PROCEEDINGS OF SCIENCE

Analysis of Common Requirements for Environmental Science Research Infrastructures

Yin Chen
School of Computer Science and Informatics, Cardiff University
Queen's Buildings, 5 The Parade, Roath, Cardiff, UK
E-mail: `Y.Chen@cs.cardiff.ac.uk`

Alex Hardisty, **Alun Preece**
School of Computer Science and Informatics, Cardiff University,UK
E-mail: `{Alex.Hardisty,A.D.Preece}@cs.cardiff.ac.uk`

Paul Martin, **Malcolm Atkinson**
School of Informatics,The University of Edinburgh,UK
E-mail: `{pmartin, mpa}@staffmail.ed.ac.uk`

Zhiming Zhao
The University of Amsterdam, Netherlands
E-mail: `z.zhao@uva.nl`

Barbara Magagna, **Herbert Schentz**
Environment Agency Austria,Austria
E-mail: `{Barbara.Magagna, Herbert.Schentz}@umweltbundesamt.at`

Yannick Legré
Grid and Cloud Institute ,CNRS, France
E-mail: `yannick.legre@idgrilles.fr`

This paper presents the initial findings from a requirements analysis of environmental science research infrastructures. A standard model, Open Distributed Processing (ODP), is used in analysis. From the perspective of the ODP *enterprise viewpoint*, 5 common communities have been identified, *data acquisition*, *data curation*, *data publication*, *data service provision*, and *data usage*. From the perspective of the ODP *engineering viewpoint*, the physical structuring mechanism for the infrastructures have been examined and 5 common sub-systems have been identified: *data acquisition*, *data curation*, *data access*, *data processing*, and *community support*. Finally, from the perspective of the ODP *computational viewpoint*, a set of required functions have been identified which are commonly provided by research infrastructures.

The International Symposium on Grids and Clouds (ISGC) 2013
March 17-22, 2013
Academia Sinica, Taipei, Taiwan

PoS(ISGC 2013)032

1. Introduction

Environmental issues will dominate the 21st century [1]. 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 high costs of such infrastructures require cooperation on sharing experiences and technologies, solving crucial common e-Science challenges together, and avoiding unnecessary duplication. The ENVRI project gathers 6 EU ESFRI[1] environmental research infrastructures (ICOS[2], EURO-Argo[3], EISCAT-3D[4], LifeWatch[5], EPOS[6], and EMSO[7]) in order to develop common data and software components and services. The results will accelerate the construction of these infrastructures. The experience will also benefit the building of other advanced research infrastructures.

The requirement analysis study of ENVRI examines the design of the 6 representative environmental infrastructures, to identify their common computational characteristics and develop an understanding of their requirements.

There are many challenges. Different research infrastructures are designed for different purposes and evolve over time. The designers describe their approaches from different points of view, in different levels of detail and using different typologies. The documentation provided is often incomplete and inconsistent. What is needed is a uniform platform for interpretation and discussion, which helps to unify understanding.

In ENVRI, we choose to use a standard model, Open Distributed Processing (ODP) [3], to interpret the design of the research infrastructures, and place their requirements into the ODP framework for further analysis.

This paper presents the initial findings from this study. Firstly, from the ODP *Enterprise Viewpoint*, we have identified 5 common *communities*, **data acquisition, data curation, data publication, data service provision,** and **data usage**; Secondly, from the ODP *Engineering Viewpoint*, we have examined the architectural characteristics of the research infrastructures, and identified 5 common *sub-systems*: **data acquisition**, **data curation**, **data access**, **data processing** and **community support**. Finally, from the ODP *Computational Viewpoint*, we have identified the functions and embedded computations commonly provided. Matrices are used for comparison. Definitions of functionalities are provided.

