepts.

Initially, examples are selected with the aim to serve **primary audience** within the community of ESFRI Environmental Research



Infrastructures. We use scenarios that are familiar to our users, and include information that may be of interest to the community and perhaps benefit their work.



To collect these examples, we used a template with 5 questions:

1. What is this use case about? *Describe the purpose of the use case, and any background information.*
2. How can the reference model be used in this use case?
3. What are the results of using the reference model? *Evidence of usefulness/utility.*
4. What are the benefits of using the reference model? *Demonstrate specific cases of things that could not have been achieved without the RM.*
5. Are there any problems with using the reference model in this use case? *Feedback from users.*

These questions proved to be helpful in organising investigation activities. We encourage readers also to use this template to structure newly developed stories and share them with us so as to inspire others.

With limited resources, only few examples are included; these will be extended when more resources are available for future investigations.

## How to Use the Guideline

A collection of examples demonstrating usage of the ENVRI Reference Model is given below. Different examples may serve different purposes. Some of them merely illustrate a different way of using the reference model (e.g., Example 5), while others also intend to introduce model concepts where many terms are highlighted with clickable links. Please click those highlighted concepts that will re-locate you to the related definitions and specifications in the Reference Model. Be sure to go through all terms marked with  -- some of them, though repeated, will guide you to a different part of the model. By visiting all linked contents, you will have explored 90% of the most important model content. (Note, terms marked with  are also model concepts which link to content you might have visited before.)

## Examples of Using the Reference Model

[Example 1: Using the Reference Model to guide research activities \(EISCAT 3D - EGI\)](#)

[Example 2: Using the Reference Model as an analysis tool \(EUDAT\)](#)

[Example 3: Using the Reference Model in documentation \(EMSO\)](#)

[Example 4: Using the Reference Model as design reference \(EPOS\)](#)

[Example 5: Using the Reference Model to drive implementations of common services \(WP4 practices\)](#)

## Conclusions

Using a number of examples, we have shown that by using the Reference Model, a ESFRI ENV RI could benefit from:

- **A set of ready-to-use terminology with a publicly-accessible reference base**, which can be used to describe requirements and architectural features of an infrastructure, and serve as a common language in communication materials; in particular, with an external community without any specific knowledge of the scientific domain being addressed.
- **A uniform framework with well-defined subsystems** of components specified from different complementary viewpoints (Science, Information and Computation), which promotes structural thinking in constructions of system architectures, and can be used as a research tool for comparison and analysis of heterogeneous infrastructures.
- **A knowledge base capturing existing requirements and state-of-the-art design experiences**. The information provided can be referred to in various system analysis tasks, to guide design and implementation activities, and to drive the development of common services.

When future resources become available, we will conduct more investigations, including:

- We will assist our users to get hand on the Reference Model and exploit new ways of using it.
- We will assist the development of the common services.
- We will use the Reference Model to bridge ESFRI ENV RIs with external communities (such as, RDA), projects (such as, GEOSS, DataOne, EUDAT and EGI), and standards (such as, INSPIRE, OGC, and the Digital Library Reference Model). These will provide ESFRI ENV RIs an overview of related technologies, and possible solutions for the integrations.
- We also have a plan to experiment with the Reference Model as a guide to train the next generation data scientists.

## Tutorials

- ENVRI Reference Model: an Overview. [[ppt](#)]
- Main Processes of the ENVRI Reference Model – Corresponding Viewpoint [[ppt](#)]

## Example 1: Using the Reference Model to Guide Research Activities (EISCAT 3D - EGI)

### Descriptions of the Example

This example explains the usage of the Reference Model in a pilot project that investigates the big data strategies for the EISCAT 3D

research infrastructure. The Reference Model serves as a knowledge base to guide various research activities.

EISCAT, the *European Incoherent Scatter* Scientific Association, was established to conduct research on the lower, middle and upper atmosphere and ionosphere using the incoherent scatter radar technique. This technique is the most powerful ground-based tool for these research applications. A next generation incoherent scatter radar system, EISCAT 3D, is being designed. The multi-static radars to be used will be a tool to carry out plasma physics experiments in the natural environment, a novel atmospheric monitoring instrument for climate and space weather studies, and an essential element in multi-instrument campaigns to study the polar ionosphere and magnetosphere. It will be a world-leading international research infrastructure, using the incoherent scatter technique to study how the Earth's atmosphere is coupled to space.

The design of the EISCAT 3D opens up opportunities for physicists to explore many new research fields. On the other hand, it also introduces significant challenges in handling large-scale experimental data that will be massively generated at great speeds and volumes. During its first operation stage in 2018, EISCAT 3D will produce 5PB data per year, and the total data volume will rise up to 40PB per year in its full operations stage in 2023. This challenge is typically referred to as a big data problem and requires solutions beyond the capabilities of conventional database technologies.

EISCAT is currently considering the use of e-Science technologies to deliver strategies for handling its big data products. Advanced e-Science infrastructure projects such as [EGI](#), [PRACE](#), and their enabling technologies are making large-scale computational capacities more accessible to researchers of all scientific disciplines. Emerging infrastructures, such as cloud systems proposed by [the Helix Nebula project](#) and by [the EGI Federated Cloud Task Force](#), or the data infrastructure to be provided by [EUDAT](#) will extend possibilities even further.

As a potential of e-science partner for EISCAT, we present EGI. EGI was established in 2010 as a Europe-wide federation of national computing and storage resources. The EGI collaboration is coordinated by [EGI.eu](#), a not-for-profit foundation created to manage the infrastructure on behalf of its participants: National Grid Initiatives and European Intergovernmental Research Organisations. Resources in EGI are provided by about 350 resource centres from the NGIs who are distributed across 55 countries in Europe, the Asia-Pacific region, Canada and Latin America. These providers operate more than 370,000 logical CPUs, 248 PB disk and 176 PB of disk capacity (June 2013 statistics) to drive research and innovation in Europe and beyond.

Since February 2013, a pilot project has been set up within ENVRI, which establishes a partnership between EISCAT, EGI and EUDAT, aiming to identify and allocate solutions that directly benefit EISCAT 3D, which can also be reused in other ESFRI projects involved in ENVRI. ENVRI WP3 has been involved in this investigation, and uses the Reference Model to guide various research activities, including;

- Analysis of the EISCAT 3D data infrastructure; Capturing requirements from the EISCAT 3D scientific community concerning applications that work with and process data.
- Analysis of EGI and EUDAT services; Identifying the gaps between the generic service infrastructures of these providers and the domain-specific requirements of EISCAT 3D.
- Provide recommendations to EISCAT 3D for the setup of up a big data strategy and a big data infrastructure for its community. Setup demonstrators/proof-of-concept systems if resources permit.

