ABSTRACT

E-Science experiments are often interdisciplinary; they require data from different domains and research infrastructures. The interoperability between research infrastructures, including not only cross invocations of services, but also integration between data schemas, processing models and management policies and security controls, is essential to enable large scale data-driven experiments. However, the construction and execution of such e-Science experiments are often driven by specific use cases and result in customized solutions to specific domains. The re-usability of the components used in these solutions is limited. Analysing functional gaps between research infrastructures and decomposing the interoperability issues into separate viewpoints promote a generic information-linking model between data and services in infrastructures. This deliverable presents a multi viewpoint framework, named as “Open e-Science Information Linking Model (OEILM)”, in the context of environmental research infrastructures to semantically link data and infrastructures.
1. COPYRIGHT NOTICE

Copyright © Members of the ENVRI Collaboration, 2013. See www.ENVRI.eu for details of the ENVRI project and the collaboration. ENVRI (“Common Operations of Environmental Research Infrastructures”) is a project co-funded by the European Commission as a Coordination and Support Action within the 7th Framework Programme. ENVRI began in October 2011 and will run for 3 years. This work is licensed under the Creative Commons Attribution-Noncommercial 3.0 License. To view a copy of this license, visit http://creativecommons.org/licenses/by/3.0/ or send a letter to Creative Commons, 171 Second Street, Suite 300, San Francisco, California, 94105, and USA. The work must be attributed by attaching the following reference to the copied elements: “Copyright © Members of the ENVRI Collaboration, 2013. See www.ENVRI.eu for details of the ENVRI project and the collaboration”. Using this document in a way and/or for purposes not foreseen in the license requires the prior written permission of the copyright holders. The information contained in this document represents the views of the copyright holders as of the date such views are published.

2. DELIVERY SLIP

<table>
<thead>
<tr>
<th>Name</th>
<th>Partner/Activity</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>From</td>
<td>See authors below</td>
<td>30-10-2013</td>
</tr>
<tr>
<td>Reviewed by</td>
<td>Yannick Legré</td>
<td>CNRS, WP5</td>
</tr>
<tr>
<td></td>
<td>Leonardo Candela</td>
<td>CNR, WP4</td>
</tr>
<tr>
<td>Approved by</td>
<td>Silvia Wissel</td>
<td>UvA, WP1</td>
</tr>
</tbody>
</table>

3. DOCUMENT LOG

<table>
<thead>
<tr>
<th>Issue</th>
<th>Date</th>
<th>Comment</th>
<th>Author/Partner</th>
</tr>
</thead>
</table>
| 1.0   | 30/09/13 | First draft                       | Zhiming Zhao, Paola Grosso, Cees de Laat(UvA)  
|       |          |                                   | Barbara Magagna, Herbert Schentz (EAA)  
|       |          |                                   | Yin Chen, Alex Hardisty (CU)  
|       |          |                                   | Paul Martin (UEDIN)  
|       |          |                                   | Robert Huber (UNIHB)  |
| 2.0   | 24/10/13 | With comments from other project members | Zhiming Zhao, Paola Grosso, Cees de Laat(UvA)  
|       |          |                                   | Barbara Magagna, Herbert Schentz (EAA)  
|       |          |                                   | Yin Chen, Alex Hardisty (CU)  
|       |          |                                   | Paul Martin (UEDIN)  
|       |          |                                   | Robert Huber (UNIHB)  |
4. APPLICATION AREA

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

5. DOCUMENT AMENDMENT PROCEDURE

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

6. TERMINOLOGY

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

7. PROJECT SUMMARY

Frontier environmental research increasingly depends on a wide range of data and advanced capabilities to process and analyse them. The ENVRI project, “Common Operations of Environmental Research infrastructures” is 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 is 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 that 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

This document describes the semantic linking framework named “Open e-Science Information Linking Model” (OEILM) developed in the ENVRI project task 3.4. The OEILM has three layers 1) the core ontology which is based on Open Distributed Processing (ODP), 2) the ENVRI Reference Model ontology which imports the core ontology and models concepts in ENVRI RM, and 3) the linking ontology which connects the reference model with the information models.
outside research infrastructures, such as description languages for underlying network infrastructures, and for domain specific data and services. The linking layer bridges the reference model ontology with five highlighted clusters of ontologies that are related to descriptions of abstract applications, services, data and metadata, middleware and physical infrastructure.

9. HOW TO READ

This document is part of the deliverables developed in the WP3 of the project: 1) D3.3, analysis of common requirements for ENVRI research infrastructures, 2) D3.4, ENVRI reference model, 3) D3.5, guideline of ENVRI reference model, and 4) D3.6 semantic linking for ENVRI research infrastructures. Readers are recommended to read the D3.3, 3.4 and 3.5 for the background knowledge of the Open Distributed Processing model and the ENVRI Reference Model.

The document is organised as follows:

Section 1 introduces the motivation and background for the ENVRI semantic linking framework.

Section 2 explains the main approach of the Open e-Science Information Linking Model. In this section, the reader will find the architecture of the semantic linking framework, and the basic development strategy.

Section 3 provides the core layer of OEILM: the ODP ontology we developed based on standards of Open Distributed Processing (ODP).

Section 4 presents the ontology of the ENVRI reference model. The ontology is derived from the natural language version of the reference model we developed in the other tasks in the project.

Section 5 discusses the linking ontology in OEILM. In this version of the deliverable, the linking ontology only focuses the abstract level.

Section 6 discusses some initial validation of the current version of OEILM.

Section 7 summarises the current development of OEILM and explains the agenda for the next phase.

The appendices are not part of the semantic framework, and provided for the convenience of the reader.

Appendix A is the glossary of the document that consists of all concepts and terms defined throughout the ENVRI RM, and OEILM.

The intended audience of this document is the ENVRI community as well as other organisations or individuals that are interested in understanding the top level technical architecture that underpins the construction of a research infrastructure.

This document uses the APA (American Psychological Association) citation style.
# TABLE OF CONTENTS

1 INTRODUCTION ............................................................................................................. 8  
1.1 Purpose and scope ........................................................................................................ 8  
1.2 Requirements and challenges ..................................................................................... 9  
1.3 Multi viewpoint modelling .......................................................................................... 10  
1.4 Interoperability and semantic linking ......................................................................... 11  
1.5 Outline ........................................................................................................................ 13  

2 OPEN E-SCIENCE INFORMATION LINKING MODEL ........................................ 14  
2.1 Basic structure ............................................................................................................. 14  
2.2 A formal view on semantic linking ............................................................................. 15  
2.3 Linking methods ......................................................................................................... 16  
2.4 Challenges and goal .................................................................................................... 17  
2.5 Our approach ............................................................................................................. 17  

3 THE CORE ODP ONTOLOGY .................................................................................... 19  
3.1 Enterprise Viewpoint concepts .................................................................................. 19  
3.2 Computational Viewpoint concepts ......................................................................... 20  
3.3 Information Viewpoint concepts ............................................................................... 21  
3.4 Engineering Viewpoint concepts ............................................................................. 22  
3.5 Technology Viewpoint concepts .............................................................................. 23  
3.6 Correspondence between viewpoints ....................................................................... 24  
3.7 Current status ............................................................................................................ 25  

4 THE ENVRI RM ONTOLOGY .................................................................................. 26  
4.1 Science Viewpoint .................................................................................................... 27  
4.2 Computational Viewpoint ........................................................................................ 28  
4.3 Information Viewpoint ............................................................................................. 29  
4.4 Correspondences ...................................................................................................... 30  
4.5 Current status ........................................................................................................... 32  

5 THE LINKING ONTOLOGY ....................................................................................... 33  
5.1 Linking to data standards and metadata ................................................................... 33  
5.1.1 State of the art ..................................................................................................... 33  
5.1.2 Linking to the ENVRI RM ................................................................................. 39  
5.2 Linking to service descriptions ................................................................................ 41  
5.2.1 State of the art ..................................................................................................... 41  
5.2.2 Linking to the ENVRI RM ................................................................................. 42  
5.3 Linking to Computing, Storage and Network in physical infrastructure .................. 44  
5.3.1 State of the art ..................................................................................................... 44  
5.3.2 Linking to the ENVRI RM ................................................................................. 47  
5.4 Summary .................................................................................................................. 48  

6 VALIDATION ............................................................................................................. 50  
6.1 Expressiveness ......................................................................................................... 50
TABLE OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1</td>
<td>Interoperability analysis from different viewpoints.</td>
<td>11</td>
</tr>
<tr>
<td>Figure 2</td>
<td>The basic structure.</td>
<td>14</td>
</tr>
<tr>
<td>Figure 3</td>
<td>The basic structure of the ODP ontology.</td>
<td>19</td>
</tr>
<tr>
<td>Figure 4</td>
<td>The basic concepts and relations among concepts in the Enterprise viewpoint.</td>
<td>20</td>
</tr>
<tr>
<td>Figure 5</td>
<td>The basic concepts and relations among concepts in the Computational Viewpoint.</td>
<td>21</td>
</tr>
<tr>
<td>Figure 6</td>
<td>The basic concepts and relations among concepts in the Information Viewpoint.</td>
<td>22</td>
</tr>
<tr>
<td>Figure 7</td>
<td>The basic concepts and relations among concepts in the engineering viewpoint.</td>
<td>23</td>
</tr>
<tr>
<td>Figure 8</td>
<td>The basic concepts in the Technology Viewpoint.</td>
<td>23</td>
</tr>
<tr>
<td>Figure 9</td>
<td>ENVRI RM ontology and ODP ontology.</td>
<td>26</td>
</tr>
<tr>
<td>Figure 10</td>
<td>science viewpoint in the ENVRI RM ontology.</td>
<td>27</td>
</tr>
<tr>
<td>Figure 11</td>
<td>computational viewpoint in the ENVRI RM ontology.</td>
<td>29</td>
</tr>
<tr>
<td>Figure 12</td>
<td>Information viewpoint in the ENVRI RM ontology.</td>
<td>30</td>
</tr>
<tr>
<td>Figure 13</td>
<td>The correspondences between different ENVRI RM viewpoints regarding instrument configuration.</td>
<td>31</td>
</tr>
<tr>
<td>Figure 14</td>
<td>Linking between ENVRI information viewpoint objects and the metadata standards.</td>
<td>41</td>
</tr>
<tr>
<td>Figure 15</td>
<td>The semantic linking between ENVRI IV objects and the service descriptions.</td>
<td>44</td>
</tr>
<tr>
<td>Figure 16</td>
<td>the relation between infrastructure description languages</td>
<td>47</td>
</tr>
<tr>
<td>Figure 17</td>
<td>Linking the infrastructure description languages.</td>
<td>48</td>
</tr>
<tr>
<td>Figure 18</td>
<td>The current definition of linking ontology.</td>
<td>49</td>
</tr>
<tr>
<td>Figure 19</td>
<td>The screen snapshot of the validation tool.</td>
<td>51</td>
</tr>
<tr>
<td>Figure 20</td>
<td>Describing processes using the OELM ontologies.</td>
<td>52</td>
</tr>
<tr>
<td>Figure 21</td>
<td>The list of processes being described using the tool.</td>
<td>53</td>
</tr>
<tr>
<td>Figure 22</td>
<td>The basic architecture of the data harmonization component (Tarasova et al., 2013).</td>
<td>54</td>
</tr>
<tr>
<td>Figure 23</td>
<td>Datasets as provided by ICOS (above) with CO2 concentrations and by EURO-Argo (below) with ocean temperature measurements</td>
<td>55</td>
</tr>
<tr>
<td>Figure 24</td>
<td>Diagram showing the data mapping process within the Information Viewpoint in RM</td>
<td>56</td>
</tr>
<tr>
<td>Figure 25</td>
<td>Describing the data harmonisation process in the task 4.2 using the ENVRI RM ontology.</td>
<td>57</td>
</tr>
</tbody>
</table>
1 INTRODUCTION

1.1 Purpose and scope

Data driven experiments, for instance using data statistics or mining technologies to extract abstract knowledge from raw materials, or semantically integrating various data sources to obtain profound casual relations between different phenomena, emerge as an important research paradigm in domains such as high energy physics, environmental sciences and bio-informatics for investigating system-level problems. A research infrastructure provides data, and advanced instruments and services for a community of users to conduct different levels of research in the field (Sorvari, 2012). ICOS\(^1\), EURO-Argo\(^2\), EISCAT-3D\(^3\), LifeWatch\(^4\), EPOS\(^5\), and EMSO\(^6\) are examples of research infrastructures in environmental sciences. These research infrastructures all interface to large scale deployed sensors or observers, and deal with very large quantities of spatially-aware data series. Collaborating services and data from different research infrastructures enables science at a system-level but also requires integration and interoperability between operations at different service layers of research infrastructures. Distinguishing the commonality between research infrastructures can not only spread the development efforts for shared functionality but also promote further interoperability between them.

