ENVRI Common Operations of Environmental Research Infrastructures

ENVRI

Services for the Environmental Community

D3.5 Guidelines of Using ENVRI Reference Model

Document identifier:	D3.5 Guidelines of Using ENVRI Reference Model
Date:	**31/10/2013**
Activity:	**WP3**
Lead Partner:	**CU**
Document Status:	**FINAL**
Dissemination Level:	**PUBLIC**
Document Link:	http://tinyurl.com/o5l7bn9

ABSTRACT

The ENVRI Reference Model, hosted at www.envri.eu/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. This guideline document has been created to guide users of the ENVRI Reference Model to explore the model and help them understand the concepts defined, so as to better apply the model to daily practices.

SEVENTH FRAMEWORK PROGRAMME

1. COPYRIGHT NOTICE

2. DELIVERY SLIP

	Name	Partner/Activity	Date
From			
Reviewed by	Ari J Asmi, Yannick Legre	UHEL CNRS	22/Oct/2013 09/Oct/2013
Approved by			

3. DOCUMENT LOG

Issue	Date	Comment	Author/Partner
1.0	30/09/13	Draft	Yin Chen, Paul Martin Alex Hardisty, Alun Preece Barbara Magagna, Herbert Schentz Zhiming Zhao, Robert Huber Ingemar Haggstrom, Ville Savolainen Malgozata Krakowian
2.0	31/10/13	Internally reviewed version to be approved by project management and submitted to the Commission	Yin Chen, Barbara Magagna, Paul Martin, Alex Hardisty, Alun Preece Herbert Schentz, Zhiming Zhao, Robert Huber, Ingemar Haggstrom, Ville Savolainen, Malgozata Krakowian

4. APPLICATION AREA

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

5. DOCUMENT AMENDMENT PROCEDURE

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

6. TERMINOLOGY

A complete project glossary is provided at the following page: http://www.ENVRI.eu/glossary.

7. PROJECT SUMMARY

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

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

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

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

8. EXECUTIVE SUMMARY

The ENVRI Reference Model provides the ESFRI Environmental Research Infrastructures with a common ontological framework and standards for the description and characterisation of computational and storage infrastructures in order to achieve seamless interoperability between the heterogeneous resources of different infrastructures.

This guideline is prepared to help Reference Model users to understand key model concepts and map the abstraction to concretions. It uses a set of practical examples to illustrate various ways of the ENVRI Reference Model. 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.

Using a number of examples, we will show that the Reference Model can benefit a ESFRI Environmental Research with:

- **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.

TABLE OF CONTENTS

1 INTRODUCTION

The development of the ENVRI Reference Model, hosted at www.envri.eu/rm, provides the ESFRI Environmental Research Infrastructures (ESFRI ENV RIs) 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 the primary audience[1] within the community of ESFRI ENV RIs. 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 Reference Model.*
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.

[1] See the Intended Audience of the Reference Model at: http://tinyurl.com/ksjmjzq

With limited resources, only a few examples are included; these will be extended when more resources are available for future investigations.

The rest of the document is arranged as follows: we firstly revisit the key concepts of the Reference Model in Section 2, then introduce a set of examples in Section 3, and we summarise our work in Section 4.

2 REVIEW OF THE ENVRI REFERENCE MODEL

2.1 The Basis

The ENVRI Reference Model (ENVRI-RM) is built using the Open Distributed Processing (ODP) framework, an international standard for distributed system specification published by ISO/IEC (ISO/IEC 10746-1, 1998). ODP provides an overall conceptual framework for specifying large or complex computing systems. It adopts the **object modelling** approach, and defines five specific **viewpoints** – abstractions that yield specifications of the whole system related to particular sets of concerns. The five viewpoints are:

- 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 the environmental science community, we refer to this in the ENVRI-RM as the *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.

The reasons for adopting ODP in ENVRI include:

- It provides a descriptive framework for specifying and building large or complex system that consist 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;
- It enables large collaborative design activities. ODP breaks down a complex design specification into separated but interlined viewpoints, which allows designers in different teams from different organisations to work in parallel and to deliver uniform specifications;
- There is a natural fit between the ODP viewpoints and interoperability requirements in e-science (Zhao 2012).
- Being an international standard, ODP offers authority and stability.

In the next subsection, we revisit key contents of the model, which are described using the ODP terms and concepts.

2.2 The Reference Model

The development of the reference model is based on a preliminary study of a collection of the representative environmental research infrastructures (Chen 2013). By examining their computational characteristics, 5 common *subsystems*[2] have been identified: *Data Acquisition*, *Data Curation*, *Data Access*, *Data Processing* and *Community Support*. The fundamental reason of the division of the 5 subsystems is based on the observation that all applications, services and software tools are designed and implemented around 5 major physical resources: the sensor network, the storage, the (internet) communication network, application servers and client devices. The definitions of the five *subsystems* are given below:

- **Data acquisition:** collects raw data from sensor arrays, various instruments, or human observers, and brings the measurements (data streams) into the system.
- **Data curation**: facilitates quality control and preservation of scientific data. It is typically operated at a data centre.
- **Data access:** enables discovery and retrieval of data housed in data resources managed by a *data curation subsystem*.
- **Data processing:** aggregates the data from various resources and provides computational capabilities and capacities for conducting data analysis and scientific experiments.
- **Community support:** manages, controls and tracks users' activities and supports users to conduct their roles in communities.

The relationships between *subsystems* are depicted in Figure 2.1.

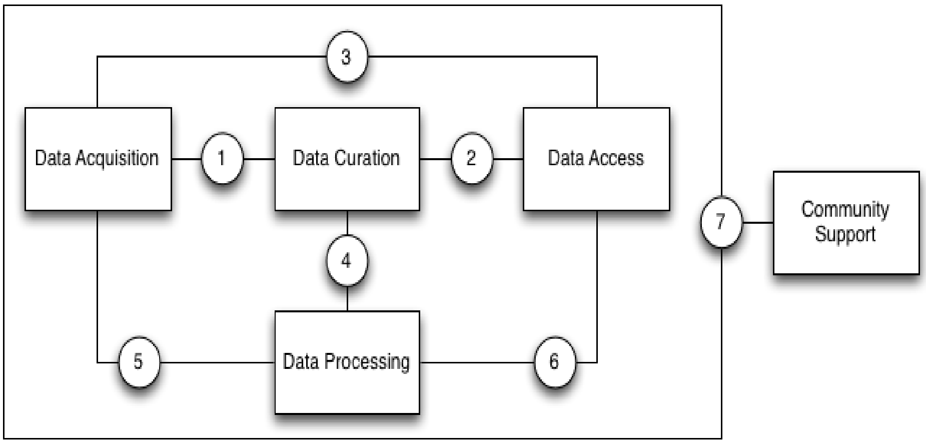


Figure 2.1: Illustration of the major points-of-reference between different subsystems

[2] Here, we define *subsystem* as a set of capabilities that collectively are defined by a set of *interfaces* with corresponding operations that can be invoked by other subsystems. An *interface* in ODP is an abstraction of the behaviour of an object that consists of a subset of the interactions of that object together with a set of constraints on when they may occur (Linington 2012).

Amongst the five *subsystems* can be identified seven major points-of-reference wherein interfaces between *subsystems* can be implemented. These points-of-reference are as follows:

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

Depending on the distribution of resources in an implemented infrastructure, some of these reference points may not be present in the infrastructure. They take particular importance however when considering scenarios where a research infrastructure delegates subsystems to other client infrastructures. For example, EPOS[3] and LifeWatch[4] both delegate data acquisition and some data curation activities to client national or domain-specific infrastructures, but provide data processing services over the data held by those client infrastructures. Thus reference points 4 and 5 become of significant importance to the construction of those projects.