[1] ESFRI, the European Strategy Forum on Research Infrastructures, is a strategic instrument to develop the scientific integration of Europe and to strengthen its international outreach.
[2] ICOS, www.icos-infrastructure.eu/, is a European distributed infrastructure dedicated to the monitoring of greenhouse gases (GHG) through its atmospheric, ecosystem and ocean networks.
[3] EURO-Argo, www.euro-argo.eu/, is the European contribution to Argo, which is a global ocean observing system.
[4] EISCAT-3D, www.eiscat3d.se/, is a European new-generation incoherent-scatter research radar for upper atmospheric science.
[5] LifeWatch, www.lifewatch.com/, is an e-science Infrastructure for biodiversity and ecosystem research.
[6] EPOS, www.epos-eu.org/, is a European Research Infrastructure on earthquakes, volcanoes, surface dynamics and tectonics.
[7] EMSO, 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 contribution of this work is threefold:

- It investigates a collection of representative research infrastructures for environmental sciences, and provides a projection of Europe-wide requirements they have; identifying in particular, requirements they have in common;
- It experiments with ODP as an approach in requirement analysis, to serve as a common language for interpretation and discussion to ensure unifying understanding;
- The results from this study can be used as an input to a design or implementation model. Common services can be provided in the light of the common analysis, 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 presents the results of analysis of the common communities; Section 4 discusses the analysis of the common architectural characteristics of studied research infrastructures; Section 5 presents the results of analysis of common functionalities; and chapter 6 concludes the study.

2. ODP Approach

Open Distributed Processing is an international standard published by ISO/IEC [3]. It provides an overall conceptual framework for building distributed computing systems [2, 3]. It defines five specific viewpoints which are abstractions that yield specifications of the whole system related to particular sets of concerns. The five viewpoints are [2, 3]:

- The *enterprise viewpoint*, which concerns the organisational situation in which business (research activity in the current case) is to take place;
- The *information viewpoint*, which concerns modelling of the shared information manipulated within the infrastructure of interest;
- The *computational viewpoint*, which concerns the design of the analytical, modelling and simulation processes and applications provided by the infrastructure;
- 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 ODP approach provides a uniform framework for discussion and comparison of various systems. In addition, since the ODP framework is created to help designers deliver a practical architecture which leads to concrete implementations, using ODP concepts to analyse existing infrastructures can help to drill down into details to discover essential problems. There is one key issue of how to denote in the ODP context the "known unknowns". We tolerate schemas and descriptions with many of these at first, and progressively push the unknowns towards detail or boundaries later.

In the next section, we discuss the main results from the analysis study using the ODP approach.

3. Common Communities

The purpose of the ODP Enterprise Viewpoint, in essence, is to capture the system requirements from the perspective of the people who perform their tasks and achieve their goals as mediated by the infrastructure. We have identified 5 communities that commonly exist in all studied infrastructures: *data acquisition*, *data curation*, *data publication*, *data service provision*, and *data usage*. The definitions of 5 common communities are given below, which are based on their objectives:

- **Data Acquisition Community**, who collect raw data and bring (streams of) measures into a system;
- **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.

The fundamental reason of the division of the 5 communities is based on the observation that all human activities which interact with an environmental science research infrastructure, from data collection to the deliverance of scientific results, can be grouped around 5 major physical resources: **the sensor network**, **the storage**, **the (internet) communication network**, **application servers**, and **client devices**. The human activities interact with applications, software tools and system operations distributed in these 5 physical resources, in order to collaboratively conduct scientific research.

4. Common Architectural Characteristics

In ODP, the purpose of the *Engineering Viewpoint* is to identify and specify the structuring mechanisms for distributed interactions and the functional elements. It concerns the architectural features of an infrastructure.