A semantic linking framework that can link various information models surrounding the data life-cycle is crucial to achieve interoperability between services in different research infrastructures.

1. First, accessing distributed data sources requires both metadata describing the content properties, and effective data distribution mechanisms over storage services. A semantic linking framework can complement the metadata with information about storage services and with information about the topology of the underlying network for further content distribution.

2. Second, e-Science experiments require not only data sources from different domains, but also computing resources and network connectivity between different infrastructures for data processing. A linking framework links semantic descriptions of data and resources, and enables a high-level planning service to query and aggregate information across different research infrastructures and to select optimal combinations of services for execution.

3. Third, the provenance of data evolution provides a historical view of the data; a semantic linking framework between data and infrastructure services extends the data provenance with runtime context information for underlying infrastructures. However, the current research infrastructures lack common information models and an effective semantic linking framework to describe the relation between infrastructure components and data.

\(^1\) ICOS, [http://www.icos-infrastructure.eu/](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.

\(^2\) EURO-Argo, [http://www.euro-argo.eu/](http://www.euro-argo.eu/), is the European contribution to Argo, which is a global ocean observing system.

\(^3\) EISCAT-3D, [http://www.eiscat3d.se/](http://www.eiscat3d.se/), is a European new-generation incoherent-scatter research radar for upper atmospheric science.

\(^4\) LifeWatch, [http://www.lifewatch.eu/](http://www.lifewatch.eu/), is an e-science Infrastructure for biodiversity and ecosystem research.


This deliverable proposes a semantic linking framework for environmental research infrastructures named the Open e-Science Information Linking Model (OEILM). OEILM contributes to the ENVRI project from the following aspects:

1. It extends the natural description of the ENVRI Reference Model (ENVRI RM) with a set of ontologies, which offers ENVRI RM users additional possibility to understand the reference model and make use of it.

2. The semantic framework provides formal description of the ENVRI RM; it allows developers to use third party reasoning tools to reason on the consistency and other properties of the descriptions based on the ENVRI RM.

3. By gluing different types of information component into the research infrastructure, the linking framework links the external components, such as domain specific data standards and infrastructure information models, with the ENVRI RM, and enables information query paths across different conceptual layers within the research infrastructures. This will lead to extended interoperability between the ESFRI ENV projects.

1.2 Requirements and challenges

The semantic linking framework glues the semantic descriptions of the data with the infrastructures and provides semantic interoperability between data and resources. It needs to address fault tolerance, optimization and scheduling of linked resources, while making a trade-off between fuzzy and full information. More specifically, the linking framework needs to take the following aspects into consideration:

1. The application aspect includes linking issues between scientific models and application scenarios from different domains; it should also capture issues such as application characteristics, workflow patterns, performance requirements, quality-of-service, security and various different policies in user communities;

2. The data aspect includes linking issues between different data and metadata standards at different phases of data evolution (raw data, transfer, calibration, and fusion);

3. The computing aspect includes linking issues related to the interoperability between operations and services that an application needs from different research infrastructures;

4. The middleware aspect includes linking issues between software middleware from different research infrastructures for data catalogue, storage and data processing related computing services;

5. The infrastructure aspect includes linking issues between underlying physical infrastructures such as network elements and topologies, and also the monitoring information of the runtime status of the underlying infrastructure.

These aspects in the linking models relate to each other but focus on linking issues from different viewpoints. For instance the first aspect focuses on the application logic of experiments enabled by the research infrastructures, and these issues are typically from the perspective of domain scientists or users of the research infrastructures. The second and third aspects include design issues of experiments: data schemas and the computing processing; the fourth and fifth aspects mainly refer to the runtime issues and resource management for supporting the experiments. Clearly distinguishing these aspects allows different users to choose relevant parts in describing and searching information.

The development of the semantic linking framework thus faces several challenges:
1. The first challenge is the lack of common terminology for data and research infrastructures. The research infrastructures are developed within different domains; the diverse data types and processing approaches result in different terminologies used in infrastructure services and resources. It has been recognized as a community effort to synchronize those terminologies, such as the working group in RDA for data foundation and common terminology (RDA, 2013). Reviewing existing standardized taxonomies and attending community efforts will be the basic approach to build terminology for the linking framework.

2. Second, persistent data identifiers (PID) and preservation enable the access of data if resources are found via the semantic linking framework; however, the assignment and management of persistent data identifiers (PID) are still in a very early stage. Persistent identifiers are important in large-scale e-Science experiments – for instance, for retrieving data that are hosted by distributed sources for a complex workflow. Moreover, it also plays a key role in the information-centric delivery paradigm of the data objects. Several PID solutions have been proposed, for instance using hierarchical or flat name spaces; putting publisher as prefix, like the URL, is a typical example of hierarchical name space. Flat name spaces involve binding the hash of an object’s content closely to its name. The interoperability between different PID types will be challenging issues in the linking framework.

3. Third, as a fast growing field, data-driven sciences are pushed by different innovative technologies; however, the utilisation of these technologies varies between experiment purposes and specific domains. These technical details complement the generic specification of the architecture of the application but also require a proper level of detail in the linking model to balance the abstraction of generic architecture and the specific functionality of different technologies.

4. Last but not least, the interoperability between legal issues between research infrastructures fundamentally enables the opening and sharing of the data among infrastructures. These issues are not technical but require effort from different organizations and legal entities. Modelling and including these non-technical issues in the linking model will require interoperability between different policies and legacy issues.

We can see that these challenges come not only from the interoperability between service layers of different research infrastructures, but also from the complexity in both data collection and processing in specific domains. Mixing these different issues in system development hampers the re-usability of any resulting implementation. It is important to decompose and classify these issues. In the next section, we present a multi-viewpoint-based approach.

1.3 Multi viewpoint modelling

The Open Distributed Processing (ODP) model captures the design and development issues in complex distributed systems from five viewpoints, namely enterprise, information, computational, engineering and technology (Linnington et al., 2011).

1. The enterprise viewpoint includes the concepts related to the usage of the system, potential use cases, involved roles, behaviours and interactions;
2. The informational viewpoint models the schemas of the data objects in the systems;
3. The computational viewpoint models the operation and the binding interfaces of the system’s logical components;
4. The engineering viewpoint describes the middleware, decomposition and encapsulation of the system’s functional components;
5. The technological viewpoint describes technology related issues in the development of the system.
These viewpoints deliver information on a system independently for specific purpose; they also correspond to each other. For instance the artefacts in the enterprise viewpoint correspond to data objects in the informational viewpoint, and the behaviours in the enterprise viewpoint correspond to operations in the computational viewpoint. The multi-viewpoint modelling approach is suitable for modelling complex research infrastructures, and we follow the ODP philosophy to build the ENVRI Reference Model (ENVRI RM), and develop the linking framework.

1.4 Interoperability and semantic linking

Linking information and knowledge fragments that represent the semantics of services and data sources related to the data life-cycle essentially enables further interoperability between RIs. It is thus important to analyse the interoperability related issues before we discuss the linking framework.

Following the basic concept of Open Distributed Processing, we can group the interoperability related research into five groups, and briefly give a review on the state of the art.

1. From the enterprise (scientific) viewpoint, the interoperability between different scientific domains and disciplines is considered as the basic approach in e-Science; for instance modelling and simulation-based computing methods extend traditional laboratory-based experimental sciences such as chemistry, physics and finance (Lillian, 2011; Debra et al., 2009). Moreover, many simulation models have also exhibited their power in different contexts, for instance the finite element simulation model shows its strong power not only in fluid dynamics but also in studying financial problems (Qiu et al., 2007). These types of interoperability comprise the basis of interdisciplinary activities in e-Science, and will remain a key issue in the future collaboration of different research infrastructures.

2. Information interoperability refers not only to data content but also metadata, information models and all kinds of descriptions of services needed by e-Science applications. The solutions can be roughly classified into two groups. The first one is to map diverse information models using connectors, for
instance the Advanced Resource Connector ARC between EGEE and NorduGrid (Ellerta et al., 2007). Semantic-level interoperability enables data exchange at the meta-level, for instance the Linked Open Data principles proposed by the EU environmental sciences is for sharing scientific data (Hausenblas, 2011), semantic web services based mediators (Arcieri et al., 2004), and semantic information pre-processing frameworks (Zhao et al., 2011). The second one is to synchronize diverse information representation mechanisms to one common model, for instance OGF promotes a solution named Grid Laboratory Uniform Environment (GLUE) (Field et al., 2008). The information interoperability potentially realizes the semantic-level information exchange between different e-Science domains and for aggregating and distributing data. The data specification team in INSPIRE identified a set of components related to the data interoperability, namely rules for application schemas, coordinate referencing and units model, identifier management, multi-lingual text and cultural adaptability, object referencing modelling, multiple representation and consistency (INSPIRE, 2011).

3. From the computational viewpoint, the application logic of an application requires sharing and reusing the services from different infrastructures. Currently, the scientific workflow management systems from scientific computing contexts, and the MapReduce-related Cloud computing models from large Internet companies such as Google are two main forces pushing the evolution of the cooperative-computing model. Although aiming at different types of data and application contexts, these two models are being merged from different perspectives to get benefits from each other; for instance MapReduce is included as an execution model in Kepler (Ching et al., 2009), and DAG is included in MapReduce to extend its flow control (Budiu et al., 2008). Workflow bus was proposed to realize engine level interoperability between different systems (Zhao et al., 2006). Computational interoperability focuses on the runtime issues of the e-Science applications.

4. Engineering viewpoint interoperability refers to cross-invocation of services among different service layers, and mainly deals with the diversity between middleware in invoking services and in managing the data involved with the service invocation. A typical approach is to provide a gateway node to map the invocation interface of both sides (Wang et al., 2007); another way is to converge development efforts to a standardized service interface for job submission, storage and network management; OGF promoted a standard named HPC basic profile (HPCBP) towards this direction (Ruiz-Alvarez et al., 2008). An important problem in middleware interoperability is how to efficiently utilize the resources from different domains after the middleware interoperability has been enabled; cross-infrastructure resource selection, reservation and allocation still remain as important challenges.

5. From the technology viewpoint, distributed resources have been aggregated and virtualized as different stacks of services built from Clusters, Grids and Clouds. There are several standardization initiatives for Cloud models, for instance the Open Cloud Computing Interface (OCCI) (OGF, 2010), and the Cloud Data Management Interface (CDMI) (SNIA, 2010). Moreover, organizational issues such as management policies for sharing data and resources, including authorization, authentication and accounting issues are included; for instance security level interoperability (Moses, 2005) remains a practical difficulty in collaborating with different infrastructures.

This interoperability issue has been recognized as a key issue in different contexts, and several reference models have been proposed. Riedel proposed an Infrastructure Interoperability Reference Model (IIRF) (Riedel et al., 2011), which is now in the OGF community effort. The main idea of the IIRF is to fill the missing links between the existing earlier interoperability related middleware or standards, in particular the interoperability between infrastructure for High performance computing and high throughput computing. Open Grid Forum set up a working group called Grid Interoperability Now! (GIN) (Riedel et al., 2009).
The IIRF focuses on the different standards in the existing Grid middleware; however, its basic concepts are compatible with our model. In addition to these models, the DL.org consortium has also proposed a model called Digital Library Reference Model (DLRM) (Candela et al., 2011) to improve the interoperability of data management in the context of digital library. The DLRM introduces the main notations characterising the whole digital library domain. Although the digital library society may emphasise on different issues from the environmental research infrastructures, the model of DLRM still shows strong relevance to the ENVRI RM. Moreover, the same project developed a technology and methodology cookbook, i.e. artefact presenting a portfolio of best practices and pattern solutions to common issues faced when developing interoperable Digital Library systems (Athanasopoulos et al., 2011).