Analysis of the common requirements of the six ESFRI ENV RIs has resulted in the identification of a number of common functionalities. As shown in Figure 2.2, these functionalities can be partitioned amongst the five subsystems. They encompass a range of concerns, from the fundamental (e.g. data collection and storage, data discovery and access and data security) to more specific challenges (e.g. data versioning, instrument monitoring and interactive visualisation).

The Minimal Model

In order to better manage the range of requirements, and in order to ensure rapid publication of incremental refinements to the ENVRI-RM, a *minimal model* has been identified which describes the fundamental functionality necessary to describe a functional environmental research infrastructure. The *minimal model* focuses on the major interaction links from raw data acquisition to the access and export of specific curated datasets, passing through stages of curation, brokering and authorisation. This core interaction chain represents the most fundamental contract between the archetypical research infrastructure and its community -- the access to scientific observations/measurements. The core interactions between data curation and data processing, as well as uploading of contributions from outside the infrastructure are also present in the *minimal model*, providing the skeleton by which additional extensions to the reference model can be attached, including alternative mechanisms for data retrieval and presentation. By initially focusing on this *minimal model*, it

[3] EPOS: http://www.epos-eu.org/

[4] LifeWatch: http://www.lifewatch.com/

then becomes practical to produce a partial specification of the ENVRI-RM which nonetheless reflects the final shape of the ENVRI-RM without the need for significant refactoring. Further development of the ENVRI-RM will focus on designated priority areas based on feedback from the contributing ESFRI representatives.

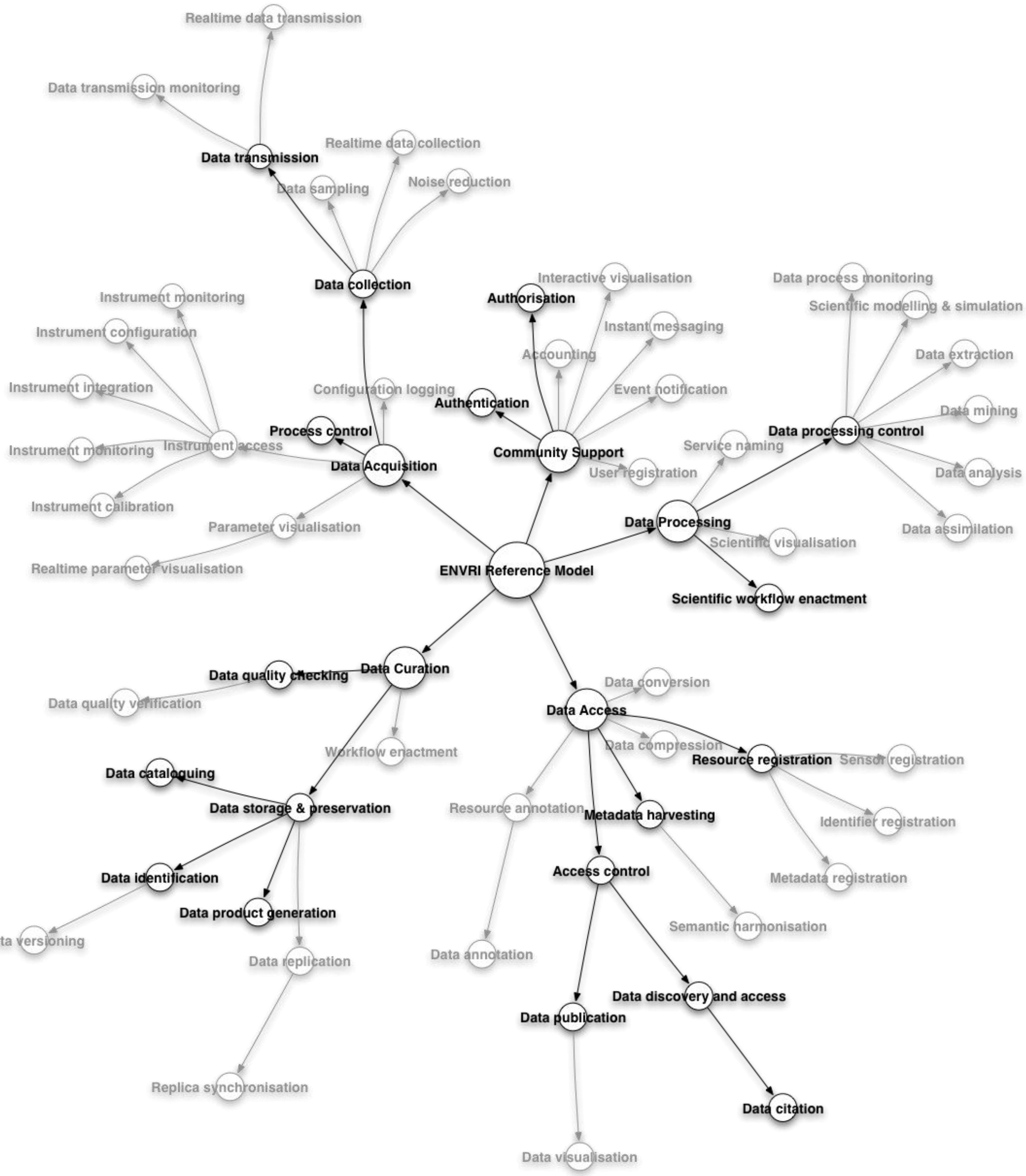


Figure 2.2: Radial depiction of ENVRI-RM requirements with the minimal model highlighted[5]

[5] The definitions of the functionalities are given at the reference model wiki site: http://miniurl.com/92Mz.

The ENVRI-RM subsystems are specified using the ODP standard framework. The ENVRI-RM defines an 'archetypical' environmental research infrastructure rather than a specific (implemented) infrastructure. Three viewpoints take particular priority: the *Science*, *Information* and *Computational* Viewpoints, which gives better focus on the core objective of ENVRI: to develop an understanding of the common requirements and to provide the design solutions for common data and operation services.

2.2.1 Science Viewpoint

The **Science Viewpoint** of the ENVRI-RM intends to capture the requirements for an environmental research infrastructure from the perspective of the people who perform their tasks and achieve their goals as mediated by the infrastructure. The key concepts defined in this viewpoint include **communities** and their **roles** and **behaviours**. 5 common communities are specified in according to the 5 subsystems: *data acquisition*, *data curation*, *data publication*, *data service provision*, and *data usage*. The definition of the communities are based on community objectives:

- **Data Acquisition** Community, who collect raw data and bring (streams of) measurements into an infrastructure;
- **Data Curation** Community, who curate the scientific data, maintain and archive them, and produce various data products with metadata;
- **Data Publication** Community, who assist data publication, discovery and access;
- **Data Service Provision** Community, who provide various services, applications and software/tools to link and recombine data and information in order to derive knowledge;
- **Data Usage** Community, who make use of data and service products, and transfer knowledge into understanding.

By analysing common requirements, use scenarios for each community are derived, community *roles* and *behaviours* are identified[6].

2.2.2 Information Viewpoint

The **Information Viewpoint** provides a common abstract model for the shared information handled by the infrastructure. The focus lies on the information itself, without considering any platform-specific or implementation details. It is independent from the computational interfaces and functions that manipulate the information or the nature of technology used to store it. It specifies the types of the information objects and the relationships between those types and how the states of these objects change as results of computational operations.