Five common *sub-systems* have been identified in accordance to the 5 common communities, and they are: **data acquisition**, **data curation**, **data access**, **data processing**, and **community support**. Here, we define each *sub-system* as a set of capabilities that collectively are defined by a set of *interfaces* with corresponding operations that can be invoked by other *sub-systems*. 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 [2, 3]. The 5 sub-systems are described as follows:

The **data acquisition sub-system** collects raw data from sensor arrays, various instruments, or human observers, and brings the measures (data streams) into the system. Note, ENVRI is concerned with the computational aspects of an infrastructure, thus, by definition, the *data acquisition sub-system* starts from the point of sensor signals being converted into digital values and received by the system. There are many related activities including defining data

PoS(ISGC 2013)032

acquisition protocols, design and deployment of the sensor instruments, and configuration and calibration of devices, which are crucial tasks for data acquisition nevertheless beyond the scope of the ENVRI investigation. The *data acquisition sub-system* is typically operated at observatories or stations. Data in the *acquisition sub-system* are normally non-reproducible, the so-called raw data or primary data. Consistent time-stamps are assigned to each data object. There are the cases that the raw data may be generated by a simulation model, in which situation the raw data may be reproducible, in terms of being regenerated. The (real-time) data streams sometimes are temporarily stored (e.g., in computer clusters), then, sampled, filtered or processed (e.g., based on applied quality control criteria). Control software is often provided to allow the execution and monitoring of data flows. The data collected at the *data acquisition sub-system* are transmitted to the *data curation sub-system*, to be maintained and archived there.

The **data curation sub-system** facilitates quality control and preservation of scientific data. It is typically operated at a data centre. Data handled at the *curation sub-system* are often reproducible in term of being able to be re-processed. Operations such as data quality verification, data identification, annotation, cataloguing and long-term preservation are often provided. Various data products are generated and provided for users which to be accessed through data *access sub-system*. There is usually an emphasis on non-functional requirements for a *data curation sub-system* including the need for satisfying performance criteria in availability, reliability, utility, throughput, responsiveness, security and scalability.

The **data access sub-system** enables discovery and retrieval of data housed in data resources managed by a *data curation sub-system*. *Data access sub-systems* often provide facilities such as data portals, as well as services to present or deliver the data products. Search facilities including both query-based and navigation-based searching tools are provided which allow users or services to discover interesting data products. Discoveries based on metadata or semantic linkages are most common. Data handled at the *access sub-system* can be either structurally and semantically homogeneous or heterogeneous. When supporting heterogeneous data, different types of data (often pulled from a variety of distributed data resources) may be converted into uniform representations with uniform semantics which can be resolved by a data discovery and access service. Services allowing harvesting of metadata and/or data, as well as services enhancing the performance by compression and packaging methods and encoding services for secure data transfer are often part of the *data access sub-system*. Data access can be open or controlled (e.g., enforced by authentication and authorisation policies). It is notable that a *data access sub-system* usually does not provide "write" operations for end users, although such operations may be provided for an administrator of a data resource.

The **data processing sub-system** aggregates the data from various resources and provides computational capabilities and capacities for conducting data analysis and scientific experiments. Data handled by the *data processing sub-system* are typically derived and recombined via the *data access sub-system*. A *data processing sub-system* normally offers operations for statistical and/or mining functions for analysis, facilities for conducting scientific experiments, modelling/simulation, and scientific visualisation. Performance requirements for processing scientific data tend to be concerned more about scalability issue, which may also be necessary to address at the infrastructure level -- for example, to make use of the Grid or Cloud technology. In this case, functionalities to interact with the physical infrastructure should be provided.

Finally, the **community support sub-system** manages, controls and tracks users' activities and supports users to conduct their roles in communities. Data handled by a *community support sub-system* typically are user generated data, control and communications. A *community support sub-system* normally supports for interactive visualisations, Authentication, Authorisation and Accounting (AAA), as well as for managing virtual organisations. The *community support* is orthogonal to and cuts across the other 4 *sub-systems*.