The diversity of the scientific domains drives interdisciplinary research in e-Science from various angles; solving diversity by providing interoperability support only promotes the evolution of scientific domains and all kinds of support technologies, but will not eventually remove such diversity. From this point of view, interoperability will be a continuous problem within the e-Science context. Different interoperability technologies as we see from the above analysis play a central role in enabling the collaborations in different contexts, but they also raise a problem for e-Science researchers: how should we deal with interoperability issues in an evolutionary vision of infrastructures and e-Science applications?

In the ENVRI project, the interoperability issue is tackled on two levels. First, a reference model is developed to guide the development of future e-Science application and service layers of infrastructures (Chen et al., 2013). Second, a linking model is used to bridge semantic gaps between the reference model and the domain specific semantic models of data, resources and infrastructures.

1.5 Outline

The rest of the deliverable is organized in 6 sections. Section 2 introduces the layered structure of OEILM. The three basic layers in OEILM are discussed in sections 3 to 5. Section 6 discusses validation of the linking framework. The last section summarizes the deliverable and indicates future work.
2 OPEN E-SCIENCE INFORMATION LINKING MODEL

Linking information and knowledge fragments that represent the semantics of services and data sources related to the data life-cycle essentially enables further interoperability between RIs. A generic and globally operational ontology applicable for all possible situations is most probably not achievable; instead, interrelating different pairs of ontologies through semantic bridges promotes an evolutionary solution for many previously unrelated ontologies. A hybrid approach is taken in the ENVRI project. The Open e-Science Information Linking Model (OEILM) employs the ENVRI Reference Model (ENVRI RM) to distinguish and formulate the commonality between the RIs, and uses a linking layer to bridge the semantic gaps between the reference model and other specific ontologies in RIs. We use the Open Distributed Processing model and semantic web technology as the basis to develop the linking issues.

2.1 Basic structure

We propose a three-layer structure for the OEILM, as shown in Fig. 2:

1. The core ontology of ODP provides basic classes and properties to describe a system;
2. The ENVRI-Reference Model ontology imports the core ontology and models the basic functional components in ENVRI research infrastructures;
3. The linking ontology connects the reference model with the information models outside research infrastructures, such as schemas for underlying physical infrastructures, and for domain specific data and service.

![Figure 2: The basic structure.](image)

In Section 1.2, we highlighted five aspects of the linking framework, namely application, data, service, middleware and technologies. Therefore, in principle, the linking ontology will match the ENVRI reference model ontology to five clusters of ontologies that are related to descriptions of applications, services, data or metadata standards, middleware, and physical infrastructures.
The linking layer ontology aims at bridging the gap between the ENVRI RM ontology with the external ontologies from different viewpoints as shown in Fig 2:

1. From the science viewpoint, the linking layer links ENVRI RM to the ontologies used to describe e-Science experiments, such as workflows, and application level quality constraints;
2. From the information viewpoint, the linking layer links the ENVRI RM to the external data and metadata standards which are used in specific ESFRIs;
3. From the computational viewpoint, the linking layer links the ENVRI RM to the ontologies of services and software components deployed in the ESFRIs;
4. From the technology viewpoint, the linking layer links the ENVRI RM to the underlying technologies, including physical infrastructures that ESFRI may employ for computing, storage and network transfer.
5. From the engineering viewpoint, the linking layer links the ontologies or information model of the middleware and architecture that ESFRI may use for the component assembling and system construction. In the current model, the linking from engineering viewpoint is not explicitly included; therefore, this part is not shown in the Fig. 2.

These clusters are related to the five viewpoints of the reference infrastructures. Since the current reference model only focuses on the science, computational and information viewpoints, we simplify the relation between them in the Fig. 2.

Semantic web technologies provide an open-world perspective on the resource descriptions, and have been widely used in different projects to model services, hardware and networks (Ghijsen et al., 2013). It will therefore be used as the basic mechanism to model the ontology in OEILM.

In the rest of the section, we will discuss how the ODP and ENVRI RM ontologies are developed, and then discuss how data and physical infrastructures are linked with OEILM.

### 2.2 A formal view on semantic linking

The key problem in developing the linking layer ontology between ENVRI RM and external ontologies is how to match different ontologies, distinguish gaps and build links between them. This problem is often investigated in the context of ontology matching, mapping or alignment. In this section, we first give a formal definition on the ontology linking, and then highlight involved issues, afterwards review the existing linking approaches.

An ontology is often defined as a tuple $o=(C, R, \leq_c, \leq_r, \sigma, A, I)$ (Suzette & Jugal, 2010), in which

- $C$ is a nonempty set of classes,
- $R$ is a nonempty set of relations;
- $C$ and $R$ are disjoint;
- $\leq_c$ is a class hierarchy, a partial order on $C$;
- $\leq_r$ is a relation hierarchy, a partial order on $R$;
- $\sigma$: $R\rightarrow C \times C$, representing relationships between classes;
A is a set of class axioms, possibly empty.
I is a set of instances of classes, possibly empty.

Ontology alignment is defined based on correspondence, which is based on a 4-tuple \( \text{corr} = (e_i, e_j, \varphi, p) \), where

- \( e_i \) and \( e_j \) are from two different ontologies \( o_i \) and \( o_j \), they are either both concepts or both relations in their ontologies
- \( \varphi \) is a nonempty set of relationships between \( e_i \) and \( e_j \)
- \( p \) is the confidence on the relationship \( \varphi \)

An ontology alignment is a set of correspondences \( \text{align} = \{ \text{corr} \} \). We define the linking ontology as the set of alignments between ENVRI RM and each external ontologies of ESFRI: \( \text{linkOnt} = \{ \text{align} \} \).

### 2.3 Linking methods

The key task in ontology alignment is to compare similarity between entities from different ontologies. Such comparisons often result in distance measurements between the entities, and are performed at different layers (Enrig, 2007):

- **At the data layer**: the comparison focuses on the data values and objects, and typically involve the following comparison techniques:
  - **Equality checks** to check if values of data types or classes are exactly same.
  - **Syntactic similarity** checks to compute the distances between strings by determining how many atomic actions are required to transform one into another, by computing the difference between numeric values.
  - **Object similarity** checks collect several specific techniques, such as based on coefficients, linkages, or mixtures.

- **At the ontology layer**: the comparison is performed between labels or concepts. More specific techniques include:
  - **Label similarity**, often computed using the syntactic similarity check technique.
  - **Concept similarity**, computed using:
    - **Taxonomic similarity** computed using the shortest distance between concepts in the taxonomic tree. The basic idea is that upper layer taxonomic concepts are more generic than the one at lower layers.
    - **Extensional similarity** computed based on similarity between instances.
    - **Structure similarity** computed based on similarity between the domains and ranges of the properties of the class.

- **At the context layer**: the comparison between entities by including the application in which the ontology entities are used. The most often used technique is to compute the usage similarity of the ontology entities.

Using these basic techniques, ontology alignment is typically conducted via an iterative process that has five steps:
1. The **feature preprocessing** step selects small set of excerpts of the overall ontology definition to describe a specific entity;

2. The **search space definition** step defines the search space in the ontology for candidate alignment.

3. The **similarity computation** step computes the similarity between two entities from different ontologies.

4. The **similarity aggregation** step aggregates the different similarity results of one entity pair. Here the different similarity results can be computed from different aspects, such as from label, sub concepts or instances. This step depends on the design of the algorithm.

5. The **interpretation** step derives the final alignments between entities using different interpretation mechanisms, including human experts.

6. Based on the interpretation, the procedure continues iteratively until no new alignments can be found.

During the past decade, there have been several approaches proposed, for instance:

1. The PROMPT-Suite environment (Bruijn et al., 2006), uses labels in the ontology as main features, and does complete search in the space, syntactic measures for similarity computation, and no similarity aggregation. The procedure ends when it reaches the threshold value given by the user. Similar systems include GLUE (Doan, 2001) and NOM (Ehrig, 2007).

2. APFEL (Alignment process feature estimation and learning) is another example. It employs machine learning techniques to perform automatic discovery of the alignment (Ehrig et al., 2005).

In the development of OEILM, we will extend the existing similarity checking approach by including use case contexts of specific ESFRI projects, and existing taxonomies employed in specific domains.

### 2.4 Challenges and goal

The linking framework faces several challenges. First, theENVRI reference model is still in its development phase; the quick evolution of ESFRI makes the development of the ENVRI reference model face the challenge of stability. The current ENVRI RM only focuses on a minimal set of research infrastructure requirements; building a linking ontology on top of such reference model should thus consider the evolution path of the reference model itself. Second, the linking framework has to face a diverse set of external ontology that the research infrastructure needs. How to couple these diverse ontologies into one coherent linking framework requires a flexible structure and open mechanisms for including new changes.

OEILM will not aim at linking the ENVRI RM with an exhaustive list of external ontologies; instead, we will focus on 1) the approach of constructing such linking ontologies, and 2) defining an open structure for OEILM that will allow for linking future external ontologies.

### 2.5 Our approach

OEILM will thus be developed via several steps:

1. Develop an ODP core ontology to describe semantics of components in complex distributed systems;

2. Develop an ENVRI RM ontology to describe the semantics of data and resources in ENVRI research infrastructures;
3. Review information models of data, metadata, resources and infrastructures in current ESFRI projects, and identify the most representative ones from the review list.

4. Define links between the reference model ontology and selected representative models identified in step 3.

5. Validate OEILM by different criteria.

In the following sections, we will follow these steps to explain how OEILM is developed.
3 THE CORE ODP ONTOLOGY

The ODP ontology defines the basic vocabulary for describing the ENVRI RM. In (Alain & Andrey, 2001), Alain et al. discussed the early work on the ODP ontology. Based on the existing work and the ODP standard, we model the basic ODP vocabulary using OWL.

![Diagram of ODP ontology](image)

Fig. 3 shows the basic structure of the ODP ontology. In this deliverable, we use the unfilled arrow to indicate the inheritance relation. We will briefly explain the main concepts defined in the current version.

3.1 Enterprise Viewpoint concepts

The Enterprise Viewpoint models the business or organization in which the system operates. The Enterprise Viewpoint concepts cover the user community, behaviour, scope, purpose and policies of the system. Fig. 4 shows the basic concepts and relations among the concepts defined in the current Enterprise Viewpoint:

- The *EV_Community* concept represents a collaboration between enterprise objects or roles that share common business purposes.
- The *EV_Object* concept represents any abstract objects that appear in enterprise viewpoint specifications.
- The *EV_Role* concept is used to identify different roles in the behaviour of a community; each role is performed by an *EV_Object* in that community. A role that is also a reference indicates a link between objects or actors at a specific point of use.
- The *EV_Policy* concept represents constraints that govern a collection of objects in the community.
- The *EV_Behavior* concept includes actions, interactions, steps, and processes. In practice, behaviour can be modelled as processes and steps within individual roles, or as interactions between roles.

Each concept has an abbreviation prefix of the viewpoint it belongs to; however, we eliminate the prefix where it can easily be inferred in order to make the diagram easier to read.
3.2 Computational Viewpoint concepts

The Computational Viewpoint defines the basic modelling concepts and structure rules for describing a system. Fig. 5 shows the basic concepts defined in the Computational Viewpoint:

- The `CV_Object` concept models different types of object from the computational viewpoint such as application objects, binding objects, computational objects, human objects, environmental objects and presentational objects (typically the user interface).
- The `CV_Behavior` concept models actions and interfaces of a specific object. There are three types of interface: streaming interfaces, operational interfaces and signal interfaces.
- The `CV_Interaction` concept models the interactions between objects; it contains three specific types: stream, binding, and signal. A flow is a sequence of interactions.
- The `CV_Rule` concept models rules that the computational system should follow, including interaction rules, binding rules, failure rules, template rules and type rules.
3.3 Information Viewpoint concepts

The Information Viewpoint concepts focus on the semantics and processing actions of information in the system. Fig. 6 shows the basic concepts defined in `InformationViewpoint_Thing`:

- The `IV_Object` concept models data that is handled by the system described using the information viewpoint.
- The `IV_Action` concept models action types that cause changes in information objects.
- The `IV_Schema` concept includes static schemas, dynamic schemas and invariant schemas. Static schemas describe instantaneous views of the information, dynamic schemas models the state changes of information objects, using for example state machines, and invariant schemas models the relation or constraints among information objects.
Figure 6: The basic concepts and relations among concepts in the Information Viewpoint.