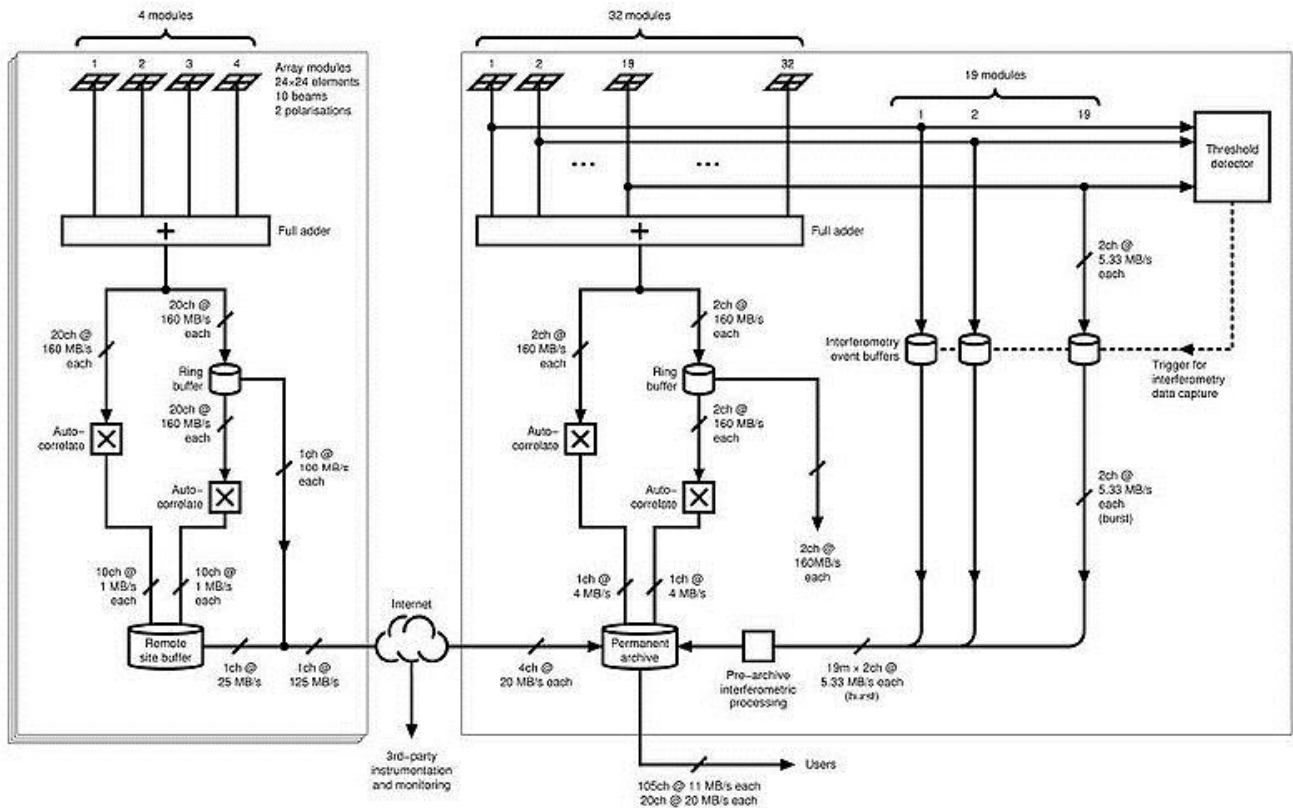
Having fulfilled these tasks, the Reference Model is proving to be useful as a knowledge base that can be referred when conducting various system analysis and design activities.

## How to Use the Reference Model


In the following, we describe how the Reference Model is used to conduct several system analysis tasks.

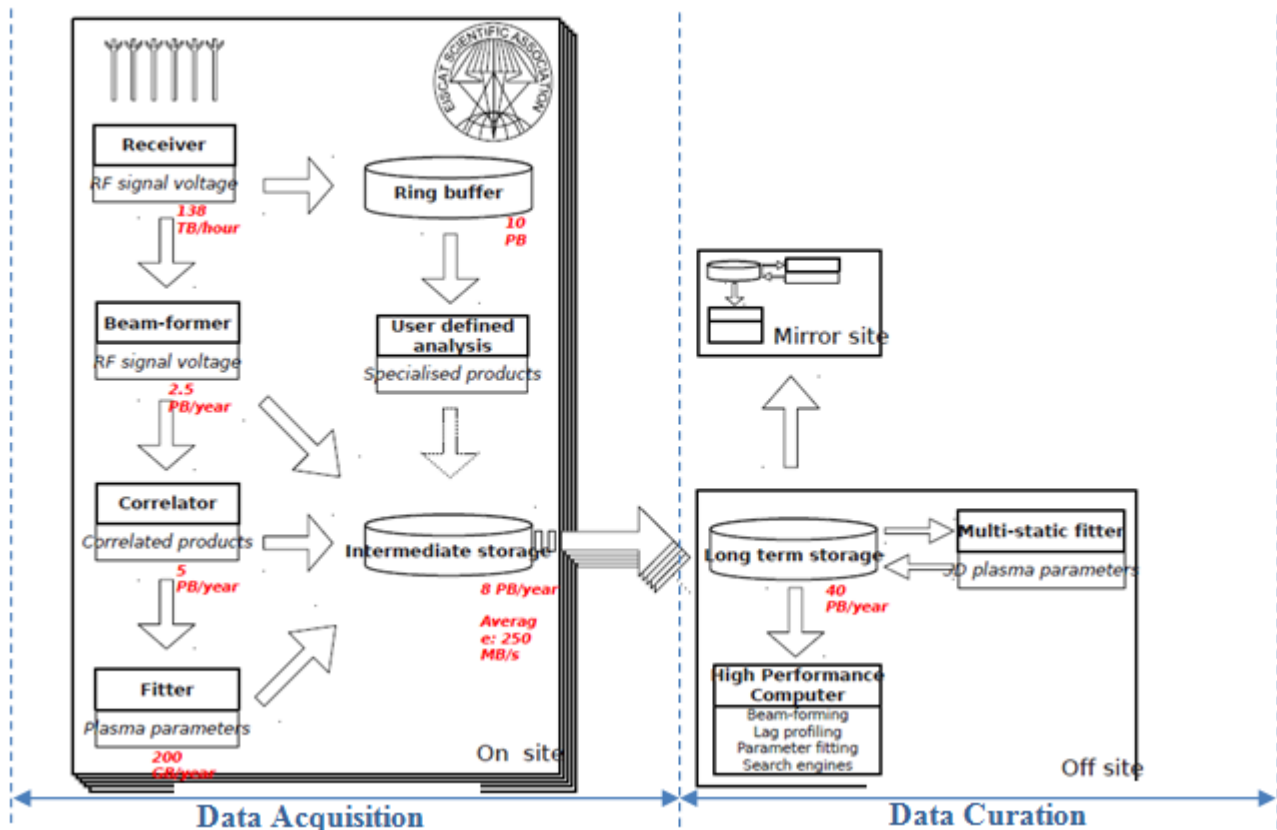
### Analysis of the EISCAT 3D Data Infrastructure

The initial challenge for the pilot project is to understand the EISCAT 3D data infrastructure. The existing design documents of EISCAT 3D has been focused on the incoherent scatter radar technologies. As shown in Figure 1, its data infrastructure is embedded within the overall design of the observatory system that is difficult for a computer scientist/technologist having little physics knowledge background to understand.







**Figure 1:** The original design of EISCAT 3D data infrastructure is embedded within the overall observatory system design

We use the  **5 ENVRI common subsystem** framework to decompose the computational elements, clarifying the boundary between the radar network and data infrastructure, which results in Figure 2. This diagram now, instead of Figure 1, is frequently used in presentations and discussions of the EISCAT 3D data infrastructure.



**Figure 2:** Using the 5 ENVRI Common Subsystem to interpret the EISCAT 3D data infrastructure makes it easy to communicate with computer scientists/technologists

Figure 2 illustrates that the EISCAT 3D functional components can be placed into 2 ENVRI common subsystems,  **data acquisition** and  **curation**. Briefly, at the  **acquisition subsystem**, the raw signal voltage data will be generated by the antenna *Receivers* at the speed of 125 TB/hr, and be temporarily stored in a *Ring buffer*. A second stream of RF signal voltages will be passed to a *Beam-former* to generate the beam-formed data (1MHz). Continually, the beam-formed data will be processed by a *Correlator* to generate correlation analysis data based on standard methods. Then, the correlation data will be delivered to a *Fitter* to produce the fitted data (1GB/year). In order to support different user requirements, EISCAT 3D will allow users to access and process the raw voltage data in the *Ring buffer* and to generate the specialised products based on self-defined analysis algorithms. Both raw data and their products will be stored in *Intermediate storage* (11PB /year), from where they will be delivered to the central site within the curation subsystem.

