A Common Reference Model for Environmental Science Research Infrastructures

Yin Chen[1], Paul Martin[2], Barbara Magagna[3], Herbert Schentz[3], Zhiming Zhao[4], Alex Hardisty[1], Alun Preece[1], Malcolm Atkinson[2], Robert Huber[5], Yannick Legré[6]

Abstract

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, www.envri.eu/rm, uses the Open Distributed Processing standard framework in order to model the "archetypical" environmental research infrastructure. The use of the ENVRI-RM to illustrate common characteristics of European ESFRI environmental infrastructures from a number of different perspectives provides a common language for and understanding of environmental research infrastructures, promote technology and solution sharing between infrastructures, and improve interoperability between implemented services.

1. Introduction

Environmental issues will dominate the 21st century. Research infrastructures which provide advanced capabilities for data sharing, processing and analysis enable excellent research and play an ever-increasing role in the environmental sciences. The ENVRI project gathers 6 EU ESFRI[7] environmental science infrastructures (ICOS[8], EURO-Argo[9], EISCAT-3D[10], LifeWatch[11], EPOS[12], and EMSO[13]) in order to develop common data and software services. The results will accelerate the construction of these infrastructures and improve interoperability among them. The experiences gained from this endeavour will also benefit the building of other advanced research infrastructures.

[1] School of Computer Science & Informatics, Cardiff University, {ChenY58, Alex.Hardisty, A.D.Preece}@cs.cardiff.ac.uk
[2] Informatics, The University of Edinburgh, {pmartin, mpa}@staffmail.ed.ac.uk
[3] Environment Agency Austria, Austria, {Barbara.Magagna, Herbert.Schentz}@umwelbundesamt.at
[4] The University of Amsterdam, Netherlands, z.zhao@uva.nl
[5] University Bremen, Germany rhuber@wdc-mare.org
[6] Grid and Cloud Institute ,CNRS, France, yannick.legre@idgrilles.fr
[7] ESFRI, the European Strategy Forum on Research Infrastructures, is a strategic instrument to develop the scientific integration of Europe and to strengthen its international outreach.
[8] ICOS, http://www.icos-infrastructure.eu/, is a European distributed infrastructure dedicated to the monitoring of greenhouse gases (GHG) through its atmospheric, ecosystem and ocean networks.
[9] EURO-Argo, http://www.euro-argo.eu/, is the European contribution to Argo, a global ocean observing system.
[10] EISCAT-3D, http://www.eiscat3d.se/, is a European new-generation incoherent-scatter research radar for upper atmospheric science.
[11] LifeWatch, http://www.lifewatch.com/, is an e-science Infrastructure for biodiversity and ecosystem research.
[12] EPOS, http://www.epos-eu.org/, is a European Research Infrastructure on earthquakes, volcanoes, surface dynamics and tectonics.
[13] EMSO, http://www.emso-eu.org/management/, is a European network of seafloor observatories for the long-term monitoring of environmental processes related to ecosystems, climate change and geo-hazards.

The primary objective of ENVRI is to agree on a reference model for joint operations. The ENVRI Reference Model (ENVRI-RM) is a common ontological framework and standard for the description and characterisation of computational and storage infrastructures in order to achieve seamless interoperability between the heterogeneous resources of different infrastructures. Fundamentally the model serves to provide a universal reference framework for discussing many common technical challenges facing all of the ESFRI-environmental research 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.

The ENVRI Reference Model is based on the design experiences of state-of-the-art environmental research infrastructures, with a view to inform future implementation. It tackles multiple challenging issues encountered by many 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) approach (ISO/IEC 10746-1, 1998), which is an international standard for distributed system specification.

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 research infrastructures are in their planning / construction phase. The high costs 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 common computational requirements and the state-of-the-art design experiences of environmental sciences research infrastructures. It is the first reference model of this kind that can be used as a basis to inspire future research.
- The model establishes a taxonomy of terms, concepts and definitions, which provides a common language for communication to unify understanding. It serves as a community standard to secure interoperability.
- The model 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.

The rest of the paper is organised as follows: Section 2 briefly introduces the ODP approach; Section 3 describes the common concepts of the ENVRI Reference Model. Only key concepts are introduced due to the space limitation; Section 4 illustrates the usages of the reference model using examples; and Section 5 summarises the work.

2. ODP Approach

The 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.

It is worth noting that ODP defines a framework for specification, but not a methodology for modelling. In the next section, we start with our modelling method based on a study of requirements, then follow with the contents of the model, which are described using the ODP terms and concepts.

3. 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*[14] 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:

[14] 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).

- **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 1.