### 3.4 Engineering Viewpoint concepts

The Engineering Viewpoint concepts include definition of mechanisms and functions required to support distributed interaction between objects in a system. Fig. 7 depicts the main Engineering Viewpoint concepts:

- The `NV_Object` concept models objects needed for describing a system from the engineering viewpoint; Basic Engineering Objects specifically correspond to computational objects from engineering point of view;
- The `NV_Structure` concept models the structure or configuration of engineering objects, including concepts such as structure configuration, node, channel, and checkpoint.
- The `NV_Function` concept models concepts used to describe functionality of engineering objects; it includes four groups: management functions, security functions, coordination functions, and repository functions.
- The `NV_Structure_Rule` concept models rules that engineering objects should follow; it includes nodes, distributed bindings, relocations, interface references, application management, failures, clusters and capsules.

These concepts model the dependencies that a system has upon the functionality offered by the platform that the system chooses for the implementation.
Figure 7: The basic concepts and relations among concepts in the engineering viewpoint.

### 3.5 Technology Viewpoint concepts

The Technology Viewpoint concepts provide mechanisms for technology issues, such as standards in software and hardware that will be used to construct the system.

- The **TV_Object** concept is an abstract class to describe the technology configuration related issues that a system requires, for instance networks, software resources, and hardware components.
- The **TV_IXIT** and **TV_Implementation_Standard** concepts are the abstract class types for modelling implementations, and testing rules and standards for system construction. Fig. 8 shows the basic TV concepts.

Figure 8: The basic concepts in the Technology Viewpoint.
3.6 Correspondence between viewpoints

The concepts in the five viewpoints are correspondent with each other. In the ODP ontology, the correspondences are modelled as properties of the corresponding concepts. Table 1-6 show the basic correspondence relations between concepts from five different viewpoints. According to the ODP standard, we can see the most correspondence are among engineering, computational and information viewpoints. In the ontology, correspondences are modelled as the property, which implies the relationship between two concepts that belong to different viewpoints.

Currently, correspondences are modelled into five groups:

- The ev_cv_correspondence property models the correspondences between concepts in enterprise and computational viewpoints, as shown in Table 1.
- The ev_iv_correspondence property models the correspondences between concepts in enterprise and information viewpoints, as shown in Table 2.
- The cv_iv_correspondence property models the correspondences between concepts in computational and information viewpoints, as shown in Table 3.
- The cv_nv_correspondence property models the correspondences between concepts in computational and engineering viewpoints, as shown in Table 4.
- The ev_nv_correspondence property models the correspondences between concepts in enterprise and engineering viewpoints, as shown in Table 5.
- The tv_nv_correspondence property models the correspondences between concepts in technology and engineering viewpoints, as shown in Table 6.

From the ODP standard specifications, the following correspondences are highlighted.

<table>
<thead>
<tr>
<th>CV Object</th>
<th>EV Object</th>
<th>Role</th>
<th>Behavior</th>
<th>Policy</th>
</tr>
</thead>
<tbody>
<tr>
<td>CV Object</td>
<td>Y</td>
<td>Y</td>
<td></td>
<td>Y</td>
</tr>
<tr>
<td>CV Behavior</td>
<td></td>
<td>Y</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>Interface</td>
<td></td>
<td></td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>Stream</td>
<td></td>
<td></td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>Bind</td>
<td></td>
<td></td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>Interaction</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Environmental contract</td>
<td></td>
<td></td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>Operation</td>
<td></td>
<td></td>
<td></td>
<td>Y</td>
</tr>
</tbody>
</table>

Table 1: The correspondences between the Enterprise Viewpoint and the Computational Viewpoint.

<table>
<thead>
<tr>
<th>IV Object</th>
<th>EV Object</th>
<th>Role</th>
<th>Policy</th>
<th>Behavior</th>
<th>Artefact</th>
</tr>
</thead>
<tbody>
<tr>
<td>IV Object</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>IV Action</td>
<td></td>
<td></td>
<td>Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Schema</td>
<td></td>
<td>Y</td>
<td>Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Invariant schema</td>
<td></td>
<td></td>
<td>Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Static schema</td>
<td></td>
<td></td>
<td>Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dynamic schema</td>
<td></td>
<td></td>
<td>Y</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: The correspondences between the Enterprise Viewpoint and the Information Viewpoint.
<table>
<thead>
<tr>
<th>IV_Object</th>
<th>IV_Action</th>
<th>Schema</th>
<th>Static schema</th>
<th>Dynamic schema</th>
<th>Invariant schema</th>
</tr>
</thead>
<tbody>
<tr>
<td>CV_Object</td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CV_Behavior</td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>State</td>
<td>Y</td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interaction</td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Environmental contract</td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3: The correspondences between the Computational Viewpoint and the Informational Viewpoint.

<table>
<thead>
<tr>
<th>NV_Object</th>
<th>Interface</th>
<th>Binding</th>
</tr>
</thead>
<tbody>
<tr>
<td>CV_Object</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>Interface</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>Binding</td>
<td>Y</td>
<td></td>
</tr>
</tbody>
</table>

Table 4: The correspondences between the Computational Viewpoint and the Engineering Viewpoint.

<table>
<thead>
<tr>
<th>EV_Object</th>
<th>Role</th>
<th>Interaction</th>
<th>Policy</th>
</tr>
</thead>
<tbody>
<tr>
<td>NV_Object</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Node</td>
<td>Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Channel</td>
<td>Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stubs</td>
<td>Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Binders</td>
<td>Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protocol objects</td>
<td>Y</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5: The correspondences between the Enterprise Viewpoint and the Engineering Viewpoint.

<table>
<thead>
<tr>
<th>TV_Object</th>
<th>NV_Object</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Y</td>
</tr>
</tbody>
</table>

Table 6: The correspondences between the Technology Viewpoint and the Engineering Viewpoint.

3.7 Current status

The ODP ontology is prototyped using Protégé (Protégé 2013), and the current version is available at http://staff.science.uva.nl/~zhiming/ontology/odp.1.owl. The concepts and properties are derived from the ISO/IEC 19793 and the book of building enterprise systems with ODP (Linington et al., 2011). It covers most of the basic definition in the ODP; however, we do not aim at a full translation of the ODP standard. Both the ODP and ENVRI RM ontology are developed in an iterative way. We start with a basic set of the ODP vocabulary, and use it to describe the ENVRI RM. When the ENVRI RM ontology requires concepts or relations that are not defined in the ODP ontology, certain refinement procedures on the ODP ontology will be triggered.
4 THE ENVRI RM ONTOLOGY

The ENVRI RM ontology is derived from the natural language description of the ENVRI RM\textsuperscript{7}. Fig. 9 shows part of the ENVRI RM ontology\textsuperscript{8}. Basically, the viewpoints in the ENVRI RM inherit the viewpoints from the ODP ontology defined in the previous section. In this section, we will describe how the concepts defined in each viewpoint fit into the ENVRI RM ontology.

The current ENVRI RM only highlights three main viewpoints of the ESFRI research infrastructures analysed:

- The \textit{Science Viewpoint} inherits the basic concepts and properties from the enterprise viewpoint of ODP, and adds the ENVRI specific parts. It presents the abstract vision on how the research infrastructures work in general.
- The \textit{Computational Viewpoint} models the common computational characteristics of the ESFRI using ODP computational viewpoint concepts. We highlight this viewpoint because it captures the internal dependencies between common operations that the ESFRI infrastructures share.
- The \textit{Information Viewpoint} models the various schemas that data has in the context of the data lifecycle supported by research infrastructure. It complements the computational viewpoint with a specific model for the evolution of research data.

There is no implication that ENVRI RM does not need the other two viewpoints; but the current ENVRI RM only focuses on this minimal set.

\begin{itemize}
\item \textsuperscript{7} Available at http://www.envri.eu/rm
\item \textsuperscript{8} Available at http://staff.science.uva.nl/~zhiming/ontology/envri_rm.1.owl
\end{itemize}
4.1 Science Viewpoint

The main concepts in the Science Viewpoint of the ENVRI RM: RM_Science_Viewpoint_Thing, which is a subclass of the ODP_Enterprise_Viewpoint. Currently, the RM_Science_Viewpoint_Thing extended three concepts from ODP enterprise viewpoint: community, community behaviour and role are three concepts. The extensions are defined as sub classes instead of instances.

- **Community** models common communities that ESFRI research infrastructures share: the Data Acquisition Community, Data Curation Community, Data Publication Community, Data Service Provision Community, and Data Usage Community.
- **Community behaviour** models the behaviour for each community:
  - Data Acquisition community: Design of measurement model, instrument configuration, instrument calibration, and data collection;
  - Data curation community: data quality checking, data preservation, data production generation, and data replication;
  - Data publication community: data publication, semantic harmonisation, data discovery and access, and data citation;
  - Data service provision community: service description, service registration, service coordination, service composition, and service orchestration;
  - Data usage community: user behaviour tracking, user profile management, user working space management, user working relationships management, user group work supporting.
- **Role** models the roles in each community:
  - Data Acquisition community:
    - Active roles: environmental scientist, measurement model designer, technician, measurer, observer and data collector;
    - Passive roles: sensor, sensor network and data acquisition sub system;
  - Data curation community:
    - Active roles: data curator and storage administrator;
    - Passive roles: data curation sub system and storage;
  - Data publication community:
    - Active roles: data originator;
4.2 Computational Viewpoint

The concepts and relations in the computational viewpoint are derived from the analysis of the typical processes and scenarios we collected from the ESFRI projects. Basically, computational objects and their interfaces are the two main outputs. In more detail:

- The CV_Object concept models different kinds of services or components in the infrastructure that provide functionality for each other or the external world. The sub-classes of CV_Object currently include: acquisition service, instrument controller, science gateway, virtual laboratory, data store controller, catalogue service, raw data controller, PID service, annotation service, coordination service, data broker, data transfer service, metadata service, process controller, process resource register, security service, and semantic broker.

- The CV_Interface concept models the signature of the service, and the interactions between objects. The sub-classes of the CV_Interface include access data, acquire identifier, annotate data, authorize action, calibrate instrument, coordinate process, query resource, register data resource, request task, request data, resolve citation, stage data, stage task, streaming data service, translate request, update catalogues, update model, update records, and update registry.

- The CV_Property models the relation between CV_Object and CV_Interface. Based on the relation of provide and consume the CV_Interface, we group properties into CV_provideInterface, and CV_consumeInterface.

Figure 11 shows part of the concepts and the relations.
4.3 Information Viewpoint

The information viewpoint concepts in the reference model include data and information involved in computations in a research infrastructure. The basic concepts are defined in two main categories:

- The IV_InformationObject concept models information entities involved in different computing processes in the research infrastructure. Information objects are derived from processes each sub-systems support.
  - Data acquisition related information objects: specification of investigation design, specification of measurements or observation and measurement result;
  - Data curation related information objects: conceptual model, QA_notation;
  - Data access related information objects: metadata, metadata catalogue, citation, concept, persistent data, data state, unique identifier, backup, and metadata state;
  - Data processing related information objects: data provenance, service, service description, and mapping rule;
  - User community related information objects: citation, data provenance, metadata catalogue, and metadata.
- The concept of IV_InformationAction.
  - Data acquisition related information actions: specify investigation design, specify measurement or observation, perform measurement or observation, store data;
  - Data curation related information actions: final review, add metadata, annotate metadata, register metadata, publish metadata, build conceptual model, setup mapping rules;
  - Data access related information actions: query data, perform mapping, setup mapping rules;
  - Data processing related information actions: do data mining.