In  **the curation subsystem**, *Long-Term Storage* will preserve the raw voltage data and their products. A *High Performance Computer* will be used for data searching and processing (e.g., beam forming, lag profiling or other correlation, and parameter fitting). Searching facilities will enable user to search over all data products and to identify significant data signatures. A *Multi-static fitter* will be installed to process the stored raw voltage data to generate the 3D plasma parameters that will then be stored back in *Long-Term Storage*. A complete copy of *Long-Term Storage* data will be established at mirror sites; related data processing and searching tools will be provided.

While it is made clear that the design specification covers 2 of 5 common subsystems described in the ENVRI Reference Model, we understand functionalities of the other 3 subsystems are currently missing. The reason of this is likely due to resource limitations. However, the absent 3 subsystems are crucial for a big data system such as EISCAT 3D. Without providing services to support data discovery, access, processing and user community, the value of EISCAT 3D big data cannot be unlocked, and expensively generated and archived scientific data will be useless.

Using the Reference Model as the analysis tool, we identified the missing pieces of the design specification, which gives the direction for future investigation.

## Analysis of EGI Enabling Services and Construction of an Integrated Infrastructure













We need to understand the functionalities of EGI services and how to integrate them to support the EISCAT 3D requirements.



A set of generic services are enabled by the EGI e-Infrastructure, including:

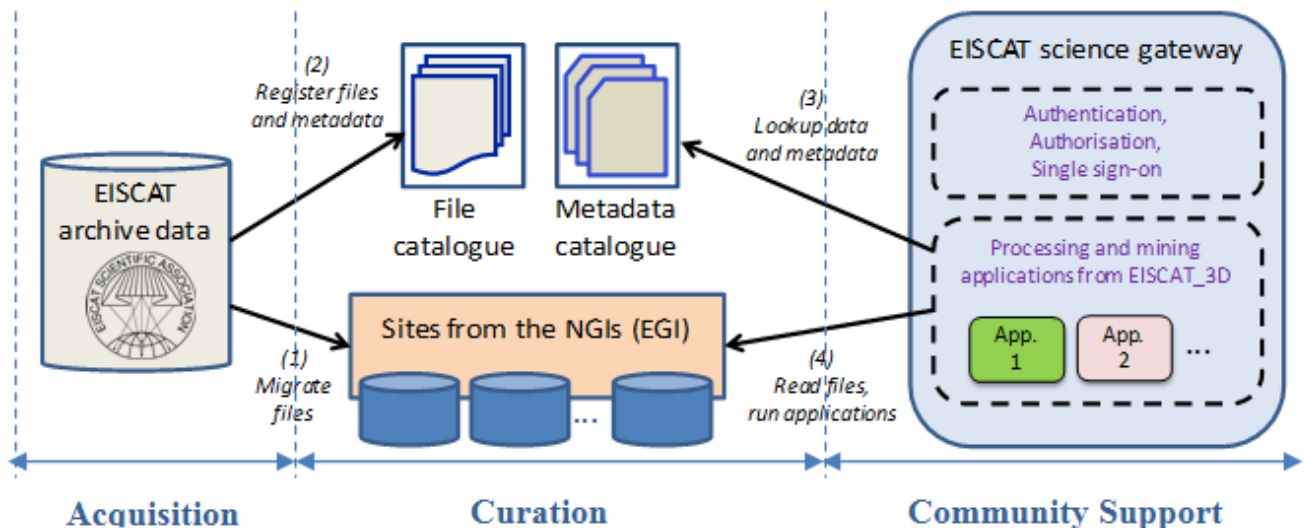
- AMGA Metadata catalogue
- LFC File catalogue
- Storage elements
- File Transfer Service
- Portal for application development & hosting (e.g. SCI-BUS)
- Access control

Showing in Table 1, by examining the functionalities of the EGI services and mapping them to the ENVRI Reference Model computational model objects, we understand these services fall into 2 ENVRI common subsystems: Curation and Community Support.

**Table 1:** Mapping EGI Services to the Reference Model Elements (from  **computational** perspective)

EGI Services	ENVRI- RM Computational Objects	ENVRI Common Subsystem
AMGA Metadata catalogue	 <b>Catalogue service</b>	 <b>Curation</b>
LFC File catalogue	 <b>Catalogue service</b>	 <b>Curation</b>
Storage elements	 <b>Data store controller</b>	 <b>Curation</b>
File Transfer Service	 <b>Data transfer service</b>	 <b>Curation</b>
Portal for application development & hosting	 <b>Virtual laboratory</b>	 <b>Community Support</b>
Access control	 <b>Security service</b>	 <b>Community Support</b>

Above analysis gives clues to a solution for integrating the EGI technologies into the EISCAT 3D data infrastructure. Depicted in Figure 3, a secondary  **data curation subsystem** (seen as the mirror site of the EISCAT 3D central archive in Figure 1) can be established using the EGI infrastructure and its services. Data from EISCAT 3D central archive (or the acquisition subsystem) can be staged into the EGI storages, and be managed using LFC File Catalogue and AMGA Metadata Catalogue. At the front end, an EISCAT science gateway can be established, seen as part of a  **community support subsystem**, to provide access control (e.g., authentication, authorisation, and single sign-on) and application portals (e.g., to which processing and data- mining applications from EISCAT 3D can be plugged in).



**Figure 3:** An integrated infrastructure of EGI and EISCAT 3D

Using the Reference Model, functional elements of both EISCAT 3D and EGI can be placed into a uniform framework, which provides a way of thinking about the construction of the integrated infrastructure.

### Evaluation of the Feasibilities of the EGI Infrastructures and Services in Supporting EISCAT 3D Requirements

Using the common framework enabled by the Reference Model, we can analyse and compare the EGI and EUDAT generic service infrastructure and the requirements from a domain-specific data infrastructure such as EISCAT 3D, and we understand that there are significant gaps in-between, including but not limited to:

- Staging services to ship scientific data from observatory networks into the EGI generic service infrastructure (and to get the data off) are missing. Such a staging service should be able to transmit both big chunk of data (up to petabyte) and continuing updates/real-time data streams during operations. Such a service should satisfy performance requirements, including:
  - Robust. Environmental scientific research needs high quality data. In particularly, during important natural events, losing observation data is unaffordable. Fault-tolerance is desirable, which requests the transmission service can be self-recover from the interruption point without restarting the whole transmission process.
  - Fast, e.g., in the case of EISCAT 3D, the 10PB ring-buffer can only hold data for about 3 days, and the big observation data need to be transferred to the archive storage fast enough to avoid being overwritten.
  - Cheap, e.g., the observatory networks are remote from the EGI computing farm. Using high-capacity pipes are possible but expensive. Software solutions such as, intelligent network protocols, optimisation, data compression, are desirable.
- Cost effective large storage facilities and long-term archiving mechanisms are urgently needed. Environmental data, in particular for climate research, need to be preserved over the long-term to be useful. Being Grid-oriented, EGI is not designed for data archiving purposes. Although large storage capabilities are potentially available through NGI participants, EGI does not guarantee long-term persistent data preservation. Curation services such as advanced data identification, cataloguing and replication are absent from the EGI service list.
- The EGI infrastructure needs to adapt in order to handle emerging big- data phenomena. The challenge is how to integrate what is new with what already exists. Services such as job schedulers need to be redesigned to take into account the trade-off of moving big data; intelligent data partitioning services should be investigated as a way to improve the performance of big data processing.
- Advanced searching and data discovery facilities are urgently needed. It is often said that data volume, velocity, and variety define big data, but the unique characteristic of big data is the manner in which the value is discovered [38]. Unlike conventional analysis approaches where the simple summing of a known value reveals a result, big data analytics and the science behind them filter low value or low-density data to reveal high value or high-density data [38]. Novel approaches are needed to discover meaningful insights through deep, complex search, e.g., using machine learning, statistical modelling, graph algorithms. Without facilities to unlock the value of big data, expensively generated and archived scientific data will be useless.
- Community support services are insufficient. The big data phenomena will eventually lead to a new data-centric way of conceptualising, organising and carrying out research activities that could lead to an introduction of new approach to conducting science. A new generation of data scientists is emerging with new requirements. Service facilities should be planned to support their needs. These together should enable the EISCAT 3D community to design new applications that are capable to work with big data, and can implement these on cutting-edge European Distributed Computing Infrastructures.
- Currently, EUDAT has taken up the role to implement a collaborative data infrastructure, however only a few services are available, storage facilities are insufficient, and policies for usage are unclear. Among our current investigations, we are investigating the possibility of integrating EUDAT services into EGI infrastructure, seen as a layer on top of the EGI federated computing facility. The analysis of the EUDAT services is included in [another usage example](#) of the Reference Model.

