ENVRI Common Operations of Environmental Research Infrastructures

ENVRI

Services for the Environmental Community

ENVRI Sustainability Document

Produced in May 2014 by the ENVRI Sustainability Team

This document is produced in May 2014 by the ENVRI sustainability team:

Antti Pursula, CSC, **Marie Sandberg**, CSC, **Samuel Keuchkerian**, CNRS, **Carmela Asero**, EGI.eu, **Ari Asmi**, UHEL, **Yannick Legre**, EGI.eu

with contributions from:

Alex Hardisty, CU, **Malcolm Atkinson**, UEDIN, **Roberto Cossu**, ESA, **Leonardo Candela**, CNR

Contact details:

ENVRI Sustainability Task leader Antti Pursula: *antti.pursula@csc.fi*

ENVRI website: http://www.envri.eu/

1. COPYRIGHT NOTICE

2. EXECUTIVE SUMMARY

The ENVRI Sustainability Document describes the main strategy for sustaining the achievements of the ENVRI project. The most viable alternatives for keeping the ENVRI project outputs operational beyond the project life time are analysed and summarised, leading up to a number of concrete recommendations for the way forward. This document is intended for ENVRI partners as well as its various stakeholders, including the European Commission, current and emerging Research Infrastructures and their user communities, service developers and e-Infrastructure providers.

The introduction gives an overview of the role of ENVRI in the research infrastructure landscape, describing collaboration within ENVRI itself as well as contribution to international cooperation initiatives such as the Group on Earth Observations (GEO), US based EarthCube and the EU funded COOPEUS project.

The e-Infrastructures and e-science initiatives EGI, EUDAT, Helix-Nebula and RDA are given special attention due to their role as key partners, stakeholders and service providers for the environmental cluster. EGI and EUDAT provide many services and capabilities that are suitable for the current and emerging Environmental Research Infrastructures and involving these in providing platforms and services for RIs will in many cases be very beneficial, especially for data discovery and access across research infrastructures. Therefore, a special subsection is devoted to describe concrete examples and views how Research Infrastructures can benefit from these services and capabilities provided by EUDAT and EGI.

The core sections of the document analyse the benefits of the results delivered by the ENVRI project and provide strategies to sustain these outcomes. The results are grouped into three areas: 1) policy level results, 2) the ENVRI Reference Model and 3) technical results.

1) The core objective of the *policy level results* is to develop and maintain a common policy for the ESFRI Environmental RIs. The main vehicle for this is the establishment of the ENVRI Stakeholders Advisory Board (SAB). By forming the ENVRI SAB, the project has brought together European Research Infrastructures and other key participants to communicate and plan long-term solutions for better collaboration and integration. One of the key outcomes of this strategic level cooperation is the development of the Environmental Research Infrastructure Strategy (ERIS), which describes the vision of the whole Earth System sciences, aiming for the year 2030.

2) The development of the *ENVRI Reference Model* hosted at www.envri.eu/rm, provides the ESFRI Environmental Research Infrastructures with a common ontological framework for the description and characterisation of ICT infrastructures and is key to maintaining interoperability between the RIs on the system level.. The ENVRI RM provides predefined professional framework to clearly define roles and processes in RI operations. This makes it far easier to design RI in the construction phase, and helps to evaluate current research infrastructures for division of tasks and finding missing or duplicated actions within the RI work. Both benefits and costs of the ENVRI RM are analysed in a preliminary Cost-Benefit analysis from an immediate (1-5 years) and an intermediate (5-10 years) time perspective. Benefits and costs often occur at different points in time – sometimes years apart, which is a

relevant factor for sustainability planning. A few case studies of the ENVRI RM used by the ESFRIs are also presented to give concrete examples of already materialised benefits.

3) The focus of *ENVRI technical results* is in data discovery and data access across communities. The main output of the project in this area are technologies that enable building platforms, solutions and services to match the Research Infrastructures' common needs and to advance integrated data-driven science. ENVRI has also produced prototypes and pilots of a few services that showcase the potential of the underlying technologies and tools.