Different research infrastructures emphasise the design and implementation of different *sub-systems*. As shown in Table 1, research infrastructures such as ICOS, EISCAT-3D, Euro-Argo, and EMSO mainly focus on data *acquisition*, *curation* and *access*. They are typical **large-scale observatory systems**. Some other research infrastructures such as EPOS and LifeWatch are built on existing systems having limited control over data resources, and focus more on data *access* and *processing*. They are **comprehensive integration infrastructures** for domain data and computations. It worthy of mentioning that generic computational research infrastructures, such as EUDAT[8] and EGI[9], are **general purpose large-scale infrastructures** for data management and processing; EUDAT tends to focus more on the functionalities related to the data *curation sub-system*, and EGI tends to focus more on the functionalities related to the data *processing sub-system*. Both EUDAT and EGI provide generic operations and services which can be used in various domains of research within either infrastructure.

Table 1: The Correlations Between the Design and Implementation Emphasis of Environmental Research Infrastructures and the Five Common Sub-systems

Sub-system	**EISCAT-3D**	**Euro-Argo**	**ICOS**	**EMSO**	**EPOS**	**LifeWatch**	**EUDAT**	**EGI**
Acquisition	Yes	Yes	Yes	Yes				
Curation	Yes	Yes	Yes	Yes			Yes	
Access	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Processing					Yes	Yes		Yes
Community Support						Partial	Partial	Partial

The *community support sub-system* is commonly requested. For example, in EISCAT-3D, there is a need to allow users to control remote radar systems to collect their own data; Both EURO-Argo and ICOS request the expertise of users to verify the quality of data; EMSO (using the PANGAEA system[10]) allows users to submit metadata; whilst EPOS and LifeWatch plan to

[8] EUDAT, eudat.eu, is a collaborative data infrastructure allowing researchers to share data and carry out research effectively.
[9] EGI, www.egi.co.uk, established to coordinate and manage the European Grid Infrastructure (EGI) federation.
[10] PANGAEA, www.pangaea.de, is an information system for archiving, publishing, and distributing georeferenced data from earth system research.

allow users to share their data, experiments and workflows. However, currently we observe few research infrastructures actively designing or implementing this sub-system. This is likely because of resource limitations. EGI offers more experience of supporting its user community. Operations, such as virtual organizations management service (VOMS) and accounting are provided through EGI infrastructure services. However, there are no generic tools or software available to support many newly emerged requirements, such as community coordination, collaboration, policy making, and user collaborative work and publication of results. We anticipate that research infrastructures with greater maturity and sufficient resources will take the lead exploring this area in the near future.

For each *sub-system*, different research infrastructures provides different functionalities. In the following, we present the results of our analysis of the common computational functions identified in each *sub-system*.

5. Common Functionalities

The ODP *Computational Viewpoint* focuses on the functionality of an infrastructure, and the services it offers.

Dividing the structure of research infrastructures into *sub-systems* helps break down the complexity in analysis. Within each *sub-system*, we use a data-oriented approach, which follows the life-cycle of data (e.g., creation, transmission, transformation, modification, processing, and visualisation) to identify key functions and embedded computations.

The analysis results in a number of common required functionalities (which is provided in [4]). 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). We further identified a minimal core set of functions which consist of the fundamental functionality necessary to describe functional environmental research infrastructures.

Table 2 lists the core set of functions[11]. Each function is defined as an **interface** which encapsulates a set of required operations or services that act upon an *object*. By ODP definition, an **object** is a model of a real-world entity, characterised by its behaviour and its state [2, 3]. The interactions between objects are supported by the corresponding computational object *interfaces*.

The value domain used in the table is defined as follows:

V ={*Yes*, *No*, *Unknown*, *Not Applicable*}, where

- *Yes*, denoted by '**Y**' -- evidence has been found that the research infrastructure provides the specified function;
- *No*, denoted by '**N**' -- evidence has been found that the research infrastructure does not provide the specified function;

[11] Due to space limitation, we only present the minimal core set of functions identified. The full list of 58 required functionalities are provided in [4].

- *Unknown*, denoted by '**U**'-- evidence hasn't been found whether the research infrastructure provides the specified function or not;
- *Not Applicable*, denoted by '**NA**'-- the specified function is out of scope of the research infrastructure's planned work.