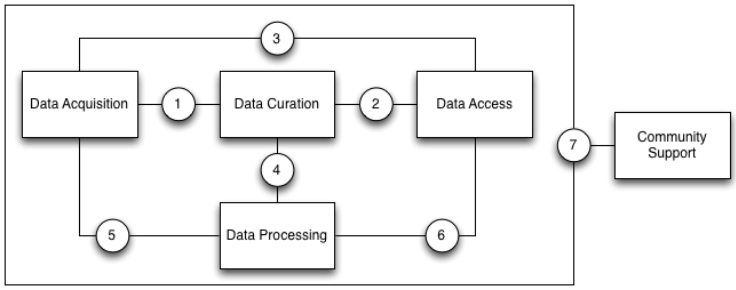

Figure 1: Illustration of the major points-of-reference between different subsystems

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[13] and LifeWatch[11] 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 environmental infrastructures has resulted in the identification of a number of common functionalities. As shown in Figure 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).

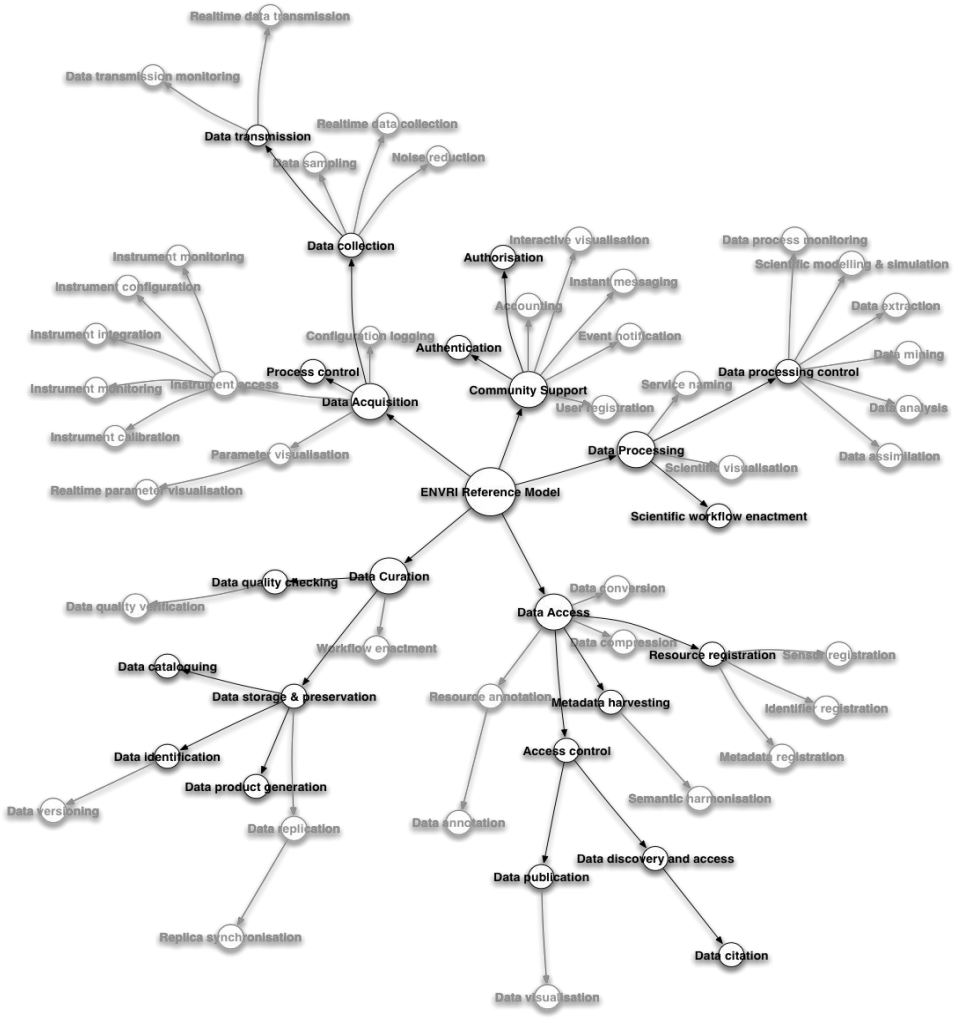

Figure 2: Radial depiction of ENVRI-RM requirements with the minimal model highlighted[15]

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 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 refactor-

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

ing. Further development of the ENVRI-RM will focus on designated priority areas based on feedback from the contributing ESFRI representatives.

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.