The various outputs of the ENVRI project all have different outlooks for sustainability. However, all of them depend ultimately on the interest of the Research Infrastructures to adopt, use, develop, support, and/or endorse these results. The success of ENVRI is much dependent on including IT partners into the consortium even if they are not directly part of Environmental RIs. The analysis shows that the current partners of ENVRI are committed to continue collaboration within the Environmental Research Infrastructure cluster, including development of ENVRI technical results and the Reference Model. The ENVRI RM is dynamic by its nature and would benefit greatly from user involvement in its further development. This naturally requires commitment from the customer community side supported preferably with public funding. Importantly, the ENVRI RM is applicable to Research Infrastructures in general and in this way the investment to its development can bring benefits widely to European research. The maintenance of the ENVRI Reference Model may also benefit from community efforts, possibly also through RDA.

Three main methods to utilize and adopt the Reference Model, tools and technologies developed in the ENVRI project are identified and summarised in the document:

1. 'As a service' model: The RIs acquire the functionality they require from an external service provider
2. Open source model: The RIs take the technology and deploy it within their operations
3. Collaborative model: The RIs and ICT providers further develop and setup services in a collaborative effort

The approach at each RI is probably a combination of two or three models. Particularly, the collaborative model is crucial for the future success of the Environmental Research Infrastructure cluster, because in absence of collaboration the activities towards interoperability and integration of data will be more difficult which will slow down the progress of Environmental science.

Six recommendations towards facilitating the sustainability of the project results are summarized at the end of the document. It includes suggestions to extend collaboration of ESFRIs Environmental cluster beyond the current ENVRI project, partly through an EC funded follow-up project to ENVRI from 2014 on. The ENVRI Reference Model is seen as a highly beneficial tool for the RIs to plan, manage and develop their operations and the maintenance and further development should be secured by collaborative efforts preferably under public funding. The recommendations also underline that the adoption of ENVRI outcomes at the Research Infrastructures requires action from the RIs themselves. The developers are willing to continue their efforts, which requires commitment from the user communities.

TABLE OF CONTENTS

1 INTRODUCTION

1.1 Purpose and abstract

The goal of the ENVRI Sustainability document is to analyse the alternatives for keeping the ENVRI project outputs operational after the end of the current project. In doing so it presents the view of the project on how its results can be sustained and fully exploited. The information contained in the document is aimed for current and emerging Research Infrastructures and their user communities, for the ENVRI project partners, for the European Commission, as well as for other stakeholders and persons in the Research Infrastructure field.

The document first outlines the role of ENVRI project within the Research Infrastructure landscape. This is followed by a brief summary of project outcomes as well as an analysis of their benefits and impacts on the current and emerging Environmental Research Infrastructures. Finally, the different alternatives for sustainability of the outcomes are outlined. The document concludes with a few concrete recommendations that will help to secure the sustainability of the outcomes.

The document is prepared by the ENVRI sustainability team that has worked in interaction with the project partners over a number of iterations of the text and at several project meetings. The Stakeholder Advisory Board of ENVRI has given valuable input and insight to the sustainability planning during the work. The final document is endorsed by the SAB and approved by the ENVRI management committee.

1.2 Project summary

The ESFRI Environmental Research Infrastructures are constructing a major part of the European research landscape in environmental science for the coming 20 to 30 years. Realising this potential will ensure that the scientific community derives full value from investments in these large-scale environmental projects, and will keep scientists at the forefront of global research as they tackle the scientific challenges ahead.

The ENVRI project, “Common Operations of Environmental Research infrastructures” is a collaboration within 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 project puts emphasis on synergy between advanced developments, not only among the infrastructure facilities, but also with ICT providers and related e-science initiatives.

The ENVRI objectives are targeted at a few priorities of common interest for the ESFRI Environmental Research Infrastructures projects, mainly to create common solutions for a range of shared problems, and to increase the interoperability of the research infrastructures to serve interdisciplinary users. The ENVRI project will not create a separate new facility, but

is addressing capabilities that can be taken up by each ESFRI infrastructure or a combination of infrastructures aiming at the delivery of common priority services to users.