### Summary

In this example, we have shown that the Reference Model could be used to conduct various system analysis tasks. Using the Reference Model we have:

- Clarified the boundary of EISCAT 3D data infrastructure and identified missing functionalities in the design;
- Provided a solution to integrate the EGI services into EISCAT 3D data infrastructure;
- Identified gaps between the EGI generic service infrastructure with the requirements from a domain specific research infrastructure,



We have shown that the Reference Model offered a research infrastructure:

- A knowledge base containing useful information could be referred in various system analysis and design activities;
- A uniform platform into which computational elements of different infrastructures could be fitted, enabling comparison and analysis;
- A way of thinking of constructions of plausible system architectures.

## Example 2: Using the Reference Model as an Analysis Tool (EUDAT)

### Description of the Example

This study case provide an example for ESFRI Environmental Research Infrastructures project managers and architects to use the ENVRI Reference Model as an analysis tool to review an emerging technology, the EUDAT data infrastructure and its service components. Such an analysis can help them better understand the newly developed technologies and decide on how to make use of the generic services provided in their own research infrastructures.

The EU-funded [EUDAT project](#) is developing a pan-European data infrastructure supporting multiple research communities. Such a generic data infrastructure is seen as a layer in the overall European scientific e-infrastructure to complement the computing layer (EGI, DEISA, PRACE) and the networking layer (GEANT).

The design activities of EUDAT are driven by use-case-based community requirements EUDAT reviews the approaches and requirements of different communities, such as linguistics ([CLARIN](#)), solid earth sciences ([EPOS](#)), climate sciences ([ENES](#)), environmental sciences ([LIFEW ATCH](#)), and biological and medical sciences ([VPH](#)), identifying common services, and provides computational solutions. Initially, 4 services are provided within EUDAT data infrastructure:

- **Safe replication:** which enables communities to replicate datasets -- using the integrated Rule-Oriented Data System ([iRODS](#)) as a replication middleware -- within data centre sites, with persistent identifiers automatically assigned to the digital objects in order to keep track of all the replicas;
- **Data staging:** which enables easy movement of large amounts of data between EUDAT storage resources and workspace areas on high-performance computing (HPC) systems to be further processed.
- **Metadata Catalogue:** which allows researchers to easily access metadata of data (or their collections) stored in EUDAT nodes. EUDAT will also harvest external metadata (which contains pointers to actual data) from stable metadata providers to create a comprehensive joint catalogue that will help researchers to find interesting data objects and collections.
- **Simple Storage:** which allows registered users to upload "long tail" data objects (large in number but small in size), and share such objects with other researchers.









We use the concepts developed in the ENVRI Reference Model to analyse the EUDAT data infrastructure and its service components. Only cursory analysis is provided, since the main purpose of the study case is to illustrate the usage of the ENVRI Reference model.



### How to Use the Reference Model








#### Analysis of EUDAT common services and components






The ENVRI Reference Model models an archetypical environmental research infrastructure (RI). As a service infrastructure, EUDAT itself is therefore not an implementation of the Reference Model, but is rather a source of implementations for instances of objects required by any RI implementing the Model.



**Table 1:** Mapping EUDAT Services to the Reference Model Elements

EUDAT Services	Computational Viewpoint	ENVRI Common Subsystem
Safe replication	 <b>Data Transfer Service</b>	 <b>Curation</b>
Staging	 <b>Data Importer</b>	 <b>Curation</b>
Metadata Catalogue	 <b>Catalogue Service</b>	 <b>Curation</b>
Simple Store	 <b>Data Store Controller</b>	 <b>Curation</b>


From the  **computational** perspective, EUDAT offers services that can be used to instantiate various objects in the Reference Model. For example EUDAT's Safe Replication facilities can implement various required services within the  **Data Curation subsystem** of an environmental RI:

-  **Acquisition:** EUDAT does not offer facilities for  **data acquisition**, relying on data already gathered by client RIs.
-  **Curation:** EUDAT can provide instances of any of the computational objects used for data curation (including  **data store controllers**,  **data transfer services** and  **catalogue services**), either in place of or complementary to instances provided by an environmental RI – the extent to which EUDAT assumes the curation role for an infrastructure will vary from case-to-case.
-  **Access:** Data access to EUDAT curated data is brokered by EUDAT, whilst the RI would broker RI-curated data. In practice the

- RI  **broker** would sit in front of the EUDAT broker, forwarding data requests that involve data delegated to EUDAT.
-  **Processing**: EUDAT do not offer data processing (beyond *metadata annotation*) as a core service; *workflow enactment* is being investigated as a future service however, which would allow a later version of the EUDAT platform to implement elements of a  **Data Processing subsystem**.
- Whilst certain aspects of EUDAT such as the Simple Store for researchers might be directly accessible as an independent  **gateway service**, in general EUDAT sits behind a client RI, its services hidden behind the RI's native services from the perspective of the RI's user community. It would be likely however that the 'virtual laboratories' by which community groups interact with an RI would be in some way augmented by EUDAT services; in particular, implementations of the  **Security Service** would integrate the EUDAT AAI service to allow seamless integration of EUDAT-held datasets with locally-held RI datasets.

The most immediately apparent conclusion that can be drawn from cursory analysis of EUDAT services in the context of the Reference Model is that EUDAT can potentially implement the entire  **Data Curation subsystem** of an environmental RI; however in practice, one would expect that an RI would retain a certain amount of data locally (particularly raw data that is expensive to transfer off-site), necessitating a more nuanced division of labour between the RI and EUDAT. In particular, EUDAT provides replication services, allowing the co-existence of RI and EUDAT data stores holding the same data, and EUDAT provides metadata (including global persistent identifier) services, allowing EUDAT to provide any  **catalogue service** (probably complementary to catalogue services maintained by an environmental RI itself). The delegation of services will be a product of negotiation between the environmental RI and the EUDAT project (some degree of automation may be possible, but likely sufficient for only smaller projects).

## Summary

The principal potential benefit of using the Reference Model in general is the ability to precisely identify components required by an environmental RI and then identify how (if at all) the RI implements those components. In the EUDAT context, EUDAT provides a number of services that implement certain components (primarily in  **Data Curation**); it should therefore be possible to identify the equivalent services in a modelled RI and determine whether or not there is a benefit to delegating those services to EUDAT. This decision may be based on cost (particularly related to economies of scale) and development time (in cases where the RI has not yet implemented the service, but may be able to use the EUDAT service instead).