Fig. 12 shows part of the concepts and their relations defined in the information viewpoint part of the ontology.
As Information Actions are also defined as proper entities in the ODP standard, they have been treated as concepts and not as properties in the ontology. Actions are performed upon Information Objects and to be able to model this, object properties have been established between these two concept groups with labels which reflect the activity. In some cases, actions act on other actions and thus a relation between those two InformationAction concepts is needed (e.g. AnnotateAction has an annotatedAction property applied to any InformationAction in the model).

4.4 Correspondences

The procedure of identifying correspondences among the viewpoints in the ENVRI RM also validates the definition of the ENVRI RM ontology. Corresponding different viewpoints can synchronise the level of details of different viewpoints and detect missing concepts or properties in the ontology. As we mentioned in the beginning, the development of the ENVRI RM starts from a minimal set of the functional components in a research infrastructure; the construction is based on a set of selected common processes identified from the ESFRI research infrastructures. Therefore, the correspondence identification also starts with this set of processes for building the ENVRI RM. For each process all concepts used in the different
viewpoints are presented side by side. For corresponding concepts between the viewpoints dashed lines with arrows at both ends which cross the viewpoints are used (see Fig. 13).

In Section 3.6, we discussed the correspondences among ODP concepts using Tables 1 to 6. Since the viewpoints in the ENVRI ontology inherit from the ODP, we apply the ODP correspondences as the starting point for analysing the correspondence among ENVRI concepts. The Enterprise Viewpoint (EV) is inherited as Science Viewpoint (SV) in the ENVRI ontology. Thus, several correspondences are highlighted in the first round analysis, for instance between SV_Role and CV_Object (table 1), between SV_Behavior and CV_Interface (CV_Behavior) (table 1), between SV_Behavior and IV_Action (table 2), between IV_Object and CV_Object (table 3), and between IV_Action and CV_Interface (CV_Behavior) (table 3).

![Diagram of Instrument configuration](image)

**Figure 13:** The correspondences between different ENVRI RM viewpoints regarding instrument configuration.

Fig. 13 shows an example of the instrument configuration process. Between Science Viewpoint and Information Viewpoint there could be identified the SV_Behavior_IV_Action correspondence by linking the Instrument Configuration (behaviour) with the Specify Measurement/Observation (action). The
IV_Action.CV_Interface correspondence could be assigned to the relation between the Specify Measurement/Observation (action) of the Information Viewpoint and the Configure Instrument Client and Server Interfaces of the Computational Viewpoint.

We will perform such analysis on all the processes included in the ENVRI RM (Chen et al., 2013). We believe such analysis will synchronize the levels of details between different ENRI RM viewpoints, and identify missing concepts in the ontology. The final analysis results will also be part of the ENVRI RM as the correspondence ontology.

4.5 Current status

The ENVRI RM imports the current version of the ODP ontology, and also constructed using Protégé. The current ENVRI RM ontology is based on the latest deliverable of the ENVRI reference model (www.envri.eu/rm). The current version is available at http://staff.science.uva.nl/~zhiming/Ontology/.

The correspondence part of the ontology requires continuous effort, and the current version only contains initial analysis.
5 THE LINKING ONTOLOGY

The linking ontology bridges the ENVRI reference model and other ontologies from different viewpoints. We have mentioned in the beginning that there will be in principle five clusters of ontologies for linking; however, in the first phase of the development we will only focus on three of them: data standards and metadata, service description, and physical infrastructure descriptions.

5.1 Linking to data standards and metadata

The data lifecycle in ENVRI RM begins with the acquisition of raw data from data collecting instruments (seismographs, weather stations, robotic buoys, human observations, etc.) that is then pre-processed and curated within a number of data stores belonging to an infrastructure or one of its delegate infrastructures. This data is then made accessible to authorised requests by parties outside the infrastructure, as well as to services within the infrastructure. The data standards and metadata vary depending on specific domains and research infrastructures. The EMSO RI, for example, distinguishes between observations and archived data; the observation data is modelled and stored according to OGC O(observation) and M(measurement) specification, while the archived data uses ASCII and Network Common Data Form (NetCDF) (Russ & Glenn, 1990) format. The metadata for both observation and archived data are collected; currently, Sensor Model Language (SensorML) (SensorML, 2013) is used to describe observation data, and the metadata for archived data sets is compatible with ISO19115 (ISO19115, 2013) or the NetCDF specification. Another example is the EuroArgo RI, in which data standards for web services and non-programmatic access are different: the NERC DataGrid (NDG) (Lawrence et al., 2005) Vocabulary for the former, and the CSV (CSV, 2013) format for the latter. The ISO 19115 content model is the basis for XML formats used for SeaDataNet metadata services (which provide access to EURO-Argo data). In this section, we will first review the existing metadata and data standards, and then discuss their semantic links to the reference model.

5.1.1 State of the art

Before we start, it is necessary to define the following concepts that used in this deliverable:

1. **Data**: Data is the representations of information dealt by information systems and users thereof (ISO/IEC 10746-2).
2. **Data set**: An identifiable collection of data (ISO19115).
3. **Data model**: Graphical and/or lexical representation of data, specifying their properties, structure and inter-relationships. (ISO11179)
4. **Data type**: A set of distinct values, characterized by properties of those values and by operations on those values (ISO 11179).
5. **Metadata**: Data about data, in scientific applications used to describe, explain, locate, or make it easier to retrieve, use, or manage an information resource.
6. **Metadata Schema**: A cognitive framework that helps organize and interpret information in metadata (derived from the definition in psychology).
7. **Data standard**: A standardized data model.
8. **Metadata standard**: A standardized metadata schema.
In this section, we will first briefly describe the metadata standards that are utilized or considered by the ESFRI projects. From the descriptions of the ESFRI research infrastructures, we select a set of representative metadata standards from the list: SensorML, ISO 19156, NetCDF, ISO 19115, CSR, Dublin Core, CERIF, CSMD, and INSPIRE.

5.1.1.1 NetCDF

Network Common Data Form (NetCDF) is a data model for array-oriented scientific data. It is a machine independent format, which provides interfaces, libraries and format to support the creation, access, and sharing of scientific data. There are two data models available: the classical model and the enhanced model.

The metadata for NetCDF include the following three main parts:
1. **Dimensions** define the size of the different variables;
2. **Variables** specify the type of data the file contains;
3. **Global attributes** provide additional information that applies to all the data in the file.

It can exist in different forms: CDL text, ncML, ncML with coordinate system extensions, and ncML-GML (Netcdf, 2011).

5.1.1.2 Dublin Core (ISO 15836)

The Dublin Core metadata element set is a vocabulary of 15 properties for describing resources. The Dublin connection comes from the original workshop that took place in the Dublin, Ohio, in 1995. The Dublin Core Metadata Initiative (DCMI) maintains a large set of metadata vocabularies and technical specifications.

The core Dublin Core has also been standardised as ISO/FDIS 15836, which was released in 2008. The core elements include: creator, subject, description, publisher, contributor, date, type, format, identifier, source, language, relation, coverage, and rights. Each element can be optional or used multiple times.

5.1.1.3 SensorML

Sensor Model Language (SensorML) is part of the Open Geospatial Consortium (OGC) Web Services (OWS). It aims at capturing fundamental properties and assumptions regarding sensor system, providing description of sensors and sensor systems for inventory management, and supporting the processing and analysis of the sensor observation (SensorML, 2013). Its use has been considered in EMSO.

In SensorML, *processes* are the basic concept, which model all components, including both physical processes, such as transducers, actuators, and processors, and sensors, and non-physical processes including pure operations such as mathematical functions. Processes can be connected as a process chain, which can also be viewed as a process, namely composite process. Each process contains inputs, outputs, and parameters. The input, output and parameters are defined as properties of the concept of process. The metadata of the process and the properties are collected in metadata groups. Currently there are five metadata groups are defined.
1. **General information** metadata group includes three properties: identifier, classifier and description;

2. **Constraint** metadata group includes national or international security constraints, valid time, and legal constraints. These constraints may or may not be considered as information suitable for the discovery process;

3. **Property** group includes characteristics and capabilities properties;

4. **Reference** group includes contact and documentation properties that are useful for human consideration;

5. **History** group is a collection of event objects that serve as the value of the history property.

SensorML is expressible in XML, UML and RDF.

### 5.1.1.4 Earth Observation Metadata Profile (ISO19156)

This International Standard arises from work originally undertaken through the Open Geospatial Consortium’s Sensor Web Enablement (SWE) activity. SWE is concerned with establishing interfaces and protocols that will enable a “Sensor Web” through which applications and services will be able to access sensors of all types, and observations generated by them, over the Web. SWE has defined Sensor Model Language (SensorML), Observation & Measurements (O&M) and other services. The version 2 standard was released in 2012. The most basic model in the standard is about observation and measurement: an observation is an event that estimates an observed property of some feature of interest using a specified procedure and generates a result (ISO19156, 2012).

Mapping metadata in the context of observations and measurements, several properties can be highlighted:

- **Metadata**: General properties for describing observation: such as data identifier, downlink, archiving information;

- **Procedure**: Describes the process or methodology used in the estimation of the result in this observation. This includes the description of the platform/instrument/sensor used for the data acquisition and of the process parameters;

- **Feature of interest**: it is the representation of the real world object the property is being estimated on;

- **Result**: this is the observation/measurement value of the property of the feature-of-interest, its format or type is dependent on the domain and on the composition of the feature-of-interest itself (so it could be GML coverage, features or result types provided by SWE Common), the result should also contain validation information;

- **Observed properties**: mandatory for the observation description;

- **Result Time**: this is the parameter/variable estimated of the feature-of-interest.

ISO 19156 has been specified as a set of XML schema.

### 5.1.1.5 ISO 19115

ISO 19115 was prepared by the TC ISO/TC211 to describe metadata for geographic information. The first edition was released in 2003 as an ISO standard (ISO19115, 2003). The standard is part of the ISO
geographic information suite of standards (19100 series). ISO/TC 211 is a standard technical committee formed within ISO and is concerned with the standardization in the field of digital geographic information.

The ISO 19115 contains a basic set of metadata that are used to describe the information of what, when, who, and where about datasets. The core metadata include three types: mandatory, mandatory under certain conditions, and optional.

- **The mandatory metadata** include: title, reference date, language, topic catalogue, abstract, contact point, and date stamp of the data set;
- **The mandatory under certain conditions metadata** include: geographic location of dataset, character set of dataset, metadata language, and metadata character set;
- **Optional metadata** include: data set responsible party, spatial resolution of the data set, distribution format, additional extent information for the dataset, spatial representation type, reference system, lineage, online resource, metadata file identifier, metadata standard name, and metadata standard version.

The ISO 19139 provides the XML implementation schema for ISO 19115 specifying the metadata record format and may be used to describe, validate, and exchange geospatial metadata prepared in XML. But there exists also a RDF representation of the ISO 19115: https://www.seegrid.csiro.au/subversion/xmml/metadata/ISO19115/iso-19115.owl.

5.1.1.6 SeaDataNet CSR metadata

The Cruise Summary Reports (CSR, formally known as ROSCOPs) are a mandatory format for chief scientists to use to submit reports after a cruise. The SeaDataNet CSR metadata elements are mainly derived from ISO 19115, but also contain elements from ISO 19115-2 and ISO 19139. The details of ISO 19115 were discussed in the previous section.

5.1.1.7 Ecological metadata language (EML)

EML is the basis of the metadata specification used by the LTER Europe community which is involved in LifeWatch. EML is a metadata specification developed by the ecology discipline and for the ecology discipline. It is based on prior work done by the Ecological Society of America and associated efforts (Michener et al., 1997, Ecological Applications). EML is implemented as a series of XML document types that can be used in a modular and extensible manner to document ecological data. Each EML module is designed to describe one logical part of the total metadata that should be included with any ecological dataset.

5.1.1.8 CERIF

The Common European Research Information Format (CERIF) is a formal conceptual model to manage the set up and interoperation between different research information systems. The concepts in CERIF in the contain several group of entities:

1. Basic entities: project, person, and organization unit;
2. Result entities: result publication, result patent, result product;
3. 2nd entities: on top of the basic entities: citation, country, Currency, Curriculum Vitae, Electronic address, equipment, event, expertise and skills, facility, funding, language, metrics, postal address, prizeaward qualification, service, medium, measurement, indicator;
4. Linking entities: the connection between two entities;
5. Additional entities and classification entities.