Given its budget and time line the ENVRI project can only provide a very limited contribution to the full construction of the ESFRI infrastructures, the selected tasks are very relevant for them. With this consideration, the key objectives of the ENVRI project are:

1. Develop and maintain a common policy for the ESFRI Environmental Research Infrastructures;
2. Design guidelines for a Common Reference Model to promote infrastructure interoperability;
3. Provide common solutions for data discovery, data integration and harmonisation for the infrastructure projects.

These three target areas are also the basis for the organisation of this document.

2 ENVRI IN THE RESEARCH INFRASTRUCTURE LANDSCAPE

2.1 ENVRI in the context of ESFRI roadmap and projects

Environmental research is by its own nature multidisciplinary and the scientific data produced is highly heterogeneous and generated by a huge ensemble of diverse instrumentation and observations. This adds complexity to data storage, preservation, and transformation via computing-intensive modelling tools for a wide range of scientific users.

Scientific questions drive a wide variety of experimental and data-driven methodologies. The environmental community recognises that long-term observations are essential to analyse patterns and trends of continuous environmental changes. This need is leading to a massive increase in the volume and transfer rates of environmental data.

Environmental science community places high emphasis on the need for adopting common solutions and tools for the management of heterogeneous data and data flows within environmental (ESFRI) Research Infrastructures.

The objective of the ENVRI project is not to build an entirely new (e-)infrastructure system. It rather aims at building collaboration within the cluster, at optimising interoperability, and at providing a portfolio of tools, services and data products available for the environmental communities, with the key support of expertise offered by main pan-European e-Infrastructures.

In the last two years several of the environmental research infrastructures achieved a different level of maturity as highlighted by the high level expert group report "Assessing the projects on the ESFRI roadmap" published in August 2013[i]. EISCAT_3D, Euro-Argo and IAGOS are likely to be ready for implementation by 2015 provided that certain specific points, identified in the report are addressed. Some more substantial effort will be needed for LifeWatch and EPOS to reach maturity by 2015. The outputs of the ENVRI project can support all environmental projects to reach their implementation targets. Additionally, they are applicable to other kinds of research infrastructures as well, as is evidenced by the discussions that have taken place between ENVRI and the other cluster projects (DASHISH, CRISP and BioMedBridges) to identify and recognise common challenges faced by all infrastructures.

2.2 Collaboration within ENVRI

Within the ENVRI cluster different levels of collaborations have been established encompassing ESFRI Environmental projects participating in ENVRI (EISCAT_3D, EMSO, EPOS, Euro-Argo, ICOS, LifeWatch), associated ESFRI projects (SIOS, IAGOS, COPAL)

and other associated initiatives (ACTRIS, ANAEE, InGos, IS-ENES). These initiatives have been invited to the Stakeholder's Advisory Board of ENVRI.

Through ENVRI, such projects have participated in a wide range of activities, also in synergy with existing e-Infrastructures, such as those coordinated by EGI, to address identified needs and to implement common solutions of the environmental ESFRI projects, with a special focus on issues such as architectures, metadata frameworks, data discovery in scattered repositories, visualisation and data curation.

ENVRI solutions will empower the users of the collaborating environmental research infrastructures and enable multidisciplinary scientists to access, study and correlate data from multiple domains for "system level" research, thus also contributing to the nine Social Benefit Areas[ii] identified and addressed by GEO-GEOSS[iii].

To achieve such objectives ENVRI has built on existing tools by extending them to provide service components for participating research infrastructures. These tools included GENESI-DEC and OpenSearch technologies and D4Science technology used in tools for harmonising, integrating and analysing data.

The work carried out within this wide collaboration requires effort and focus on ensuring basic level of interoperability within the infrastructures in the cluster and the development of common elements for a reference model, which all of the facilities can adopt to align their data collection, storage and processing needs. The ENVRI reference model provides a well-defined vocabulary and framework for discussing, designing and organising the construction of the required ICT e-Infrastructures.

2.3 Links to e-Infrastructure and e-Science initiatives

Since its launch, ENVRI has worked closely with foundational e-Infrastructures such as EGI (European Grid Infrastructure) and EUDAT (European Data Infrastructure) both considered as key partners, stakeholders and service providers for the ENV cluster. During the third year of the project, collaboration has been established with Helix Nebula to investigate services offered through the Helix Nebula platform which are of potential interest to ENVRI. Initial links have also been established with RDA (Research Data Alliance), mainly through the European Support project RDA Europe.