Modelling in this viewpoint in the ENVRI context employs a data-oriented approach which follows the lifecycle of scientific data (from raw to published and processed data) as information objects in each subsystem identifying their behaviour changes when events or action occur. The model captures common issues challenging many environmental research infrastructures such as, data enrichment including attribution of unique identifiers necessary for unambiguous identification and tracking of data provenance, association of metadata, semantic annotation, quality assessment, semantic mapping, and data discovery. The model

[6] The definitions of these concepts can be found at www.envri.eu/rm.

has been continuously refined by examining the feasibility of implementations and applying community feedback.

The resulting model consists of a set of **information objects** managed and processed by the common subsystems, a set of **action types** which are events that cause the states changes of the information objects, and a set of constraints on these objects. The model also specifies the **static schemata** and the **dynamic schemata**. The *static schemata* defines instantaneous views of the information objects at a certain stage of the data lifecycle defining a minimum set of constraints for data sharing. On the other hand, the *dynamic schemata* captures how the information object evolve as the system operates, specifying the allowable state changes as the effects of the actions. For example, Figure 2.3 shows a simple *dynamic schemata*, which specifies allowable *states* (represented by boxes) of an information object after applying certain *actions types* (represented by arrows between states).

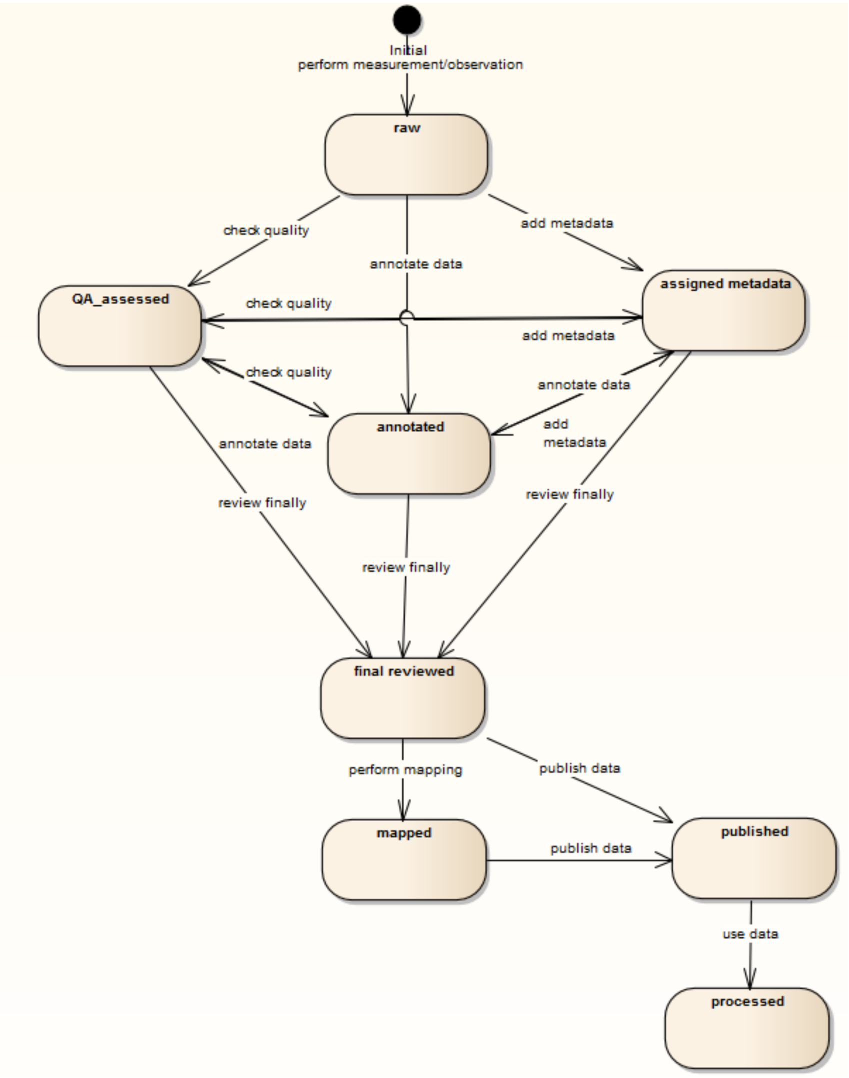


Figure 2.3: A simple example of *Dynamic Schemata*

2.2.3 Computational Viewpoint

The **Computational Viewpoint** of the ENVRI-RM accounts for the major computational objects expected within an environmental research infrastructure and the interfaces by which they interact. Each object encapsulates functionality implemented by a service or resource within the infrastructure. This encapsulation occurs at the conceptual level rather than the implementation level; it is admissible for the functions of a given object to be distributed across multiple computational resources in an implemented infrastructure, should that suit the infrastructure's physical architecture. Each object provides a number of interfaces by which functions can be invoked on the object, or by which the object can invoke the functions of other objects. By linking client and server interfaces, a network of interactions between objects can be built that demonstrates the computational dependencies of different parts of an infrastructure. These bindings can then be further specified in order to determine the particular operations and information streams supported by the interaction between interfaces, as well as the information objects that must be present.

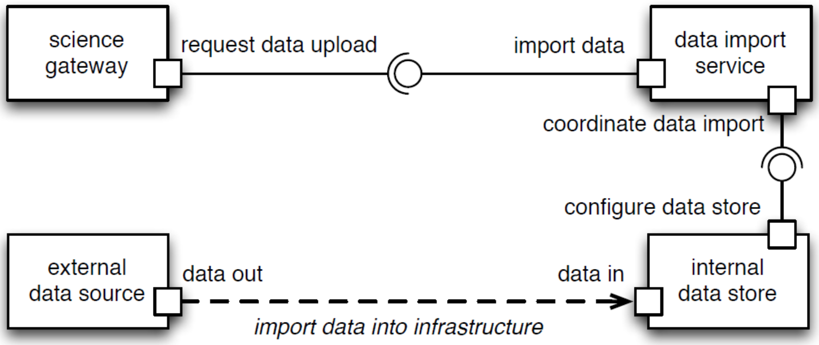


Figure 2.4: An example of interactions between interfaces of computational objects. Operations are represented by ball-and-socket connections; the client provides the socket while the server provides the ball. Data streams are represented by thick dashed lines in the direction of the data-flow.

For example a (simplified) brokered upload interaction might take the form illustrated in Figure 2.4. Four computational objects are identified: the *science gateway* encapsulating user-afforded functionality; the **data import service** handling import requests into the infrastructure; an **internal data store** controller managing access to a particular data store in the infrastructure; and an **external data source** controller from which data is to be extracted. In this instance, the **data import service** manages access via its **import data** and **coordinate data import** operational interfaces, responding to a request from the science gateway and invoking a selected data store respectively; this exchange of requests between objects can be further specified using, for example, a suitable UML sequence diagram. Once the data transfer has been validated and configured, the data can be pulled from the data source to the data store via compatible stream interfaces.

Each of the five essential *subsystems* of the ENVRI-RM must provide a number of computational objects of the kind illustrated above to be distributed across an infrastructure's technical architecture. For each of those objects, suitable interfaces must be identified and the most important interactions between those interfaces described. In the ENVRI-RM, the interfaces between *subsystems* are given particular attention, as many critical functions intercede in the movement of data between *subsystems*.