## Example 3: Using the Reference Model in documentation (EMSO)

### Descriptions of the Example

Researchers and architects of an ESFRI Environmental Research Infrastructure often encounter requests to describe their infrastructure, to introduce its particular architectural features, or to explain system requirements. The Reference Model offers a set of ready-to-use terminology with explicit definitions, which can be applied to various documentations. This example tells how the Reference Model has been used as a common language in writings to communicate with a community other than environmental science.

The **Research Data Alliance** (RDA) is established to accelerate international data-driven innovation and discovery by facilitating research data sharing and exchange, use and re-use, standards harmonization, and discoverability. This will be achieved through the development and adoption of infrastructure, policy, practice, standards, and other deliverables.

ENVRI has been actively supporting the RDA activities and made various contributions. In particular, ENVRI has been accepted as one of the use cases by the RDA Data Foundation and Terminology (DFT) working group, which has been set up to gather emerging requirements as well as to test research outcomes.

In preparing the use case, researchers and architects from two ENVRI-participating research infrastructures, EMSO and EPOS, used the terms and concepts defined in the Reference Model to describe architectural features of their research infrastructures. The resulting document from EMSO is presented below.

## How to Use the Reference Model

The European research infrastructure EMSO is a European network of fixed-point, deep-seafloor and water column observatories deployed in key sites of the European Continental margin and Arctic. It aims to provide the technological and scientific framework for the investigation of the environmental processes related to the interaction between the geosphere, biosphere, and hydrosphere and for a sustainable management by long-term monitoring also with real-time data transmission. Since 2006, EMSO has been on the ESFRI (European Strategy Forum on Research Infrastructures) roadmap; it entered its construction phase in 2012. Within this framework, EMSO is contributing to large infrastructure integration projects such as ENVRI and COOPEUS. The EMSO infrastructure is geographically distributed in key sites of European waters, spanning from the Arctic, through the Atlantic and Mediterranean Sea to the Black Sea. It is presently consisting of thirteen sites that have been identified by the scientific community according to their importance respect to Marine Ecosystems, Climate Changes and Marine GeoHazards.

The data infrastructure for EMSO is being designed as a distributed system. Presently, EMSO data collected during experiments at each EMSO site are locally stored and organized in catalogues or relational databases run by the responsible regional EMSO nodes. The EMSO data architecture is currently adapted to the ENVRI Reference Model. As shown in Figure 1, according to the ENVRI-RM it includes the 5 ENVRI common subsystems. Concepts and terms defined in ENVRI-RM are used to illustrate the currently practiced common data management strategies for real time as well as archived data within the EMSO distributed data management system.

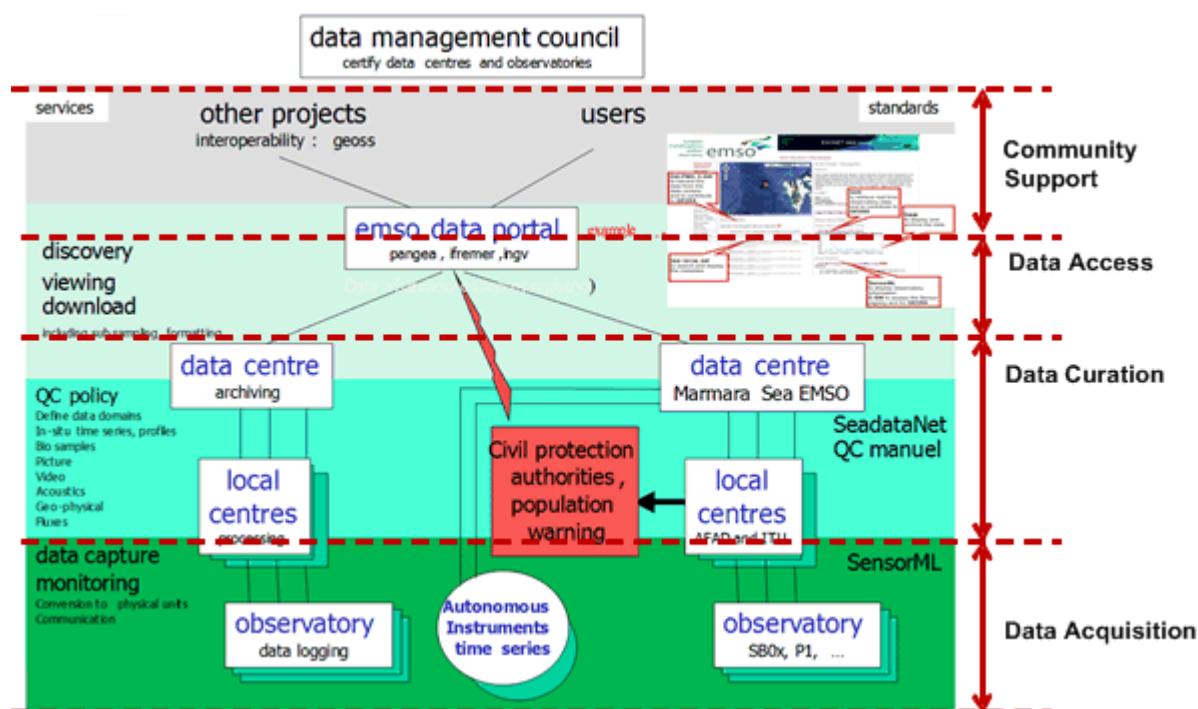


Figure 1: EMSO Distributed Data Management System

### Data Acquisition

The EMSO **data acquisition sub-system** collects raw data from EMSO's marine observatories, which represent sensor arrays of varying geometry and various instruments or human observers, and brings the measures (data streams) into the system. **Set-up** and **design of each observatory** is specified depending on the scientific demands and includes **specification of sampling designs and measuring method**.

Depending on the deployment situation and nature of collected data, EMSO data is collected in real-time or delayed mode. Both **data collection** methods are performed by the regional nodes of EMSO that are responsible for the operation of marine observatories. Marine observatories have to deal with many technological challenges due to their extreme, deep sea deployment locations. Therefore data acquired by marine observatory sensor systems is most often temporarily staged within the instruments or the observatory's internal storage systems, and real-time transmission of data is only provided by observatories that are connected by submarine cables or permanent satellite connections. Whereas real time data are immediately available, the staged data becomes available for these systems only after visits during dedicated ship expeditions when the instruments are recovered or maintained. In addition, data are acquired through laboratory studies performed on material or samples collected at marine observatory sites such as multidisciplinary analyses of water samples, sediment cores, tow or trap catches.

Depending on the instrumentation and observatory design, on-site quality control and data filtering is applied, generally followed by a transformation process which converts the instrument specific data format into a transmission format required by EMSOs **data curation** and **data processing** systems at the regional data centre nodes. The data collected by the **data acquisition sub-system** are transmitted to the **data curation sub-system**, to be maintained and archived there.

### Data Curation




The EMSO **data curation sub-system** facilitates data curation, **quality control** and **preservation** of scientific data. It is operated at the data centres responsible for archiving the data acquired by the EMSO regional nodes. Three major data centres are currently offering these services for EMSO data: UniHB (PANGAEA), INGV (MOIST) and IFREMER (EUROSITES).


**Data import services** are provided by these institutions which either transfer the above mentioned data transmission format into an archival format or provide editorial tools and interfaces to ingest delayed mode data and laboratory analysis into their systems. Data which are intended to be transferred to the regional nodes data archives are quality checked, linked with an appropriate set of **metadata** according to international standards and persistently identified, depending on the archives internal standards and procedures. EMSO offers **catalogue services** and **metadata export services** for each regional node. The node systems PANGAEA and MOIST services based on metadata standards such as ISO19115, GCMD-DIF and extended Dublin Core, for EUROSITES data, metadata is extracted from NetCDF files via a central EMSO service. **Data export services** are not yet fully implemented at all EMSO nodes, however it is planned to provide



NetCDF export services for each node. The regional archives are responsible for cataloguing and long term preservation of these data that are provided for users via EMSO's *data access* and *discovery* subsystems.