The analysis results below are a snapshot of the current states of research infrastructures. This is because the design of a research infrastructure evolves over time and is subject to change. For example, the design study of EISCAT-3D finished in 2009. Its succeeding project, the EISCAT-3D Preparatory Phase (EISCAT-3D PP) examines the feasibility of the design and prepares for implementation starting in 2014. During the EISCAT-3D PP, many parts of the design are likely to be re-evaluated and re-designed, e.g., due to infeasibility for implementation. This is a common issue for most if not all other research infrastructures. The investigation in ENVRI is based on existing knowledge of infrastructure designs, keeping up to date with new developments over the life-time of the project.

The list of functionalities is not intended to be complete. The consistency and completeness of this list will be examined in future work.

Table 2: A Core Set of Common Functions

A	**Data Acquisition Subsystem**							
No	**Functions**	**Definitions**	ICOS	EPOS	EMSO	EISCAT-3D	LifeWatch	EURO-Argo
A.1	Process Control	An interface that provide operations to receive input status, apply a set of logic statements or control algorithms, and generate a set of analogy and digital outputs to change the logic states of devices.	Y	U	U	Y	NA	N
A.2	Data Collection	An interface that provides operations to obtain digital values from a sensor instrument, associating consistent timestamps and necessary metadata.	Y	Y	Y	Y	NA	Y
A.3	Data Transmission	An interface that provides operations to transfer data over communication channel using specified network protocols.	Y	Y	Y	Y	NA	Y
B	**Data Curation Subsystem**							
No	**Functions**	**Definitions**	ICOS	EPOS	EMSO	EISCAT-3D	LifeWatch	EURO-Argo
B.1	Data Quality Checking	An interface that provides operations to detect and correct (or remove) corrupt, inconsistent or inaccurate records from data sets.	Y	Y	U	Y	NA	Y
B.2	Data Identification	An interface that provides operations to assign (global) unique identifiers to data contents.	Y	Y	Y	U	NA	U

B.3	Data Cataloguing	An interface that provides operations to associate a data object with one or more metadata objects which contain data descriptions.	U	Y	Y	U	NA	U
B.4	Data Product Generation	An interface that provides operations to process data against requirement specifications and standardised formats and descriptions.	Y	Y	Y	Y	NA	Y
B.5	Data Storage & Preservation	An interface that provides operations to deposit (over long-term) the data and metadata or other supplementary data and methods according to specified policies, and make them accessible on request.	Y	Y	Y	Y	NA	Y
C	**Data Access Subsystem**							
No	**Functions**	**Definitions**	ICOS	EPOS	EMSO	EISCAT-3D	LifeWatch	EURO-Argo
C.1	Access Control	An interface that provides operations to approve or disapprove of access requests based on specified access policies.	U	Y	U	Y	U	U
C.2	Metadata Harvesting	An interface that provides operations to (regularly) collect metadata (in agreed formats) from different sources.	U	U	Y	N	U	N
C.3	Resource Registration	An interface that provides operations to create an entry in a resource registry and insert resource object or a reference to a resource object in specified representations and semantics.	–	–	–	–	–	–
C.4	Data Publication	An interface that provides operations to provide clean, well-annotated, anonymity-preserving datasets in a suitable format, and by following specified data-publication and sharing policies to make the datasets publically accessible or to those who agree to certain conditions of use, and to individuals who meet certain professional criteria.	Y	U	Y	U	Y	Y
C.5	Data Citation	An interface that provides operations to assign an accurate, consistent and standardised reference to a data object, which can be cited in scientific publications.	N	U	Y	N	U	N
C.6	Data Discovery and Access	An interface that provides operations to retrieve requested data from a data resource by using suitable search technology.	Y	Y	Y	Y	Y	U
D	**Data Processing Subsystem**							
No	**Functions**	**Definitions**	ICOS	EPOS	EMSO	EISCAT-3D	LifeWatch	EURO-Argo
D.1	Data Assimilation	An interface that provides operations to combine observational data with output from a numerical model to produce an optimal estimate of the evolving state of the system.	Y	U	U	U	U	NA
D.2	Data Analysis	An interface that provides operations to inspect, clean, transform data, and to provide data models with the goal of highlighting useful information, suggesting conclusions, and supporting decision making.	Y	Y	Y	Y	Y	NA