EGI provides integrated computing services to European researchers, sourced from distributed and federated resources. EGI.eu is an IT partner in ENVRI. EUDAT provides services for research data management, access and preservation, as part of its pan-European Collaborative Data Infrastructure (CDI). Helix Nebula set up a pan-European Public-Private Partnership, involving e-Infrastructures such as GEANT and EGI together with commercial cloud providers, to establish a sustainable European cloud computing infrastructure, meeting the needs of European scientific research. RDA aims at building the social and technical bridges that enable open sharing of data. This is achieved mainly through the development, adoption, and deployment of infrastructure, policy, practice and

standards. The emphasis of RDA is on the adoption and deployment, rather than on the development of a new infrastructure.

The progress of collaboration between ENVRI and the above mentioned e-Infrastructures is summarised in the following subsections.

2.3.1 EGI

EGI supports the digital European Research Area through a pan-European e-Infrastructure for grid and cloud computing based on federated resources and services from the National Grid Initiatives (NGIs).

ENVRI has been working with EGI.eu on data integration, harmonisation and publication facilities, dissemination and training, sustainability plans and liaison with external stakeholders outside of the ESFRI environmental research infrastructures.

An Environmental & Biodiversity "virtual team" has been created within EGI to collect information and to develop an extensive overview about Environmental and Biodiversity research communities present in the participating NGIs as well as Environmental ESFRI participants in any country/NGI they belong to. The virtual team has worked on identifying Environmental and Biodiversity communities' needs with respect to EGI resources and services.

For this activity two ESFRIs (EISCAT-3D, EURO-ARGO) with data processing requirements were selected. In cooperation with EISCAT-3D and EURO-ARGO representatives in ENVRI, EGI.eu investigated the most suitable solutions for data pre-processing of primary data and post-processing toward publishing. The main goal of this activity is to demonstrate on the real case how ESFRI's can benefit from using the EGI e-Infrastructure.

A "Sustainability of Environmental Research Infrastructures" workshop is planned at the EGI Community Forum on 20 May 2014[iv] to discuss common challenges within the research infrastructure communities and ways of sustaining solutions in the long term. The workshop will provide an overview of the common solutions developed in ENVRI including the ENVRI Reference Model for interoperability and software components for data discovery, access, integration and publication.

2.3.2 EUDAT

ENVRI project has continuously been interacting with the EUDAT project[v]. EUDAT aims to tackle the specific challenges of data management, and to ensure a coherent approach to research data access and preservation to realise a vision of a multidisciplinary collaborative data infrastructure.

The work carried on with EUDAT has mainly been focusing on identification of data infrastructure requirements including integrated data discovery across various catalogues and (near) real-time data handling. For instance, the requirements for handling real-time and near real-time data – areas identified as common, prioritised challenge in ENVRI – were discussed in EUDAT 2nd User Forum in March 2013, and a multi-disciplinary working group on dynamic data has been established[vi].

ENVRI intends to continue collaboration with EUDAT on aspects of common interest such as managing the growing amount of data, improving interoperability between infrastructures and across disciplines, as well as promoting collaboration and clarifying roles and responsibilities between research infrastructures and e-infrastructures.

EUDAT provides currently four services aimed at supporting research data management: B2Find for finding research data, B2Safe to safely replicate research data, B2Share for storing and sharing of research data, and B2Stage to transfer data to computation. These services have been applied as a pilot for EISCAT data within ENVRI project and are currently being piloted at a wider scale by EPOS.

2.3.3 Helix – Nebula

Helix Nebula[vii] established a very innovative Public Private Partnership for a Scientific Cloud Computing Infrastructure for Europe in 2011. The partnership includes large European Research Organisations, pan-European e-Infrastructures and leading commercial providers in Europe, working together on a new hybrid cloud computing paradigm to federate public and private resources within an architecture based on open standards, to meet the high requirements of data intensive science and thus bolster the data driven economy.