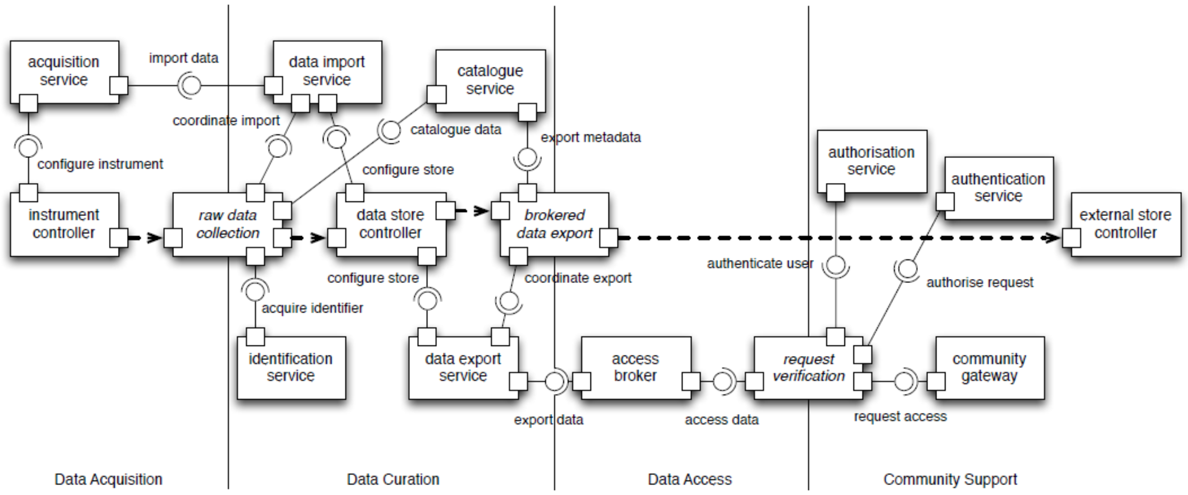


Figure 2.5: A subset of the core interactions involved in the acquisition and access of data. Using the notation of Figure 2.3, as observed in the majority of studied cases. The chains of bindings between objects indicate the intermediaries involved in interactions between objects in different subsystems.

Figure 2.5 illustrates the computational objects involved in basic data acquisition, curation and access, positioned with respect to four of the five research infrastructure *subsystems*. Client/server interface labels have been merged for clarity. Multi-party interactions are coordinated via *binding objects*[7] (such as **raw data collection** and **brokered data export**) that serve to simplify such interactions by abstracting aside implementation-specific details of the coordination such as how information and control is passed between objects when three or more parties are involved.

Thus the archetypical research infrastructure is considered here as having a brokered, service-oriented architecture. Core functionality is encapsulated in a number of service objects that control various resources present in the infrastructure. Access to most of these services by external entities is overseen by various brokers that validate requests and provide an interoperability layer between heterogeneous components --- this is particularly important for federated infrastructures, which may not be able to enforce a core set of standards on all data and services being integrated.

2.3 Summary

The section review the ENVRI Reference Model which exists to illustrate common characteristics of environmental sciences research infrastructures and establishes a taxonomy of terms, concepts and definitions in order to provide a common language and understanding, promote technology and solution sharing and improve interoperability.

[7] A *binding object* is an ODP computational object, which supports a binding between a set of other computational objects (Linington 2012).

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

In the next section, we use a set of practical examples to illustrate the usage of the Reference Model.

3 EXAMPLES OF USING ENVRI REFERENCE MODEL

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 4), 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.)

3.1 Example 1: Using the Reference Model as a Research Tool to Guide Research Activities

3.1.1 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 *E*uropean *I*ncoherent *Scat*ter 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[8],

[8] EGI: www.egi.eu

PRACE[9], 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[10] and by the EGI Federated Cloud Task Force[11], or the data infrastructure to be provided by EUDAT[12] 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[13] 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 and EGI, 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;
- Analysis of EGI enabling services and construction of an integrated infrastructure;
- Identifying the gaps between the EGI generic service infrastructure with the functional requirements of the domain specific research infrastructure, EISCAT 3D.

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.

3.1.2 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 3.1, 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.

[9] PRACE: http://www.prace-project.eu

[10] HelixNebula the Science Cloud: http://helix-nebula.eu/

[11] EGI Federated Cloud: http://go.egi.eu/cloud

[12] EUDAT: www.eudat.eu

[13] June 2013 statistics

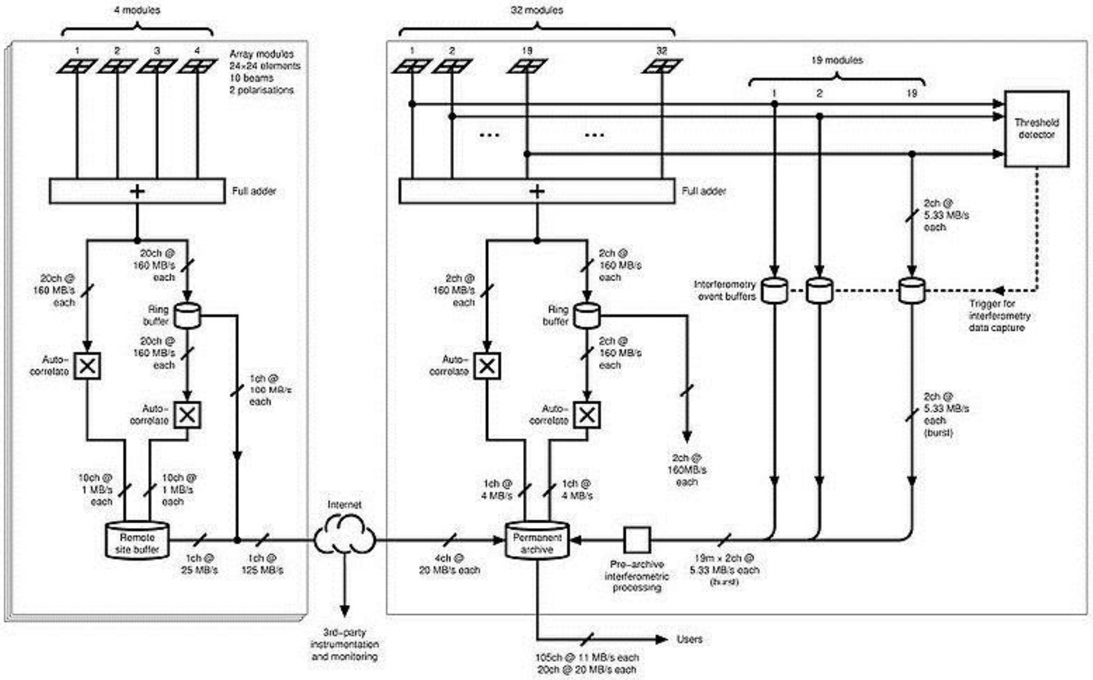


Figure 3.1: The original design of EISCAT 3D data infrastructure is embedded within the overall observatory system design