### Data Access and Discovery


The EMSO  **data access sub-system** enables discovery and retrieval of data housed in data resources managed by a *data curation sub-system*. EMSO offers  **data discovery** via a common  **metadata catalogue** and web portal which can be visited at <http://dataportals.pangaea.de/emso>. The portal is based on the brokerage system panFMP (<http://www.panfmp.org>) and uses Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) or simple file transfer via FTP/HTTP to harvest metadata from EMSOs distributed regional node data archives and their archival systems PANGAEA, MOIST and EUROSITES.

The EMSO data portal offers machine-human as well as machine-machine search facilities and discovery services based on the collected  **metadata**. This includes a simple web-based user interface, a data search engine, which is offered at the EMSO data portal in a Google like style. In addition the data portal offers a common discovery service following the OpenSearch specification including the OpenSearch-Geo extension. A Open Geospatial Consortium (OGC) Catalogue Service for Web (CS-W) interface is currently under development.

A centralized data export service for these archived data is not implemented or planned, therefore, unless each EMSO data archive offers its own NetCDF data transformation service (see above) data requests are not yet processed by the EMSO data portal but are redirected to the hosting data archives which provide their own data access services for data retrieval.

Access to real time data is also offered via the EMSO data portal. EMSO has chosen to implement core standards of the OGC Sensor Web Enablement (SWE) suite of standards, such as Sensor Observation Service (SOS) and Observations and Measurements (O&M) to deliver real time data. These interfaces and formats are used to offer a common, web based SOS client which provides interactive visualizations of real time data.

### Data Processing

Centralized  **data processing sub-systems** that aggregate the data from various resources and provides computational capabilities and capacities for conducting data analysis and scientific experiments are not yet implemented for EMSO. Once more regional EMSO nodes and their data archives support NetCDF data export, it has been envisaged to introduce data visualization and plotting services at the EMSO data portal following the ESONET example. However presently, data processing services such as visualization, mining, as well as statistical services, are exclusively provided by each regional node and its responsible data centre.

### Community Support

Centralized  **community support sub-system** services to  **manage**, control and  **track** users' activities and supports users to conduct their roles in communities are not yet implemented or planned for EMSO.

### Summary

The EMSO example demonstrates how to use the common language defined by the Reference Model in documentation to communicate with the RDA community.

It has been recognised there is a common challenge when communicating with external organisations or communities -- "*your 'model' is not my 'model', your 'data' is not my 'data'*". With a public accessible reference base, an external community who has little domain knowledge, such as the RDA, is able to understand the specific descriptions of EMSO by looking up the terminology in the Reference Model. In a way, using the Reference Model, the communication efficiency can be improved.

The ENVRI Reference Model provides a set of ready-to-use terminology, in principle:

- Terms in the Science Viewpoint can be used for describing requirements, use scenarios, and human activities;
- Terms in the Information Viewpoint for describing information objects handled in a system, their action types, constraints, states, and lifecycles; and
- Terms in the Computational Viewpoint for describing functionalities, computational components, interfaces and services.

A reader may have noticed there are some terms in the writing that are different from the ones linked back in the Reference Model. For example, "**Set-up** (... of each observatory)" is linked to "**Instrument Configuration**". The intention is to show that in practice, to pursue the fitness, significance or beauty of the writing, an author may use different vocabularies to express the same concept. However, one can link them to the related concepts and definitions in the Reference Model to indicate the precise meanings. In this sense, using the Reference Model is different from using a dictionary – referring to the Reference Model places more emphasis on conceptual relativity.

### Example 4: Using the Reference Model as design reference (EPOS)

#### Descriptions of the Example




Although it looks similar, this example provides a different perspective on the usage of the Reference Model to that of [Example 3](#).

The ENVRI Reference Model is characterised by being both an ontology and a model. While [Example 3](#) demonstrates how to make use of its ontological framework in documentation, in this example, we exploit its representation as a model, which enables structural thinking and is more useful in the construction of an infrastructure and the organisation of design activities.

The European Plate Observing System (EPOS) is the European integrated solid earth sciences research infrastructure; a long-term plan to integrate existing national research infrastructures for seismology, volcanology, geodesy and other solid earth sciences. One of EPOS' goals is to provide the technical and legal framework by which to automate discovery and access to datasets and services provided by existing national (and trans-national) research institutions and monitoring networks throughout Europe. Another goal is to provide a standard set of core services by which researchers and other interested parties can interact with the federated infrastructure independently of the any particular data centre or national infrastructure. By providing such a common service interface and federation of resources, EPOS will be able to provide greater access to data recorded by existing and future monitoring networks, laboratory experiments and computational simulations, and foster greater cross-disciplinary research collaborations.

EPOS was included in the European Strategy Forum on Research Infrastructures (ESFRI) Roadmap in December 2008 and is currently in its Preparatory Phase; EPOS is scheduled to enter its Construction and Operational Phase in 2015.

EPOS is an infrastructure that intends to integrate several existing infrastructures which in the past have generally been constructed on a national scale only. There already exist established data centres with established working practices and monitoring networks. The challenge for EPOS is to provide a lightweight service layer that can be placed over these existing established infrastructures whilst disguising the underlying heterogeneity of components; this challenge is at least partially mitigated by the existence of certain protocols and data formats that are already standard in some parts of (for example) seismology, and a general drive within EPOS to further extend standardisation throughout its constituent institutions -- though it is not clear how extensively this level of standardisation will apply to all of the (currently highly disparate) earth sciences covered by EPOS' remit.

It is intended that ENVRI contribute in some way to the design of the EPOS Core Services, whether by the production of useful tools (via ENVRI WP4) or by the application of the ENVRI Reference Model (ENVRI-RM) for infrastructure layout and design (via ENVRI WP3). Focusing on the latter, ENVRI-RM should be able to simplify the design problem by breaking it down into well-defined subsystems of components specified from different complementary viewpoints (principally  **Science**,  **Information** and  **Computation**).

## How to Use the Reference Model

Following the guidance of the  **ENVRI common subsystems** the EPOS design issues can be broken down as follows:

### Data Acquisition

Data acquisition is performed by EPOS' constituent 'client' infrastructures; existing monitoring networks and laboratories, collected by data centres and presented for discovery and access to the EPOS integration layer. Many of these client systems operate in real-time (for example the continuous data streams produced by seismograph networks), **requiring concurrently active data curation facilities (storage, persistent identification and metadata assignment).**

### Data Curation

Data is principally curated within existing data centres that publish their datasets according to some agreed protocol. These data centres have their own data collection policies, but EPOS intends to promote the adoption of common metadata in order to ease interoperability, based on a three-level model consisting of discovery metadata (using extended qualified Dublin Core) which is derived from contextual metadata (using CERIF, the Common European Research Information Format), which points to detailed metadata (domain-specific and associated with a particular service or resource). EPOS will also provide a global persistent identification mechanism for continuous data streams and discrete datasets (the latter possibly using the mechanisms produced by the EUDAT project).

### Data Access, Brokering and Processing

Given global persistent identification and metadata, as well as the use where possible of standard data formats, it is intended that tools be produced to search over and extract specific datasets from different sites based on geospatial (and other) requirements. This along with tools for modelling, processing, data mining and visualisation form the data-oriented integration layer of the EPOS Core Services. These sit atop the 'thematic layer' of the Services, which divide services by domain and forms (for example seismology, volcanology and geodesy as well as satellite data, hazard maps, geomagnetic observatories and rock physics laboratories).