5.1.1.9 CSMD

The Core Scientific Meta-Data Model (CSMD) was originally developed for study-data oriented models to support data collected within a facility’s scientific workflow. The CSMD is organised around the concept of study, which is a body of scientific work on a particular subject of investigation. A study typically contains one or more investigations e.g. experiments, observations, measurements and simulations. The results from these investigations usually are collected from different stages: raw data, analysed or derived data and end results suitable for publication.

The core entities in the CSMD include:

1. **Investigation**: forms the fundamental unit of the model, with a title, abstract, dates, and unique identifiers referencing the particular study. Also associated with an investigation are the facility and instruments used to collect data;
2. **Investigator**: describes the people involved in the study, together with their institution and role in the study (e.g. principle investigator, research student);
3. **Topic and Keyword**: provide controlled and uncontrolled vocabulary to annotate and index the investigation;
4. **Publication**: provides links to publications associated with (motivating or derived from) the investigation;
5. **Sample**: information on the material sample under investigation within the study. The model has fields for a sample’s name, chemical formula and any associated special information about it, such as specific safety information on a toxic material;
6. **Dataset**: one or more datasets can be associated with an investigation, representing different runs or analyses on the sample. Initially a raw data set can be attached to the investigation, but subsequently, analysed datasets can also be associated;
7. **Data file**: the CSMD takes a hierarchical view of data holdings, as data sets may contain other datasets as well as units of storage, typically data-files. Each data-file has more detailed information, including its name, version, location, data format, creation and modification time, and fixity information such as a Checksum;
8. **Parameter**: parameters describe physical entities associated with the investigation, such as temperature, pressure, or scattering angle, describing either the parameters of the sample, the environment the data was collected in, or the parameters being measured. Thus parameters are associated with the sample, dataset or the data file, and have names, units, values, and allowable data ranges;
9. **Authorisation**: the CSMD can associate conditions on investigations and data sets, so that user specified access conditions can be specified. Thus the authorisation entity can record which user in which role can access data on specific investigations.

### 5.1.1.10 INSPIRE

The INSPIRE program aims at the interoperability and harmonization of spatial data sets and services within Europe. The INSPIRE identifies 20 components in the data interoperability, namely principles, terminology, reference model, rules for application schemas and feature catalogues, spatial and temporal aspects, multi-lingual text and cultural adaptability, coordinate referencing and units of measurement model, object referencing modeling, identifier management, data transformation, portrayal model, registers and registries, metadata, maintenance, data and information quality, data transfer, consistency between data, multiple representations, data capturing rules, and conformance.

The INSPIRE model gives requirements for each of the interoperability components, and the recommendation of using certain standards for each component.

1. Reference model interoperability will use ISO 19101;
2. Data specification will conform to ISO 19141;
3. Application schemas, general feature model and spatial object types will be compliant to ISO 19109;
4. The feature catalogues should contain the information in the corresponding application schemas in accordance with ISO 19110;
5. The spatial characteristics of a spatial object should be expressed using standards of ISO 19109 (8.7 or 8.9);
6. The temporal schemas of the application schemas should be expressed using ISO 19108;
7. The coverage function in an INSPIRE application schema should be compliant with ISO 19123;
8. The units of measurements shall be described using the model contained in ISO 19136;
9. The metadata is an important component in the INSPIRE together with the other parts. The metadata covers three levels: discovery, evaluation and use. Being associated with individual spatial objects, the metadata descriptions are often part of the application schemas. The metadata elements should be compliant to ISO 19109 and 19115;
10. The data and information quality should be described using a model compliant to ISO 19157;
11. The data capture rules associate to each spatial object should be described in conformance with ISO 19131.

### 5.1.1.11 Summary

We reviewed a list of metadata standards in the current research infrastructure. The goal is to distinguish the gap between the metadata standards and the ENVRI RM. From the above discussion, we can summarize:

1. Metadata standards are proposed for different purposes in the lifecycle of the data, for instance for annotating data acquisition, querying data, and sharing data between communities. They may have different emphasis on the vocabularies.
2. Different forms of metadata standards exist, such as RDF, XML schema, or pure set of vocabularies.
3. If a research infrastructure covers different phases in the lifecycle of data applications, more than one metadata standard maybe involved in its development.

In order to distinguish the gap between ENVRI RM and these metadata standards, we shall:
1) Highlight the coverage of the metadata in terms of the sub-systems we defined in the ENVRI RM;
2) Identify the minimal sub set of the metadata from the standards that can be mapped to the ENVRI RM.
3) Based on the typical scenarios that the research infrastructures will support, we distinguish the possible gaps between the ENVRI RM and the metadata standards.

In the rest of the section, we will discuss these issues respectively.

5.1.2 Linking to the ENVRI RM
We shall follow the basic linking approach, and do the following steps:
1) **Select basic concepts for linking.** Since the data and metadata related information in the reference model is mainly modelled from the information viewpoint, we group the concept set at each phase of the data lifecycle: acquisition, curation, access, processing and community support. As the first phase, we treat each select metadata as a single group.
2) **Similarity check.** The similarity check will be based on the coverage of the metadata and the data lifecycle, and identify the related information objects in the reference model.
3) **Propose linking relations** to connect the information viewpoint objects and the metadata standards.

5.1.2.1 Select basic concepts for linking
We start with the analysis on how metadata we collected from ESFRI related to components in the ENVRI RM. The sub systems in the current ENVRI RM are derived from the functional partitions based on the lifecycle of the research data. Therefore, the metadata utilized in the ESFRI should necessarily cover these sub systems. Table 1 shows the analysis of the coverage.

We roughly classify the metadata standards (see main elements reviewed above) based on their potential usage on each sub system. It is not a precise classification, because the level of details of the metadata standards is different. However, such classification allows us to understand potential utilization of a metadata standard in the RI. For instance, the SensorML does not contain relevant items with data processing and community support, while Dublin Core main focuses on the publication (access) of the data but not the other sub systems.

<table>
<thead>
<tr>
<th>Metadata standards</th>
<th>Acquisition</th>
<th>Curation</th>
<th>Processing</th>
<th>Access</th>
<th>Community</th>
</tr>
</thead>
<tbody>
<tr>
<td>SensorML</td>
<td>Y</td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NetCDF</td>
<td></td>
<td></td>
<td>Y</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>ISO19115</td>
<td>Y</td>
<td>Y</td>
<td></td>
<td></td>
<td>Y</td>
</tr>
<tr>
<td>ISO19156</td>
<td>Y</td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CSR</td>
<td>Y</td>
<td></td>
<td>Y</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
In the information viewpoint, metadata related concepts are used in three sub systems: acquisition, curation and access.

<table>
<thead>
<tr>
<th>Process in data lifecycle</th>
<th>IV_Objects</th>
<th>IV_Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data acquisition</td>
<td>Specification of measurements or observations</td>
<td>Specify measurement and observation</td>
</tr>
<tr>
<td></td>
<td>Specification of Investigation Design</td>
<td>Specify Investigation Design</td>
</tr>
<tr>
<td>Data curation</td>
<td>Metadata catalogue</td>
<td>Register metadata</td>
</tr>
<tr>
<td></td>
<td>Metadata</td>
<td>Annotate metadata</td>
</tr>
<tr>
<td></td>
<td>Metadata catalogue</td>
<td>Annotate data</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Add metadata</td>
</tr>
<tr>
<td>Data access</td>
<td>Metadata</td>
<td>Publish metadata</td>
</tr>
<tr>
<td></td>
<td>Metadata catalogue</td>
<td>Data Query</td>
</tr>
</tbody>
</table>

Table 8. Metadata related information objects and actions.

We highlight the metadata and metadata catalogue concepts as a starting point for linking. Most of the other related concepts can be linked to the external metadata standards via these two concepts.

5.1.2.2 Similarity check

In the information viewpoint, metadata is defined as data about data, in scientific applications used to describe, explain, locate, or make it easier to retrieve, use, or manage an information resource. From the definition, we can see that the metadata in ENVRI RM are the extra description of specific data. And a metadata catalogue as a collection of metadata, usually established to make the metadata available to a community. A metadata catalogue has an access service. The catalogue is a service that collects metadata. The metadata description should be compliant to certain standards or their extensions, such as the external metadata standards we reviewed above.

5.1.2.3 Propose linking relations

The relations between the concepts in Table 2 and the metadata standards can be modeled at different levels. As we mentioned before, at this stage, we only focus on the relationship between the ENVRI RM and the entire metadata standard definition, instead of on relationships with individual concepts in each standard.

The most important two concepts are: IV_Metadata and IV_Metadata_Catalogue. The metadata standards in the Table 1 can be sub-classes of the metadata standards that can be used to create metadata. Therefore,
in the linking layer, a concept OEILM:Metadata_Standard is added, as shown in Fig. 14. We use the name space of OEILM to indicate the linking ontology, and ENVRI to indicate the name space of the ENVRI RM. The property of OEILM:compliant is defined to describe the relation between IV_Metadata and the standards that IV_Metadata is compliant to.

Figure 14: Linking between ENVRI information viewpoint objects and the metadata standards.

5.2 Linking to service descriptions

5.2.1 State of the art

In the service oriented architecture, service descriptions and registries are important components for publishing and discovering services. Universal Description Discovery and Integration (UDDI) is an early standard (UDDI, 2004); it models the services descriptions from three categories: 1) white paper for addresses, contacts and known identifiers, 2) yellow paper for industry categorizations, and 3) green paper for technical information of the services in the context of business logics. Web Service Description Language (WSDL) (WSDL, 2001) is a core standard in the web service architecture for describing the interface of services. OWL-S (OWL-S, 2004) is an important semantic standard to extend the service description. Service profile, grounding and model are three important concepts associated with services. Together with UDDI, a client can discover the services from the registry and obtain the interface description of the service. However, the adoption of UDDI in industry is very limited; and most of efforts ended around 2007. The category mechanisms in UDDI for entry points are overly complicated for just the simple task of discovering the address of a service. The original ambition of having a world-wide registry mechanism for promoting services faces challenges of trust, while most of the service developers are only aiming at limited audiences. Nevertheless, having a centralized information pool for service information is still important for sharing services.

In this section, we review some solutions are currently used in the ESFRI or relevant projects.
5.2.1.1 gCube

gCube (Candela et al., 2008) is a software system supporting the creation and operation of Hybrid Data Infrastructures enabling the building of Virtual Research Environments (VREs) (Candela et al., 2013). The gCube system can be deployed to exploit different Grid infrastructures, including facilities provided by the European Grid Initiative (EGI) (formerly operated by European Grid for E-sciencE project (EGEE)) (EGI, 2013).

The basic components in gCube are services, and gCube provides dynamic environment for registering, deploying, discovering and invoking services. A registry service in gCube manages information about resources and the gCube Host Nodes that provide the resources. From the registry, a consumer service can query a resource using the properties such as category name, resource name, version and entry point.

The description of a resource contains:

1) Basic parameters fields: resource id, resource name, secondary type, and description;
2) Plugin parameters fields: varying depending on the type of plugin selected. Different external metadata can be used here.

The gCube framework is currently used in ENVRI for prototyping common operations that ESFRI projects require.

5.2.1.2 CineGrid Description Language

Semantic web technologies have been recognized as important mechanisms for describing service categories, relationships between services, and relationships between services and resources. CineGrid Description Language (CDL) (CineGrid, 2013) is an example of such solution, which is developed by the CineGrid project for describing CineGrid data exchange services.

In CDL, services are categorized into different types, and relations between services and host nodes are explicated by relation cdl:provided_by or cdl:provideService. Several basic class types are defined: storage, visualizer, transcode, indexing and authentication.

5.2.1.3 Summary

In the data lifecycle, services and software components are tools for users to acquire, curate, access and process data. The functionality description and categorization of services and software components are dependent on the technical choice of resource discovery and integration. From the analysis, we can see that semantic technologies are playing important roles in service description and discovery; the description languages or schemas are intended to be open in order to include extensions of new metadata or semantic descriptions.