PoS(ISGC 2013)032

D.3	Data Mining	An interface that provides operations to support the discovery of patterns in large data sets.	Y	U	N	N	Y	NA
D.4	Data Extraction	A interface that provides operations to retrieve data out of (unstructured) data sources, including web pages ,emails, documents, PDFs, scanned text, mainframe reports, and spool files.	Y	U	U	Y	Y	NA
D.5	Scientific Modelling and Simulation	An interface that provides operations to support of the generation of abstract, conceptual, graphical or mathematical models, and to run an instance of the model.	Y	Y	U	U	Y	NA
D.6	(Scientific) Workflow Enactment	A interface that provide operations to support composition and execution a series of computational or data manipulation steps, or a workflow, in a scientific application. Important processes should be recorded for provenance purposes.	N	U	N	N	Y	NA
D.7	Data Processing	An interface that provides operations to initiate the calculation and manage the outputs to be returned to the client.	N	U	N	N	Y	NA
E	**Community Support Subsystem**							
No	**Functions**	**Definitions**	ICOS	EPOS	EMSO	EISCAT-3D	LifeWatch	EURO-Argo
E.1	Authentication	An interface that provides operations to verify a credential of a user.	Y	Y	Y	Y	Y	Y
E.2	Authorisation	An interface that provides operations to specify access rights to resources.	Y	Y	Y	Y	Y	Y

6. Conclusions

The goal of the ENVRI investigation was to identify the common requirements of the environmental science research infrastructures. Throughout the study, ODP has been used as the analysis framework, which serves as a uniform platform for interpretation and discussion to ensure a unified understanding. From the ODP *Enterprise Viewpoint*, we have identified 5 common *communities*; From the ODP *Engineering Viewpoint*, we have examined the architectural characteristics of the research infrastructures and identified 5 common *sub-systems*: **data acquisition**, **data curation**, **data access**, **data processing** and **community support**. Finally, from the ODP *Computational Viewpoint*, we have identified the common functions and embedded computations they provided.

The results from this study can be delivered as an input to a design or an implementation model. Common services can be provided based on the common analysis, which can be widely applicable to various environmental research infrastructures.

There are several elements which could be extended in future work:

- Only 3 ODP viewpoints have been explored. In future work, analysis with respect to the ODP *Information Viewpoint* can be conducted wherein the requirements for a common information model can be investigated; and from the ODP *Technology Viewpoint*, shared technologies and standards can be identified.

PoS(ISGC 2013)032

- The analysis with respect to the ODP *Enterprise viewpoint* has only examined the *communities* involved. It will be useful to further derive the community *behaviours* (in term of activity processes), and to address community *policy* issues.
- More substantially, how the findings from this study can be better used to support constructions of research infrastructures is still an open question. The next phase of the work has been to develop the reference model based on these viewpoints, sub-systems and functionalities that can be used as basis for further dialogue, common engineering design and implementation.

References

[1] European Commission, Directorate-General For Research and Innovation, *Global Change: Towards global research infrastructures", Publications Office of the European Union*, ISBN: 978-92-79-23363-0, 2012.

[2] P. F. Linington, Z. Milosevic, A. Tanaka, and A. Vallecillo, *Building Enterprise Systems with ODP: An Introduction to Open Distributed Processing*, Chapman & Hall/CRC Press, 2011.

[3] *ISO/IEC 10746-1: Information technology—Open Distributed Processing – Reference Model: Overview*, ISO/IEC standard, 1998.

[4] Y. Chen, *Analysis of Common Requirements for Data Processing*, ENVRI Deliverable D3.3, 2013. http://envri.eu/rm