Because EPOS is making a concerted effort to integrate data standards and services, the resultant infrastructure should be less reliant on the brokering model than otherwise expected; the homogenisation of resources means that it will not be so necessary to maintain interfaces between heterogeneous resources required to be interoperable.

EPOS also intends to provide access for researchers to high-performance computation facilities as provided by such infrastructure projects as PRACE.

### Community Support

EPOS intends to provide training facilities to its research demographic; it is as yet unclear if EPOS intends to provide any kind of 'social' aspect to its core services (annotation of datasets, record of individual researchers' interactions with the infrastructure, etc.). It is a goal however of EPOS to promote best practices and reward participation, as well as to increase the visibility of research results produced using EPOS services. This implies that community support will become an increasingly important aspect of the EPOS infrastructure as the basic integration challenge it faces becomes solved.

## Summary

Like EPOS, ESFRI Environmental Research Infrastructures are characterised as large-scale distributed complex systems involving numbers of organisations across different European countries. Design and implementations become large collaborative activities subject to change and are evolving, which bring significant challenges. Considering the difficulty of ensuring efficiency and productivity, it is not only what to do but how to do it that is important. We observe no approach is currently in use to assist the organisation of the design activities.

The ENVRI Reference Model captures common requirements of a collection of representative environmental research infrastructures, providing a projection of Europe-wide requirements they have, which in potential can be served as a technology roadmap to position and orchestrate collaborations in design and developments. It provides well-defined subsystems of components specified from different complementary viewpoints (Science, Information and Computation), which can help break down the complexity and simplify the design problems, enabling designers to deliver a practical architecture that leads to concrete implementations. It offers a descriptive framework for specifying uniform distributed systems, allowing designers from different organisations to carry out design activities in parallel.

## EPOS/ENVRI modelling

[Link to EPOS/ENVRI modelling notes.](#)

### EPOS/ENVRI Modelling


A (very) generic overview of EPOS data access and (RI-internal) processing in terms of ENVRI-RM computational objects.

Two current possible avenues of development:

- **Instantiation** – identifying specific services created or in the process of being created that are instances of the computational objects above.
- **Case study**– take a usecase (one out of FutureVolc?) that allows a story to be told about how the user could interact with data via EPOS in terms of the ENVRI-RM.

## Example 5: Using the Reference Model to explain the technology details of common services (WP4 practices)




### Descriptions of the Example

ENVRI working package 4 responses to deliver common services to support the constructions of ESFRI ENV RIs. Initially, the implementations focus on a  **data access subsystem** that supports integrated data discovery and access. In order to help ESFRI project managers, architects, and developers understand the design and implementation of these services, this example uses the terms and concepts from the Reference Model to explain the technology details of these services.




### How to Use the Reference Model




We start with the semantic harmonisation service developed by the team in Task 4.2 [39]. The development is conducted to support the [use case "Iceland Volcano Ash"](#). The goal is to support scientists to analyse Iceland behaviour using data provided by different research infrastructures during a specific time period.

#### Science Viewpoint

Defined by the Reference Model Science Viewpoint, the  **semantic harmonization** is a  **behaviour** belong to the  **data publication community**, which captures the business requirements of unifying similar data (knowledge) models based on the consensus of collaborative domain experts to achieve better data (knowledge) reuse and semantic interoperability.


#### Computational Viewpoint


A data publication community interacts with a  **data access subsystem** to conduct user roles. The computational specification of the data access subsystem is given in Figure 1. The model specifies a **data access subsystem** which provides  **data broker** 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).


**Figure 1:** Computational specification of data access subsystem

#### Definitions

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.

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.

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 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.

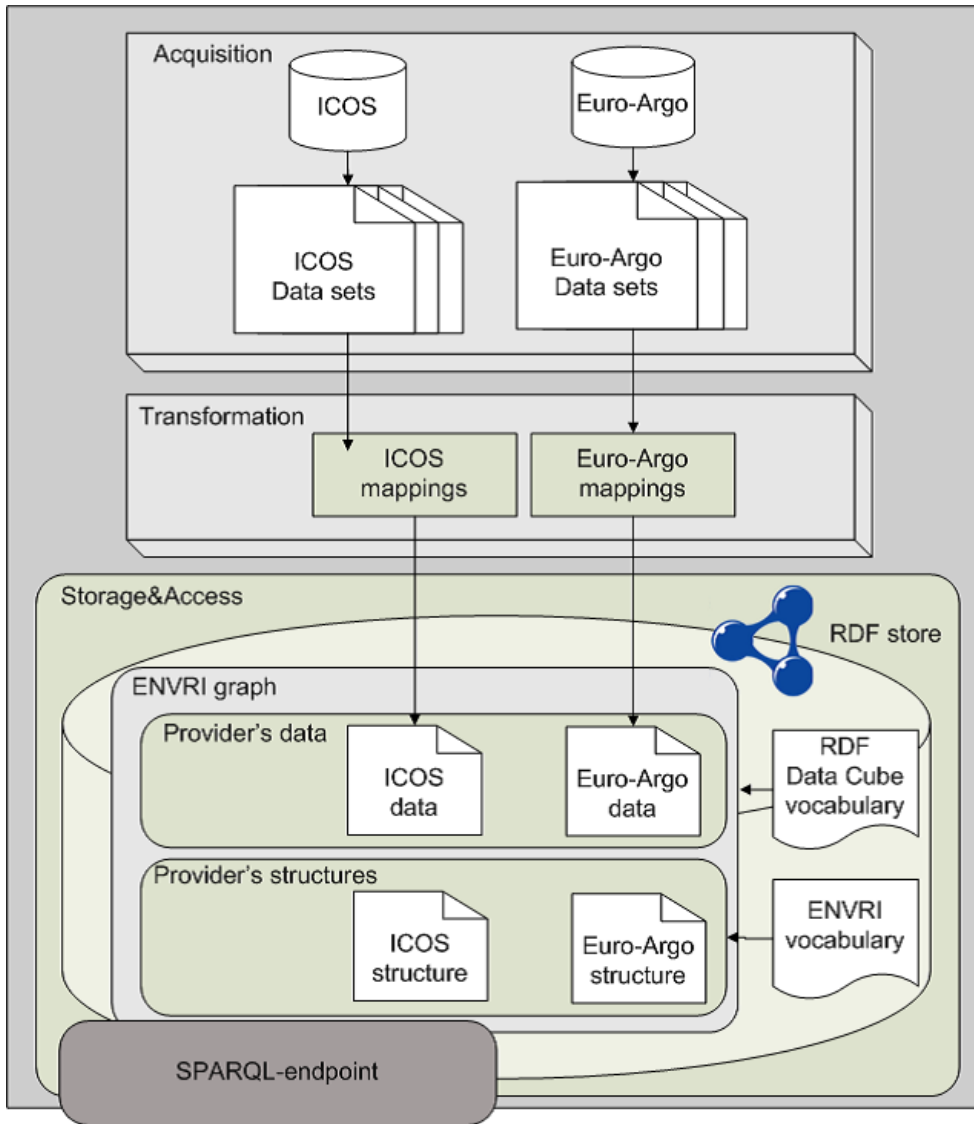
*Please click the links to find out the specification details of these computational objects and the interactions between them.*

The implementation conducted by WP4 T4.2 is an **instantiation** of the above computational objects specified in the Reference Model, that uses existing software components and developed approaches to enable integration and harmonization of data resources from cluster's infrastructures and publication according unifying views.