5.2.2 Linking to the ENVRI RM

In the ENVRI reference model, services and software tools are typically modeled as computational objects in the computational viewpoint. The linking between the ENVRI reference model ontology and the service description and registration related information models will be mainly through the computational viewpoint, since these services are mainly used to instantiate the processes defined in the computational viewpoint.
Following the same approach we did for metadata, we will also investigate the linking between the reference model and the service descriptions via three steps.

5.2.2.1 Select basic concepts for linking

The services and software components in a research infrastructure are typically modeled in computational viewpoint.

<table>
<thead>
<tr>
<th>Process in data lifecycle</th>
<th>CV_Objects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data acquisition</td>
<td>Acquisition service, Instrument controller, Field laboratory</td>
</tr>
<tr>
<td>Data curation</td>
<td>Catalogue service, Annotation service, Data store controller, Data transfer service, Data transporter</td>
</tr>
<tr>
<td>Data access</td>
<td>Semantic laboratory, Semantic broker, Virtual laboratory, Experimental laboratory, Data broker</td>
</tr>
<tr>
<td>Data process</td>
<td>Coordination service, Process controller</td>
</tr>
<tr>
<td>User community</td>
<td>Science gateway, Virtual laboratory, Security service, Experimental laboratory, PID service, Field laboratory</td>
</tr>
</tbody>
</table>

Table 9. Service related objects in the computational viewpoint.

In the current reference model, an explicit registry service is defined for hosting metadata of resources, namely catalogue service. The computational objects in general model the basic software services and tools. For the first linking, we highlight the catalogue service and abstract computational objects as the components for linking.

5.2.2.2 Similarity check

As the first step, we consider each external description language or schema as a single entity, and distinguish their relation to the objects in the reference model. The service description standards and the information models used by the catalogue services are currently not explicitly modeled in the computational viewpoint. However, these two standards are important for technical decisions for developing brokers of data and services. Extra concepts are needed to bridge this gap between the catalogue service and the external description standards.

5.2.2.3 Propose linking relations

We propose two abstract concepts in the linking ontology: catalogue description standard, and service description standard. The catalogue service and the computational objects in the reference model will have a relation called compliant to the concept of description standards defined in the linking ontology. The abstract standards will be the super class of the external service standards. Fig. 15 shows the basic idea.
5.3 Linking to Computing, Storage and Network in physical infrastructure

The information model for network topologies, devices, and QoS requirements for data services play an important role in the management of the infrastructure, and high level application planning (Zhao et al., 2012). During the past few years, infrastructure modelling has been highlighted as an important issue in several projects such as CineGrid (CineGrid, 2013), NOVI (NOVI, 2013) and Geysers (Geysers, 2013). In the CineGrid project, the UvA team has developed two ontologies for describing CineGrid services and network topologies respectively using semantic web technologies. As we mentioned, the CineGrid Description Language (CDL) (CineGrid, 2013) describes the services and resources on top of the network infrastructure\(^9\). The Network Description Language (NDL) (van de Ham, 2010) models the different levels of a network infrastructure: physical, domain, layer and topology\(^{10,11}\). Via the Open Grid Forum (OGF) working group, the NDL has been standardized as Network Modeling Language (NML). It has been adopted in the Network Service Interface (NSI) as the topology exchange schema. The Infrastructure and Network Description Language (INDL) was newly developed in the NOVI and Geysers projects to include the virtualized resources and energy model of the infrastructure. In this section we will review these existing models and discuss how they can be linked to the ENVRI reference model.

5.3.1 State of the art

Most of the information models in the current ESFRI focus on the data and data related services, there are little about the lower level of physical infrastructures such as network. On the other hand, we can see that advanced network infrastructure plays an important role in the e-Science environment to provide high quality connections between largely distributed data sensors, and computing and storage elements. By enabling large-scale data movement between different devices and nodes, applications which require extremely large-scale data movement can be enabled. However, the quality of the network connections and services has rarely been taken into account by current applications, because 1) traditional IP-based networks provide limited reservation capability for workflow engines; 2) the existing e-Science

---

\(^9\) Available at: [http://cinegrid.uvalight.nl/owl/cdl/2.0](http://cinegrid.uvalight.nl/owl/cdl/2.0)

\(^{10}\) Available at: [http://cinegrid.uvalight.nl/owl/ndl-domain.owl](http://cinegrid.uvalight.nl/owl/ndl-domain.owl)

\(^{11}\) Available at: [http://cinegrid.uvalight.nl/owl/ndl-topology.owl](http://cinegrid.uvalight.nl/owl/ndl-topology.owl)
applications assume available network connections as non-changeable services, and seek customized solutions at software level to optimize computing processes and data storage; and 3) the existing applications mainly consider the functionality of the e-Science services, and limited support has been provided for including (network) quality requirements for the services in the composition, enactment and execution of workflows.

The recent emergence of advanced network infrastructures, especially in research and educational networks, have seen a gradual shift in the type of services offered to end users and applications. Circuit-switched services play an increasingly important role in the network infrastructure, and optical and photonic devices are being employed to create circuits between service points. In this section, we will take a look at recent progresses made in the information modeling of network and infrastructures.

5.3.1.1 Network information modeling

The Network Markup Language Working Group (NML-WG) within the Open Grid Forum has gathered together experts in the area of network topology descriptions and is working towards a first standard. The modeling effort done in NDL (Jeroen, 2010) has largely been incorporated in the NML schemas, but has been enlarged and expanded by adopting concepts and models from the PerfSONAR community (Perfsonar, 2013).

The NML defines the following basic concepts:
1. Node is a machine part of the network, such as router, or a regular computer;
2. Port describes how a Node is connected to the network;
3. Link is the connection between two Ports;
4. Service defines the capabilities of a Node or Port; examples are a SwitchingService or an AdaptationService respectively.

Currently, the NML schema has been applied and extended in the OGF Network Services Interface (NSI) working group (NSI, 2013). This group has created an inter-domain interface for circuit provisioning. This NSI topology description extension allows the description of inter-domain topologies, the service-plane aspects, as well as the peering relations between different domains for circuit provisioning.

5.3.1.2 Virtual resource modeling

The goal of the Infrastructure and Network Description Language (INDL) is to capture the concept of virtualization in computing infrastructures and to describe the storage and computing capabilities of the resources. The INDL ontology is built upon the NML ontology and it uses the nml:Node concept as the basic entity for describing a resource in a computing infrastructure. INDL can be used as a stand-alone model (i.e. without any network descriptions), or it can be used in combination with NML by importing the NML ontology into the INDL definition. In the latter case, all NML concepts will become available to the user of INDL (Ghijsen, 2013).

Virtualization is modeled using the VirtualNode concept, which is modeled as a subclass of nml:Node (i.e. a virtual node inherits all properties of a node). A virtual node is also implemented on a node. This allows us to create layers of virtualization stacked on top of each other. The internal components of a node
are modeled by defining \texttt{nml:Node} to consist of a number of \texttt{NodeComponents}. The \texttt{NodeComponent} is an abstract class which describes the following essential components of machines which are of interest to the user: \texttt{MemoryComponent}, \texttt{ProcessorComponent}, and \texttt{StorageComponent}. Additional work is needed to investigate the modeling of disk partitioning, different file formats and the use of disk-images.

5.3.1.3 Software defining networking modeling

Software defined networking (SDN) is a newly developed network technology that decouples the control and data transfer of the network devices, and allows applications to manipulate network behaviour at a high level, such as modifying the underlying network packet forwarding table and controlling the network quality of service. There are several SDN technologies available; OpenFlow (Kotani et al., 2012) and Network Service Interface (NSI) (NSI, 2013) are two typical examples.

The OpenFlow protocol was proposed to separate the control plane from the switch and to enable the remote control on packet routing and other high-level behaviours. It is an open standard for enabling new ways of managing networks such as bandwidth and provisioning, access controlling, latency controlling, and constructing non-IP networks (Kotani et al., 2012). After being standardized in 2011, more and more switch vendors are adopting it in the new products.

Network Service Interface was created as a result of collaborative development of network and application engineers primarily associated with the Research and Education (R&E) community. The NSI provides not only protocols for different domains to exchange topology information for inter-domain path selection, but also software services for selecting, reserving and provisioning network paths. The NSI Working Group was established in the Open Grid Forum in 2008 (NSI, 2013), and since then the group has been working mainly on the NSI Connection Service standard.

These SDN technologies complement research infrastructure with several new opportunities for support data intensive applications, for instance, 1) customizing network connectivity between services by defining suitable flow forwarding table, or by reserving dedicated links, 2) virtualizing the network resources for different resource partition by tuning the match between network slice and the requirements of computing and storage nodes, and 3) controlling the network quality of service by either advanced reservation of links or controlling the package flows.

In environmental sciences, data quantities are getting increasingly large, it is thus very important to include the SDN in the ENVRI reference model. In our recent work, the SDN controller is modeled as service type and in the extended version of CDL (Jiang et al., 2013).

5.3.1.4 Information centric network

Other development of network infrastructures, such as programmable network and Information Centric Network (ICN) provide new opportunities delivering data content and planning complex application workflows. For instance, there are several on-going projects aim at this goal, such as PSIRP (Psirp, 2013), 4WARD (4Ward, 2013), and PURSUIT (Pursuit, 2013). Linking the state of the art of the underlying infrastructure with data and high-level services allows application schedulers to utilize the latest
infrastructure technologies to optimize application development. The research on content centric network is still in the early stage. In the project, we will investigate the usage of ICN in the data distribution and delivery in research infrastructure.

5.3.1.5 Quality of service, energy consumption

The total energy consumption, the energy sources used and pragmatically the cost of power are becoming increasingly important factor in the planning, design and operation of ICT infrastructures. The use of virtualization and the possibility to rely onto IaaS offerings provide a new dimension when trying to manage power and increase greenness of an infrastructure.

The Energy Description Language (EDL) (Zhu, 2012) is an OWL ontology that describes all the energy related parameters and QoS attributes that are needed for power management in such large scale (virtual) infrastructures. EDL supports several power management scenarios; among others for example:

5.3.1.6 Summary

In this section, we reviewed the information models of infrastructure. These models focus on different levels of issues in the infrastructure: network topology, network device, resources, virtual resources, and the quality of the services. Fig. 16 shows the basic relation of the models and available solutions.

5.3.2 Linking to the ENVRI RM

Linking application level information with the underlying infrastructure details will allow brokers to select data sources and services with consideration of network details. In this section, we discuss the possible links between infrastructure information models and the reference model.
5.3.2.1 Selected basic concepts for linking

In the previous sections, we discussed that the data related information in the ENVRI RM is typically modeled in the information viewpoint, and the software tools and services related information are modeled in the computational viewpoint. These ENVRI RM concepts can be linked with the infrastructure description through several means: via the technology viewpoint as a new technology, via the computational viewpoint as new services, or other options. In the current model, the technology viewpoint has not yet been constructed – therefore we will investigate the linking first via the computational viewpoint.

Looking back at Table 2, we asserted that an explicit catalogue service is defined in the current reference model for hosting metadata of resources, and computational objects in general model the basic software services and tools. For the first linking, we highlight the computational objects as the component for linking.

5.3.2.2 Similarity check

From the review on the infrastructure information models, we can see a clear hierarchical structure. The concept of service and device are two important classes for linking different layers. In the reference model, the physical devices are not directly modeled. Checking the similarity, the computational objects are similar concepts in the ENVRI RM of services.

5.3.2.3 Propose linking relations

To bridge the semantic gaps between computational objects with the infrastructure description, the linking ontology uses two abstract classes: service and host. These two concepts indicate the basic mapping between application level software and the physical devices. Fig. 17 shows the basic linking relations.

![Diagram of linking the infrastructure description languages.](image)

The service in the CDL will be an equivalent class to the service in the linking ontology. The concept of resource in the INDL will be equivalent to the host defined in linking ontology.

5.4 Summary

In this section, we discussed the three clusters of concepts that can be linked to the ENVRI reference model. Currently, the development of the linking ontology is still in an early phase. The selection of the
concepts for similarity checking demonstrates the linking only at a rough level. The development will be conducted in an iterative way; the evolution of the reference model and its utilization for future use cases will push the refinement of the linking ontology. Fig. 18 shows the current linking ontology.