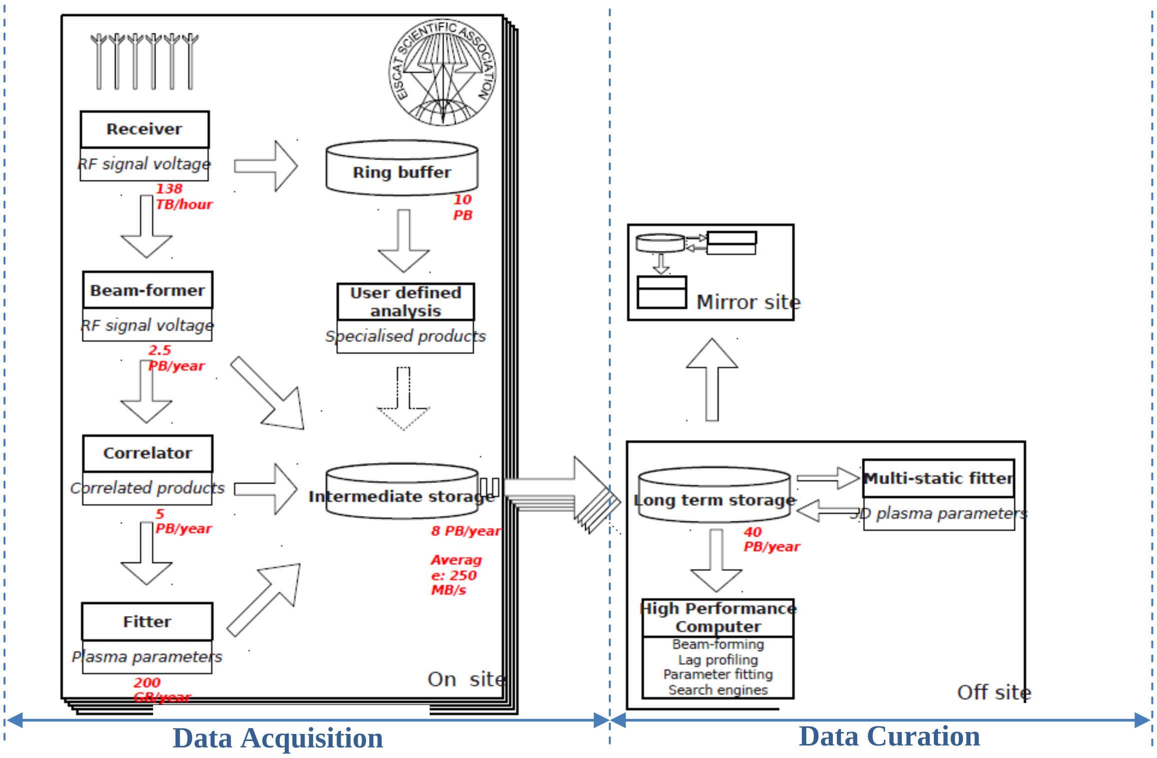


Figure 3.2: The EISCAT 3D data infrastructure maps to 2 of 5 ENVRI common subsystems. Using the 5 ENVRI Common Subsystem to interpret the EISCAT 3D data infrastructure makes it easy to communicate with computer scientists/technologists.

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 3.2. This diagram now, instead of Figure 3.1, is frequently used in presentations and discussions of the EISCAT 3D data infrastructure.

Figure 3.2 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 (including, acquisition and curation), we understand functionalities of the other 3 subsystems (access, processing, and community support) 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 3.1, 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 3.1: Mapping of the EGI Services to the Reference Model elements (from computational perspective)

EGI Services	ENVRI- RM Computational Objects	ENVRI Common Subsystem
AMGA Metadata catalogue	Catalogue service	Curation
LFC File catalogue	Catalogue service	Curation
Storage elements	Data store controller	Curation
File Transfer Service	Data transfer service	Curation
Portal for application development & hosting	Virtual laboratory	Community Support
Access control	Security service	Community Support

Above analysis gives clues to a solution for integrating the EGI technologies into the EISCAT 3D data infrastructure. Depicted in Figure 3.3, a secondary data curation subsystem (seen as the mirror site of the EISCAT 3D central archive in Figure 3.2) 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).

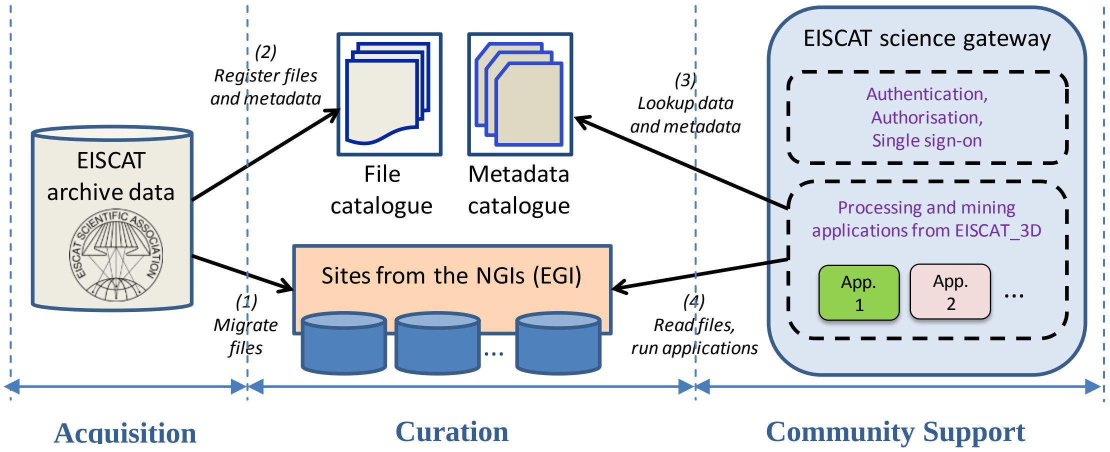

EISCAT archive data
EISCAT SCIENTIFIC ASSOCIATION
(2) Register files and metadata
File catalogue
Metadata catalogue
(3) Lookup data and metadata
EISCAT science gateway
Authentication, Authorisation, Single sign-on
Processing and mining applications from EISCAT_3D
App. 1
App. 2
...
Sites from the NGIs (EGI)
...
(1) Migrate files
(4) Read files, run applications
Acquisition
Curation
Community Support

Figure 3.3: 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 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 (Oracle 2012). 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 (Oracle 2012). Novel approaches are needed to discover meaningful insights through deep, complex search, e.g., using machine learning,

statistical modelling, and 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.

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 (see Example 2 below) of the Reference Model.

3.1.3 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 the EISCAT 3D data infrastructure and identified the 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 the domain specific research infrastructure, EISCAT 3D.

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.

In the next, we will explain a different way of using the Reference Model in the analysis of the EUDAT services.

3.2 Example 2: Using the Reference Model as An Analysis Tool

3.2.1 Descriptions 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[14] 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[15], PRACE[16]) and the networking layer (GEANT[17]).

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[18]), solid earth sciences (EPOS[19]), climate sciences (ENES[20]), environmental sciences (LIFEWATCH[21]), and biological and medical sciences (VPH[22]), 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[23]) 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.

3.2.2 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

[14] EUDAT: www.eudata.eu

[15] DEISA: http://www.deisa.eu

[16] PRACE: http://www.prace-ri.eu/

[17] GEANT: http://www.geant.net

[18] CLRTI: Common Language and Resource Technology Infrastructure: http://www.clarin.eu/

[19] EPOS: European Plate Observing System: http://www.epos-eu.org/

[20] European Network for Earth System Modelling: https://is.enes.org/

[21] LIFEWATCH: http://www.lifewatch.eu

[22] Virtual Physiological Human: http://www.vph-noe.eu/home

[23] iRODS: https://www.irods.org/

Reference Model, but is rather a source of implementations for instances of objects required by any RI implementing the Model.

Table 3.2: Mapping of the EUDAT Services to the Reference Model elements

EUDAT Services	ENVRI- RM Computational Objects	ENVRI Common Subsystem
Safe replication	Data Transfer Service	Curation
Staging	Data Importer	Curation
Metadata Catalogue	Catalogue Services	Curation
Simple Store	Data Store Controllers	Curation

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 a catalogue service (probably complementary to any catalogue service 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).