Figure 2 depicts the computational components deployed in the prototype implementation. The service receives users' requests via the [SPA RQL-endpoint](#). Then, it can automatically retrieve and integrate real measurement data collections from distributed data sources. The current prototype focuses on datasets from two different ESFRI projects:

- ICOS, which is organized by atmospheric stations which perform measurements of the CO<sub>2</sub> concentration in the air and
- EURO-Argo observations that were provided in separate collections grouped according to the float that performed measurements of the ocean temperature.





The prototyped service uses two semantic models to provide mapping between representations: the [RDF Data Cube vocabulary](#) and the [ENVRI vocabulary](#). The ENVRI vocabulary is derived from the OGC and ISO "Observations & Measurements" standard ([O&M](#)), [SWEET](#) and [GeoSparql Vocabulary](#).



**Figure 2:** The Deployed service components for semantic harmonization [39]

Table 1 provides the mapping between Reference Model computational objects and the deployed service components. Among them, the *Transformation* component serves as a **data broker** to negotiate data access with data stores within heterogeneous research infrastructures. An (instance of the) **semantic broker** is implemented using the RDF store technology which provides the semantic mappings and translations.

**Table 1:** Mapping of the deployed service components to the Reference Model computational objects

RM Computational Objects	Deployed Service Components
 <b>Data Broker</b>	Transformation (ICOS mappings, EuroArgo Mappings)
 <b>Experiment Laboratory</b>	SPARQL-endpoint
 <b>Semantic Broker</b>	Provider's data (ICOS data, EuroArgo data) Provider's structures (ICOS structure, EurArgo structure)
 <b>Semantic Laboratory</b>	RDF Data Cube Vocabulary, ENVRI Vocabulary

In the following, we explain the design of the information model of the semantic harmonisation service.

### Information Viewpoint

Analysing the environmental data schema results in identifying the common structural concepts, the ENVRI vocabulary, which include the terms such as “metadata attributes”, “observation”, “dataset”. Data retrieved from the different sources are firstly mapped to this uniform semantic model. Figure 3 gives two examples,

and shows how datasets of ICOS and EuroArgo can be mapped to the ENVRI vocabulary, respectively.

Metadata Attributes							Observations		Dataset
Site	Year	Month	Day	Hour	Date	CO2	SD	Flag	
mhdall	2011	1	1	0	2011.000000	400.135			0.526
mhdall	2011	1	1	1	2011.000114	399.893			0.670
mhdall	2011	1	1	2	2011.000228	401.878			1.073
mhdall	2011	1	1	3	2011.000342	400.474			0.499
mhdall	2011	1	1	4	2011.000457	402.787			2.611
mhdall	2011	1	1	5	2011.000571	406.205			3.125

Metadata Attributes							Observations		Dataset
PLATFORM	ARGOS_ID	DATE	LAT	LONG	PRES (decibar)	TEMP (degree_C)			
4900679	37751	2010-03-01T03:24:00Z	48.817	-31.648	4.4	11.595			
4900679	37751	2010-03-01T03:24:00Z	48.817	-31.648	8.8	11.617			
4900679	37751	2010-03-01T03:24:00Z	48.817	-31.648	19.1	11.629			
4900679	37751	2010-03-01T03:24:00Z	48.817	-31.648	29.3	11.656			
4900679	37751	2010-03-01T03:24:00Z	48.817	-31.648	38.4	11.566			
4900679	37751	2010-03-01T03:24:00Z	48.817	-31.648	48.8	11.649			
4900679	37751	2010-03-01T03:24:00Z	48.817	-31.648	58.3	11.638			

**Figure 3:** Datasets as provided by ICOS (above) with CO2 concentrations and by EURO-Argo (below) with ocean temperature measurements

Semantic mappings are based on observation statements. For example, the following observation statement declares the measurements about “air”:

*“Observation of the CO2 concentration in samples of air at the Mace Head atmospheric station which is located at (53\_20'N, 9\_54'W): CO2 concentration of the air 25m above the sea level on Jan 1st, 2010 at 00:00 was 391.318 parts per million”.*

“Air” is represented as the concept of air in GEneral Multi-lingual Environmental Thesaurus (GEMET) by assigning the URI to it (entity naming). The GEMET concept of air is then defined as an instance of envri:FeatureOfInterest (entity typing).

The mapping rules are specified by using the Data cube plug-in for Google Refine. The mappings are executed to obtain RDF representations of the source data files. As such they are uploaded to the [Virtuoso OSE RDF store](#) and are ready to be queried at a SPARQL-endpoint.

The data harmonization process described above is captured by the Reference Model. As shown in Figure 4, the Information Viewpoint models the mapping of data according to **mapping rules** which are defined by the use of **local** and **global conceptual model**. Ontologies and thesauri are defined as **conceptual models**, and those widely accepted models such as, GEMET, O&M, Data Cube, are declared **global conceptual models** whereas the ENVRI vocabulary is specified as a **local** one, because it has been developed within the current project without being yet accepted by a broad community.

**Figure 4:** The RM Information specification related to the semantic harmonisation









Describing a process using the ENVRI Reference Model concepts is to instantiate the concepts that can be mapped to the process. Figure 5 illustrates the instantiation (all boxes with a dashed line) of the ENVRI Reference Model concepts focusing at the harmonization process described above. The same could be demonstrated for the EuroArgo dataset with the feature of interest being ocean. For each part of the observation mapping rules have to be defined to be able to query both datasets at a certain time period.







**Figure 5:** Mapping of the deployed information model with that of the the Reference Model

The tables below show the mapping between the harmonisation process and the concepts in the ENVRI RM information viewpoint. The example shows that both bottom up (from the applied operation to the model description) and top down approaches (from the model definitions back to the applied solution) can lead to a better understanding of the Reference Model itself and of how components should work properly in a complex infrastructure.


**Table 2:** Mapping between the Reference Model  **Information objects** and those in the deployed service

Information Object in RM	Component/Object in Task 4.2
 Specification of measurement and/or observation	Observation of the CO2 concentration in samples of <b>air</b> at the Mace Head atmospheric station which is located at (53_20'N, 9_54'W): CO2 concentration of the air 25m above the sea level on Jan 1st, 2010 at 00:00 was 391.318 parts per million
 Mapped data	GEMET:245 is instance of FeatureOfInterest class
 Global conceptual model	GEMET, O&M, DataCube
 Local conceptual model	ENVRI vocabulary
 Local concept	FeatureOfInterest (ENVRI vocabulary)
 Global concept	Component Property, GEMET:245, FeatureOfInterest (O&M)
 Mapping rule	GEMET:245 create as instance of FeatureOfInterest class
 Published data	ICOS data CO2 of air, EuroArgo data ocean temperature

**Table 3:** Mapping between the Reference Model  **Action Types** and those in the deployed service

Information Action Tyoes in RM	Operation in Task 4.2
 Build local conceptual model	Build ENVRI vocabulary as extension of DataCube and on basis of O&M concepts
 Setup Mapping rule	Define rule: GEMET:245 create as instance of FeatureOfInterest class
 Perform Mapping	Perform Mapping using Google Refine
 Query Data	SPARQL query: <a href="http://staff.science.uva.nl/~ttaraso1/html/queries/Q1.rq">http://staff.science.uva.nl/~ttaraso1/html/queries/Q1.rq</a>

## Summary

This example demonstrate the feasibility of the design specifications of the reference model. Instances of selected model components can be developed into common services, in this case, a  **data access subsystem** that supports integrated data discovery and access. Data products from different environmental research infrastructures including, measurements of deep sea, upper space, volcano and seismology, open sea, atmosphere, and biodiversity, can now be pulled out through a **single data access interface**. Scientists are using this newly-available data resource to study environmental problems previously unachievable including, the study of the climate impact caused by the eruptions of the Eyjafjallajökull volcano in 2010.

END.