3.1 Science Viewpoint

The *Science Viewpoint* of the ENVRI-RM intends to capture the requirements for an environmental research infrastructure from the perspective of the people who perform their tasks and achieve their goals as mediated by the infrastructure. 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**, who collect raw data and bring (streams of) measurements into an infrastructure;
- **Data Curation**, who curate the scientific data, maintain and archive them, and produce various data products with metadata;
- **Data Publication**, who assist data publication, discovery and access;
- **Data Service Provision**, who provide various services, applications and software/tools to link and recombine data and information in order to derive knowledge;
- **Data Usage**, 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[16].

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

[16] Due to space limitation, the definitions of these concepts can be found at www.envri.eu/rm.

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 defines the *dynamic schemata* and the *static schemata*. The *dynamic schemata* captures how the information object evolve as the system operates, specifying the allowable state changes as the effects of the actions. On the other hand, 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.

3.3 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.

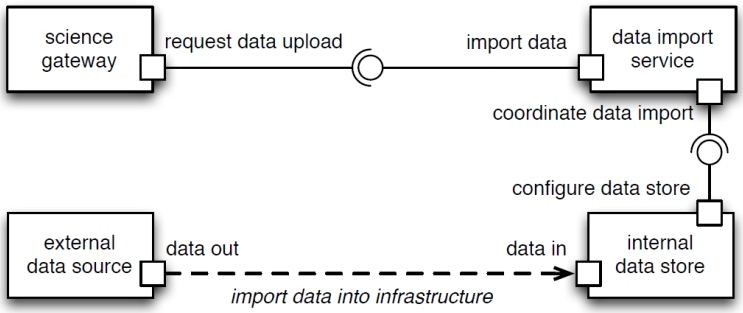

Figure 3: An example of interaction between interfaces of computational objects

For example a (simplified) brokered upload interaction might take the form illustrated in Figure 3. 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 4 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*[17] (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.

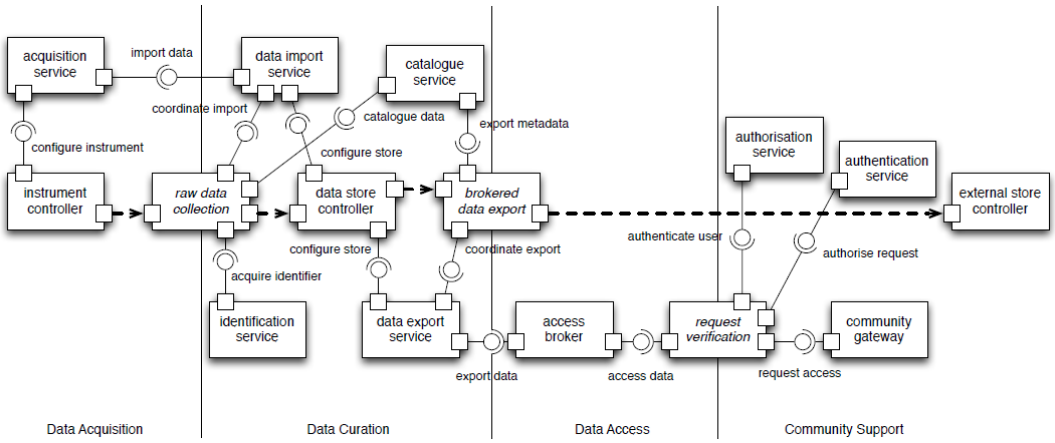

Figure 4: A subset of the core interactions involved in the acquisition and access of data

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.

4. Examples of Usage

The ENVRI-RM is in the process of being adopted by the participating environmental research infrastructures. The European research infrastructure EMSO[13] 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. The EMSO

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

data architecture is currently adapted to the ENVRI Reference Model. According to the ENVRI-RM it includes all 5 common subsystems. Concepts and terms defined in the 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.

The European Plate Observing System (EPOS)[12] 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. The challenge for EPOS is to determine how to integrate existing networks and data centres, and provide standard models for metadata and persistent identification of resources. The ENVRI-RM contributes to the design of the EPOS Core Services by simplifying the design problem, breaking it down by subsystem and demonstrating certain necessary dependencies between science, data and computation via exploration of how the EPOS architecture maps into the ENVRI-RM's complementary viewpoints and distinct subsystems.

To demonstrate the feasibility of the design specifications of the reference model, selected model components have been developed to implement a data access subsystem with 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, for the first-time be retrieved through a single data access interface. Scientists are using this new available resource to study environmental problems that were not previously possible to study. These include: the study of the climate impacts caused by the eruptions of Eyjafjallajökull volcano in 2010, and alien species invasion phenomenon around the Sicily Island.

5. Conclusion and Future Work

The paper describes 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.

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.

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.

6. Bibliography

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

Chen, Y. (et al) (2013): Analysis of Common Requirements for Environmental Science Research Infrastructures, ISGC 2013.

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

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.