3.2.3 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).

3.3 Example 3: Using the Reference Model to Describe the Architectural Features of an Infrastructure

3.3.1 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)[24] community 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.

[24] RDA: https://rd-alliance.org/

3.3.2 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 3.4, according to the ENVRI-RM it includes the 4 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.

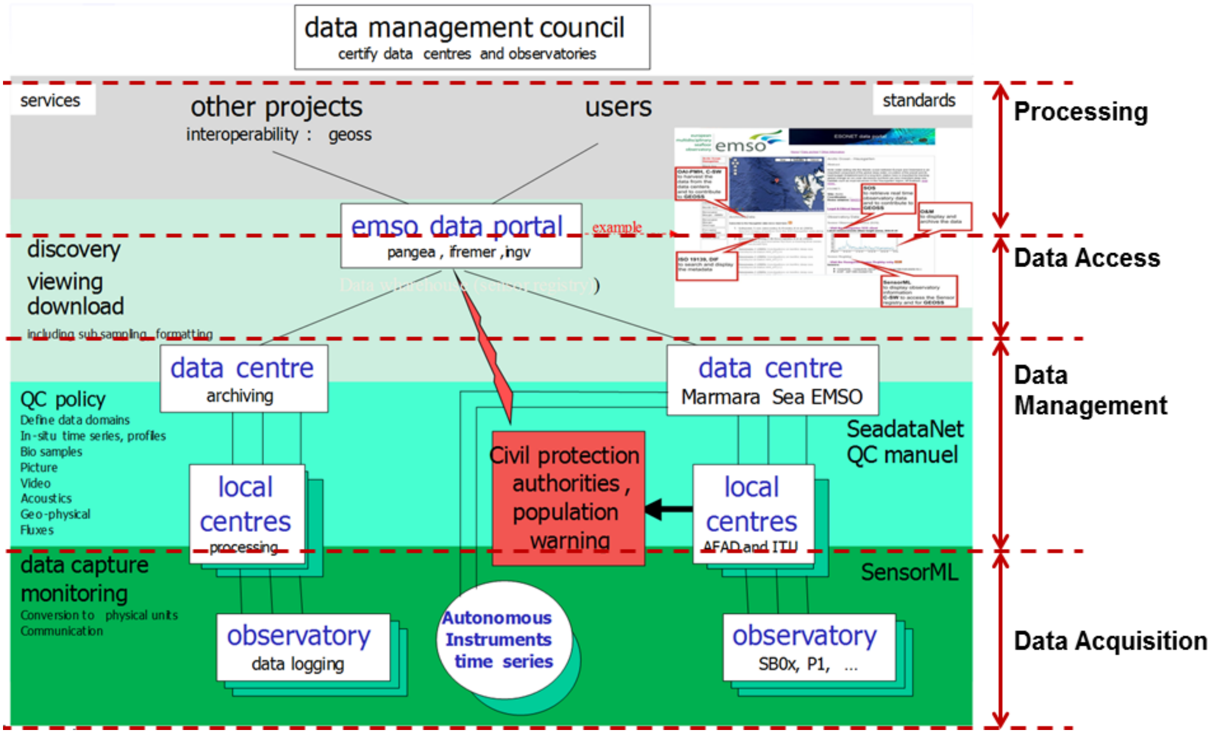


Figure 3.4: EMSO distributed data management system

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 acquired 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 EMSO's *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.*

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[25]. The portal is based on the brokerage system panFMP[26] and uses Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) or simple file transfer via FTP/HTTP to harvest metadata from EMSO's 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 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. An 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 that 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.

[25] EMSO web portal: http://dataportals.pangaea.de/emso

[26] panFMP: http://www.panfmp.org

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.

3.3.3 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 that has little domain knowledge, such as 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, constrains, 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 a single 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.

3.4 Example 4: Using the Reference Model as Design Reference

3.4.1 Descriptions of the Example

Although it looks similar, this example provides a difference 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 that 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**).

3.4.2 How to Use the Reference Model

Following the guidance of the ENVRI common subsystems the EPSO 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 interoperation, 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.

3.4.3 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 subjective 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 efficient 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.

3.5 Example 5: Using the Reference Model to Explain the Technology Details of Common Services

3.5.1 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.

3.5.2 How to Use the Reference Model

We start with the semantic harmonisation service developed by the team in Task 4.2 (Tarasova, 2013). The development is conducted to support the use case "Iceland Volcano Ash"[27]. 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 3.5.

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

[27] The basic scenario is described in http://staff.science.uva.nl/~ttaraso1/html/envri.html#model.

with data and processing services) and 💡 semantic laboratories (by which the community can update semantic models associated with the research infrastructure).

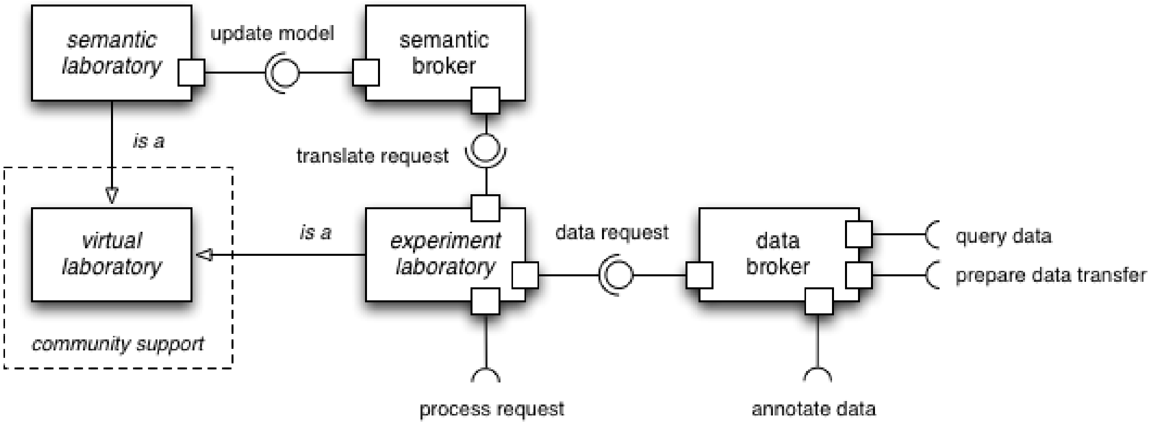


Figure 3.5 Computational specification of data access subsystem

☞ **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 3.6 depicts the computational components deployed in the prototype implementation. The service receives users' requests via the SPARQL-endpoint[28]. 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 CO2 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[29] and the ENVRI vocabulary[30]. The ENVRI vocabulary is derived from the OGC and ISO "Observations & Measurements" standard (O&M)[31] , SWEET[32] and GeoSparql Vocabulary[33].