Figure 18: The current definition of linking ontology.
6 VALIDATION

OEILM is developed and validated in an iterative way. The ontology will be validated in terms of three different criteria:

- **Expressiveness.** Experts from different RIs will be invited to provide natural language descriptions of the scenarios in RIs using OEILM. The feedback will be collected and analyzed to improve the ontology and the reference model itself.

- **Interoperability.** The descriptions of different RIs will be compared to discover the commonality of the operations. The resulting feedback will be used to improve the scope of the reference model and the ontologies in OEILM.

- **Application.** OEILM will be tested in application planning systems or resource brokers to validate the linking between data and resource descriptions across RIs.

In this section, we will briefly describe the ongoing validation work.

6.1 Expressiveness

A web based collaborative ontology editing tool has been prototyped and used in the project to validate ontologies. Fig. 19 shows a screenshot of the tool. The left side of the tool provides the tree structure of the ontology, and the central part of the tool allows users to draw and drop the concept displayed on the right side and instantiate an instance of the concept. The tool can save the composition online and allows other users to view and open it. Using the tool, OEILM will be released to different users in the ENVRI community for feedback.
The tool is online at [http://envriontology.appspot.com/main/](http://envriontology.appspot.com/main/).

### 6.1.1 Consistency between natural language descriptions

The ENVRI reference model uses a lot of processes in the data lifecycle that we identified from analysis of different research infrastructures. The natural language descriptions of those processes will be an important input for checking the expressiveness of the OEILM ontology. In this way, we can validate if the OEILM ontology is consistent with the natural-language description of the reference model. Fig. 20 shows an example of the data acquisition process. The top left shows the acquisition process described in the natural language of the reference model, and the bottom right shows the semantic composition of the process using the OEILM ontology.
We have performed a check for all the processes described for the computational viewpoint in http://confluence.envri.eu:8090/display/ERM/Computational+Viewpoint. From the exercise, we did find some inconsistencies with the ontology definition. All these issues have been solved by adding missing properties or concepts, or by correcting typographic errors in the concept or property names. With this refinement, we can see that the current OEILM ontology can describe all the processes covered by the reference model. Fig. 21 shows the list of processes being described using the ontology.

The final consistency level between ontology and the ENVRI RM is what we expected, because the ontology is derived from the same process descriptions from which we derive the reference model. The ontology instance file generated by the validation procedure can also be used by existing ontology-based composition and reasoning tools.
6.1.2 Describing the processes developed by the other project teams

In the project, a list of common operations performed in ESFRI research infrastructures is being prototyped in work package 4. Describing these operations and processes using the OEILM ontology will not only show the usability of the ontology but also validate the consistency between these prototypes and the ENVRI RM. Currently, the open search process and the data integration are two processes being focused.

6.1.2.1 Example scenario of data harmonization

We start with the data harmonization component developed by the team in the task 4.2 of the project (Tarasova et al., 2013). Data harmonization is a process of making data from heterogeneous data sources compatible, comparable and interoperable and, thus, useful for further data integration. This task uses existing software components and develops approaches to enable integration and harmonization of data resources from cluster's infrastructures and publication according unifying views.

The development is conducted in the use case "Iceland Volcano Ash" (ENVRI, 2011). The goal is to support scientists to analyse Iceland behaviour using data provided by different research infrastructures during a specific time period. The basic scenario is described in http://staff.science.uva.nl/~ttaraso1/html/envri.html#model.

The data harmonisation component is prototyped based on the RDF Data Cube (The RDF Data Cube vocabulary http://www.w3.org/TR/2013/) approach and uses the ENVRI vocabulary (http://data.politicalmashup.nl/RDF/vocabularies/envriCR-vocab-data-cube-20130625/). The ENVRI
The harmonization component 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.

![Diagram showing the basic architecture of the data harmonization component](image)

**Figure 22:** The basic architecture of the data harmonization component (Tarasova et al., 2013).

To retrieve the data from different sources they have to be first mapped to a harmonized model (the ENVRI vocabulary). Following the analysis of the environmental data different common structural concepts are identified (e.g. observation, dataset, metadata attributes – see Fig. 23).
Mapping can be explained on the following observation statement focusing on the term “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 to it the URI http://www.eionet.europa.eu/gemet/concept?ns=1&cp=245 to it (entity naming). The GEMET concept of air is then defined as an instance of envri:FeatureOfInterest (entity typing12). 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 (Virtuoso, 2013) and are ready to be queried at a SPARQL-endpoint (via http://data.politicalmashup.nl/sparql/).

---

12 According to the guidelines for O&M the feature of Interest (FoI) (INSPIRE Cross Thematic Group, 2013) should be in this case the air bubble surrounding the intake modelled as a SpatialSamplingFeature, derived from SamplingFeature. This insights should be considered when reviewing the ENVRI vocabulary ontology.
6.1.2.2 Data harmonisation in the ENVRI RM

The data harmonization process described above can be mapped directly onto the Reference Model. The Information Viewpoint models within the data curation subsystem the mapping of data according to mapping rules which are defined by the use of local and global concept models. Ontologies and thesauri are defined as conceptual models, all widely accepted models as GEMET, O&M, Data Cube are declared global conceptual models whereas the ENVRI vocabulary is specified as a local one, because it has been developed within the current project without being yet accepted by a broad community. Although Fig. 24 shows that this process before data are published, this process can be in fact done at any stage of the data life cycle. In the example described above it is performed after being published and before being queried.

![Diagram showing the data mapping process within the Information Viewpoint in RM](image)

6.1.2.3 Describing the process using the ENVRI RM ontology

Describing a process using the ENVRI RM ontology is to instantiate the concepts that can be mapped to the process. Fig. 25 illustrates the instantiation (all boxes with a dashed line) of the ENVRI RM ontology focusing at the harmonization process described above. The same could be demonstrated for the EuroArgo dataset with the feature of interest being ocean. For each part of the observation mapping rules have to be defined to be able to query both datasets at a certain time period.

Fig. 10 and 11 show 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.

Figure 25: Describing the data harmonisation process in the task 4.2 using the ENVRI RM ontology.

<table>
<thead>
<tr>
<th>Component/Object in Task 4.2</th>
<th>Information Object in RM</th>
</tr>
</thead>
<tbody>
<tr>
<td>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</td>
<td>Specification of measurement and/or observation</td>
</tr>
<tr>
<td>GEMET:245 is instance of FeatureOfInterest class</td>
<td>Mapped data</td>
</tr>
<tr>
<td>GEMET, O&amp;M, DataCube</td>
<td>Global concept model</td>
</tr>
<tr>
<td>ENVRI vocabulary</td>
<td>Local concept model</td>
</tr>
<tr>
<td>FeatureOfInterest (ENVRI vocabulary)</td>
<td>Local concept</td>
</tr>
<tr>
<td>Component Property, GEMET:245, FeatureOfInterest (O&amp;M)</td>
<td>Global concept</td>
</tr>
<tr>
<td>GEMET:245 create as instance of FeatureOfInterest class</td>
<td>Mapping rule</td>
</tr>
<tr>
<td>ICOS data CO2 of air, EuroArgo data ocean temperature</td>
<td>Published data</td>
</tr>
</tbody>
</table>

Table 10. Components/Objects in Task 4.2 versus Information Objects in the ENVRI RM
Operation in Task 4.2 | Information Action in RM
--- | ---
Build ENVRI vocabulary as extension of DataCube and on basis of O&M concepts | Build local conceptual model
Define rule: GEMET:245 create as instance of FeatureOfInterest class | Setup Mapping rule
Perform Mapping using Google Refine | Perform Mapping
SPARQL query: http://staff.science.uva.nl/~ttaraso1/html/queries/Q1.rq | Query Data

Table 11. Operations in Task 4.2 versus Information Actions in the ENVRI RM

6.2 Interoperability
An ongoing work is to describe different ESFRI infrastructures using the OEILM ontology. It will be an important step for validating the reference model and for promoting it to the ESFRI projects. It will be a follow up of the work described in 6.1, and the results will also be an important way to promote the reference model to the ESFRI membership.

Currently, the description of research infrastructures is still ongoing. One important strategy is to start with a research infrastructure from which we can obtain sufficient information, and then promote it as example to the other infrastructures when trying to construct more indepth descriptions.

This will be an important task for the next phase.

6.3 Application
The validation of OEILM for real applications is still in a very early phase. The technical prototype will be on a limited configuration of the test-bed. Applying network quality analysis in application planning for research infrastructures will bring a new dimension of performance tuning for applications. OEILM allows application brokers to search software components and data sources by network-level quality characteristics. However, it also requires that the network provide controllability over its quality, using means such as Software Defined Networking. We have constructed a test-bed using the Openlab facility provided by the University of Amsterdam. A prototype for a data delivery service based on OEILM and the test-bed are under construction (Jiang, 2013).

In the next phase, we will also investigate if we can apply OEILM to the brokers of the gCube environment that is currently used in development by other teams in the ENVRI project.

6.4 Discussion
The development and validation of OEILM is still in progress. Currently, the first version of OEILM is available. We have partially validated the ODP and ENVRI RM ontologies. Since the OEILM ontology
heavily depends on the reference model itself, any refinement on the reference model will require further iterations of development in OEILM. Greater collaboration with different development teams in the ENVRI project will be very important for developing and validating the linking model.

From the current status of the work and the validation so far, we can see the OEILM can match the expressivity of the natural-language description of the ENVRI reference model. It brings a stable foundation for the development of a linking ontology. The main objective of the next phase will be refining and validating the linking ontology, and applying it to the real use cases that the other ENVRI development teams are currently focused on.
7 SUMMARY

7.1 Current status

In this deliverable, we introduced the ongoing work in the project of ENVRI. A semantic linking framework called OEILM is proposed to enhance the coupling between RI and information about data and the underlying storage, computing and network elements. From the current status of the work, we can summarize:

1. Interoperability is an important requirement for supporting semantic integration between data and services between different research infrastructures. An effective reference model synchronizes terminologies defined in different environmental RIs, and guides the further development of common operations and functional components in the infrastructure.
2. Modeling a distributed system like a research infrastructure requires decomposition of modeling issues based on different stakeholder viewpoints. The Open Distributed Processing model provides a suitable mechanism to do this.
3. Semantic web technologies provide an open framework for modeling linking between different elements in research infrastructures. A semantic linking framework is important for realizing interoperability between research infrastructures, and the Open e-Science Information Linking Model is a step towards the direction.

The research of OEILM is still ongoing; the development of the framework still faces several challenges. As a rapidly growing field, the development of services in data oriented research infrastructures is driven by research activities both inside and outside the intended scientific domain. How to model the as-yet-unknown facets of future ESFRI projects and keep the ENVRI reference model open and extensible will remain important questions which must be answered if the model is to maintain its value for end-users. Moreover, the data-driven e-Science experiments that the ESFRI infrastructures intend to support often require customized processing services for special research purposes; how to balance the constraints of developing new services and adapting existing ones from other research infrastructures will be important issue when promoting the ENVRI RM and OEILM to the developer community targeted by the ENVRI project.

7.2 Future work

In the next phase of the work, we will concentrate on several issues:

1. First of all, the linking between the reference model and other metadata standards, and the service ontology, will continue. This will not only contribute a richer set of the linking solutions to the ESFRI projects, but will also be the next step for refining the linking ontology.
2. Second, the evaluation of OEILM will be continued. Domain scientists from different ESFRI projects will be invited to evaluate the linking ontology. In the meantime, applying OEILM to the use cases will be another important validation step.
3. Third, OEILM will be incorporated into the dissemination materials produced for the ENVRI reference model. It will also be considered as an important step for disseminating results to the ESFRI community.
8 RELATED PUBLICATIONS


9 REFERENCES


Tarasova T., Argenti M., and Marx M.: Semantically-Enabled Environmental Data Discovery and Integration: demonstration using the Iceland Volcano Use Case. To appear in proc. of the 4th Conference on Knowledge Engineering and Semantic Web (KESW), Saint-Petersburg, Russia, 2013


WSDL (2001) http://www.w3.org/TR/wsdl


## A. Glossary

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