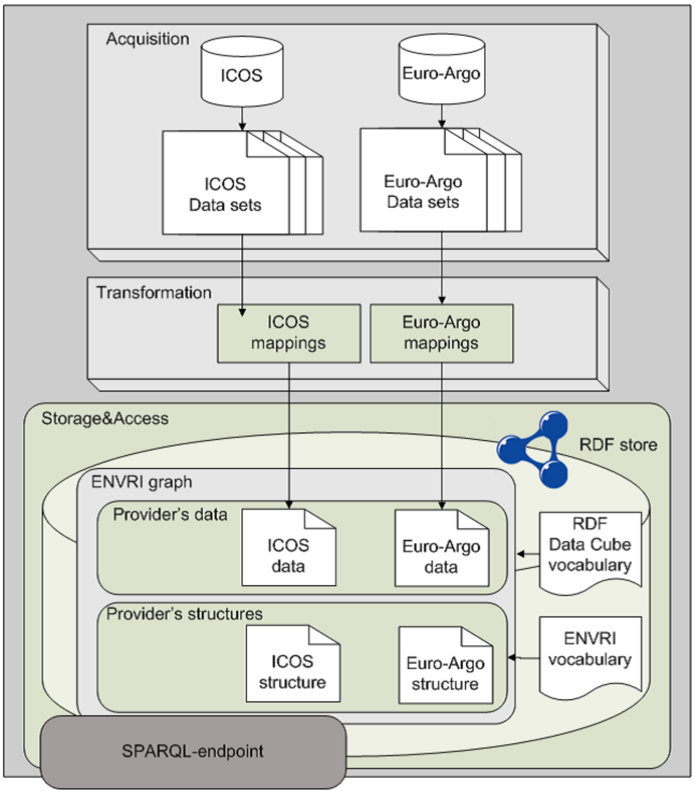


Figure 3.6: The Deployed service components for semantic harmonization (Tarasova, 2013).

[28] SPARQL-endpoint: http://data.politicalmashup.nl/sparql/

[29] RDF Data Cube Vocabulary: http://www.w3.org/TR/2013/

[30] ENVRI Vocabulary: http://data.politicalmashup.nl/RDF/vocabularies/envriCR-vocab-data-cube-20130625/

[31] OGC O&M Standard: http://www.opengeospatial.org/standards/om

[32] SWEET: http://sweet.jpl.nasa.gov/

[33] GeoSparql Vocabulary: http://www.opengeospatial.org/

Table 3.3 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.

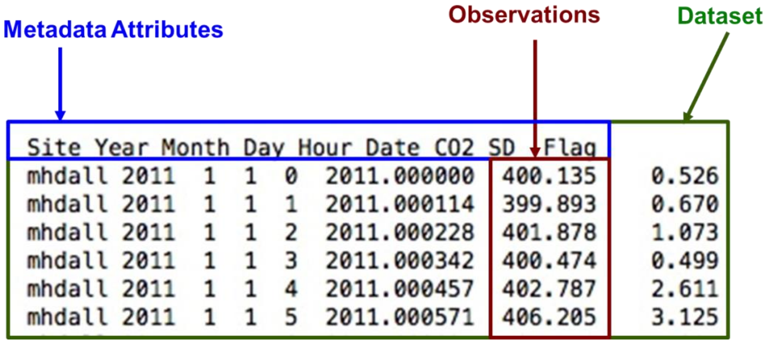
Table 3.3: Mapping of the deployed service components to the Reference Model computational objects

RM Computational Objects	Deployed Service Components
Data Broker	Transformation (ICOS mappings, EURO-Argo Mappings)
Experiment Laboratory	SPARQL-endpoint
Semantic Broker	Provider's data (ICOS data, EURO-Argo data) Provider's structures (ICOS structure, EURO-Argo structure)
Semantic Laboratory	RDF Data Cube Vocabulary, ENVRI Vocabulary

In the following, we explain the design of the information model of the semantic harmonisation service.

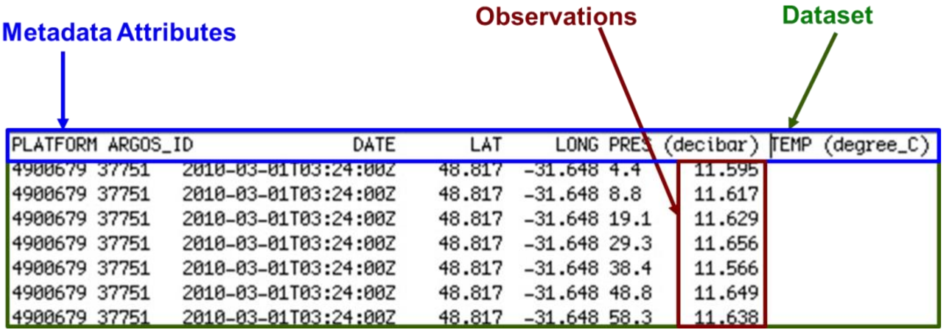
Information Viewpoint

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 3.7 gives two examples, and shows how datasets of ICOS and EURO-Argo can be mapped to the ENVRI vocabulary, respectively.



```
 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
```

(a) Datasets provided by ICOS with CO2 concentrations



```
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
```

(b) Datasets provided by EURO-Argo with ocean temperature measurements

Figure 3.7: Mapping of the datasets from ICOS and EURO-Argo to the ENVRI Vocabulary (terms in colours)

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[34] 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[35] 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 3.8, 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 model**s** 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.

Describing a process using the ENVRI Reference Model concepts is to instantiate the concepts that can be mapped to the process. Figure 3.9 illustrates the instantiation (all boxes with a dashed line) of the ENVRI Reference Model information model concepts focusing at the harmonization process described above. The same could be demonstrated for the EURO-Argo 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.

[34] http://www.eionet.europa.eu/gemet/concept?ns=1&cp=245

[35] http://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main/

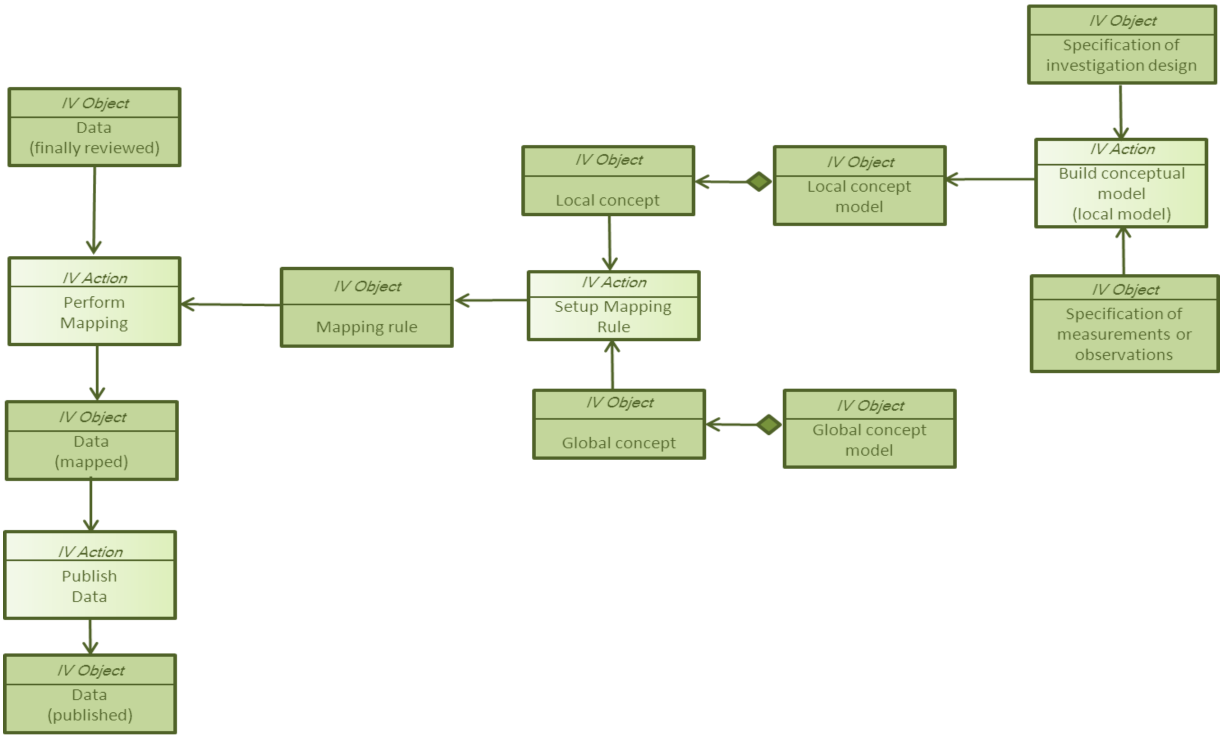


Figure 3.8: The RM Information specification related to the semantic harmonisation

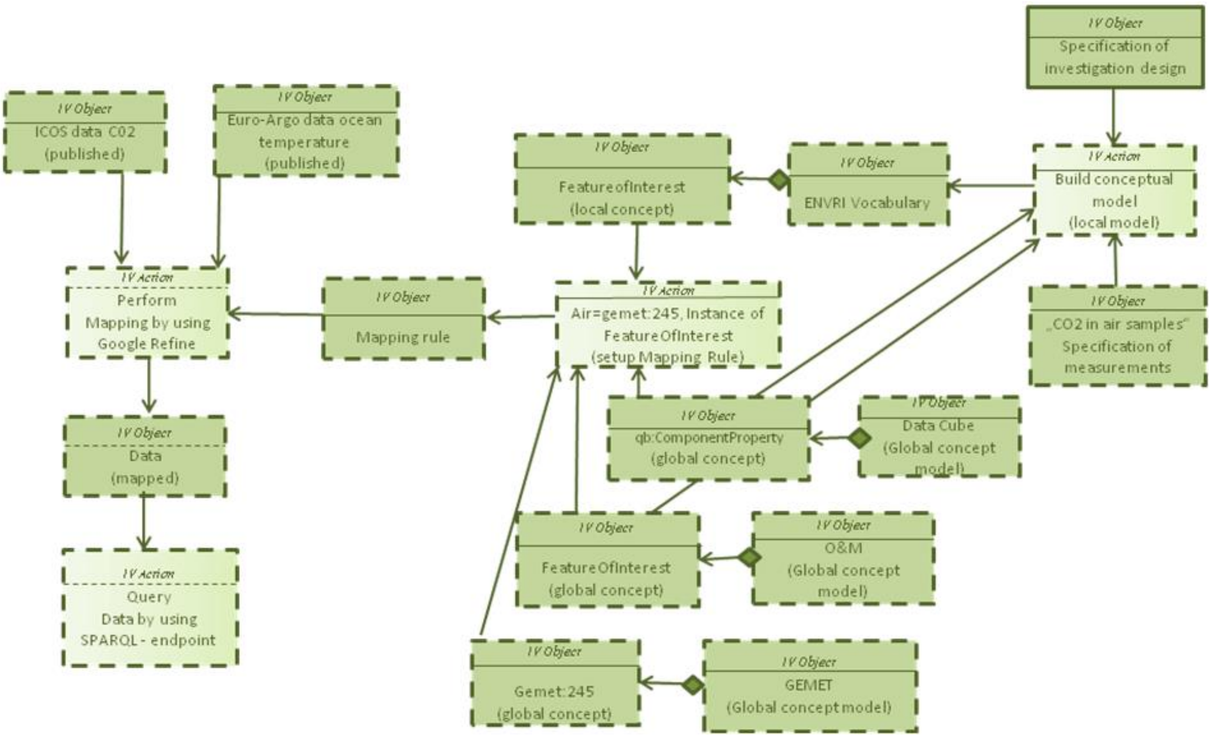

IV Object
Specification of investigation design
IV Object
ICOS data C02 (published)
IV Object
Euro-Argo data ocean temperature (published)
IV Object
FeatureofInterest (local concept)
IV Object
ENVRI Vocabulary
IV Action
Build conceptual model (local model)
IV Action
Perform Mapping by using Google Refine
IV Object
Mapping rule
IV Action
Air=gemet:245, Instance of FeatureOfInterest (setup Mapping Rule)
IV Object
„CO2 in air samples" Specification of measurements
IV Object
Data (mapped)
IV Object
qb:ComponentProperty (global concept)
IV Object
Data Cube (Global concept model)
IV Action
Query Data by using SPARQL - endpoint
IV Object
FeatureOfInterest (global concept)
IV Object
O&M (Global concept model)
IV Object
Gemet:245 (global concept)
IV Object
GEMET (Global concept model)

Figure 3.9: Mapping of the deployed information model with that of the Reference Model

Table 3.4 and 3.5 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 3.4: Mapping between the Reference Model **Information objects** and those in the deployed service

Information Object in RM	Component/Object in the Deployment
Specification of measurement and/or observation	For example: "*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*"
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, EURO-Argo data ocean temperature

Table 3.4: Mapping between the Reference Model **Action Types** and those in the deployed service

Information Action types in Reference Model	Operations Provided by the Deployed Services
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 [36]

[36] SPARQL query: http://staff.science.uva.nl/~ttaraso1/html/queries/Q1.rq

3.5.3 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** is implemented that supports integrated data discovery and access. Data products from different environmental research infrastructures, initially including ICOS and EURO-Argo, can now be pulled out through a single data access interface[37]. 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.

[37] ENVRI Data Portal: http://89.202.228.220/

4 CONCLUSION

In this guideline, we have used a set of practical examples to illustrate various ways of using the ENVRI Reference Model. We have chosen scenarios familiar to the ESFRI Environmental Research Infrastructures community, and provided analysis that may be of interest and benefit to their daily work. Throughout the document, we have incrementally introduced model concepts and terminology and guided the reader to explore 90% of the most important contents described by the Reference Model.

Using a number of examples, we have shown that by using the Reference Model, a ESFRI Environmental Research Infrastructure 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 continue the investigation of the big data pilot project with EISCAT 3D and report back new discoveries. We have also received the requests from ICOS, and we will help ICOS to apply the model concepts to their daily tasks, e.g., in various system analysis and design tasks, or to guide implementations. In a similar way, we will interact with other ESFRI ENV RIs. In particular, we are interested in how to use the Reference Model framework to improve the interoperability among the heterogeneous infrastructures.
- We will assist the development of the common services. A more detailed guide of how to use the Reference Model to drive the implementations (in WP4) will be provided.
- 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.

5 REFERENCES

Chen, Y. (et al) (2013): "Analysis of Common Requirements for Environmental Science Research Infrastructures", International Symposium on Grid & Clouds 2013, 17-22 Mar 2013, Taipei, Taiwan.

ISO/IEC 10746-1 (1998): "Information technology—Open Distributed Processing – Reference Model: Overview", ISO/IEC standard.

Linington, P. (et al) (2011): "Building Enterprise Systems with ODP: An Introduction to Open Distributed Processing", Chapman & Hall/CRC Press.

Zhao, Z. (et al) (2012): "OEIRM: An Open Distributed Processing Based Interoperability Reference Model for e-Science", Network and Parallel Computing. Springer Berlin Heidelberg 2012. 437-444.

Oracle (2012): "Oracle Information Architecture: An Architect's Guide to Big Data", An Oracle White Pater in Enterprise Architecture August 2012.

Tarasova, T. (et al) (2013): "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.

APPENDIX: TEMPLATE FOR COLLECTION OF THE EXAMPLES OF USING ENVRI REFERENCE MODEL

<Title>

1. What is this use case about?

Describe the purpose of the use case, and the background information.

2. How to use the reference model in this use case?

3. What's the result 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 of using the reference model in this use case?

Feedback from users