Helix Nebula suppliers are now launching the Helix Nebula marketplace (HNX) offering cloud IaaS services and planning to evolve toward an Information as a Service model, initially developed around Earth Observation data under the ESA Super Site Exploitation Platform.

First contacts between ENVRI and Helix Nebula started at the end of 2013. Possible use of Helix Nebula services by ENVRI and related requirements will be discussed at the Helix Nebula and ENVRI sustainability workshops both co-located with the EGI Community Forum in May 2014.

2.3.4 RDA

ENVRI has initiated discussions with the RDA Europe project[viii], the European support project to the global Research Data Alliance[ix] (RDA) initiative. The mission of RDA is to build the social and technical bridges that enable open sharing of data, mainly through the development, adoption, and deployment of infrastructure, policy, practice and standards. ENVRI participated actively in the RDA launch in March 2013, and has been part of various RDA Working Groups and Interest Groups. In particular, ENVRI has been accepted as one of the use cases by the RDA Data Foundation and Terminology (DFT) Working Group, which has been set up to gather emerging requirements as well as to test research outcomes. In preparing the use case, researchers and architects from two ENVRI-participating research infrastructures, EMSO and EPOS, used the terms and concepts defined in the ENVRI Reference Model to describe architectural features of their research infrastructures. This example is described further in the subsection 4.1.3 “Already materialised benefits for ESFRIs”.

2.4 International collaboration

In general, ENVRI contributes in the European support to the international initiative of the Group on Earth Observations (GEO) that is coordinating efforts to build a Global Earth Observation System of Systems (GEOSS). GEO was established in February 2005 by the

Third Earth Observation Summit in Brussels in response to calls for action by the 2002 World Summit on Sustainable Development and the Group of Eight (G8) leading industrialised countries. GEO is a voluntary partnership of governments and international organisations providing a framework where these partners can develop new projects and coordinate their strategies and investments. The collaboration is for example visible by ENVRI project's participation in GEO European Projects' Workshop (GEPW7), on 8-9 April 2013 in Barcelona.

In the context of specific EU and US collaboration on Environmental Research Infrastructures, ENVRI is involved in the EarthCube initiative by NSF to build a Community-Driven Data and Knowledge Environment for the Geoscience, and in the COOPEUS project, funded under the Research Infrastructures action of the 7th Framework Programme for Research and Innovation of the EU. COOPEUS brings together scientists and users involved in Europe's major environmental related research infrastructure projects, i.e., EISCAT, EPOS, LifeWatch, EMSO, and ICOS, with their US counterparts that are responsible for the NSF funded projects AMISR, EARTHSCOPE, DataONE, OOI and NEON.

The intention is that by interlinking these activities new synergies are generated that will stimulate the creation of a truly global integration of existing infrastructures. The key to this integration process will be the efficient access to and the open sharing of data and information produced by the environmental research infrastructures. COOPEUS will also serve as test-bed for new standards and methods. Therefore links to the ENVRI project in Europe and EarthCube in the US, both developing relevant architectures, are considered pivotal for the success of this international initiative.

ENVRI has been represented at a joint EUDAT and COOPEUS Strategic workshop on future harmonisation of data sharing among Research Infrastructures at EGU 2013. The workshop brought together EU and US-infrastructure projects and their international collaboration partners involved in improving international data sharing among environmental research infrastructures. The aim was to identify common goals and objectives with regard to harmonisation of data sharing among these research infrastructures and define necessary steps to demonstrate the benefits of improved data infrastructure collaboration. The workshop brought together the EUDAT, COOPEUS, ENVRI, EarthCube and RDA Europe communities.

The mentioned partners with whom ENVRI interacts find value in this interaction and would need to build an alternative means of establishing and maintaining commonalities with the Research Infrastructures if ENVRI were not sustained.

3 RESULTS DELIVERED BY THE ENVRI PROJECT

3.1 Policy level results

The aim of the policy activities of the ENVRI project can be summarized as to develop and maintain a common policy for the ESFRI Environmental Research Infrastructures. The main vehicle to achieve the goal is the establishment of the ENVRI Stakeholders Advisory Board (SAB), the high-level discussion and decision body of ENVRI. The SAB members are the coordinators of Research Infrastructure projects acting on behalf of the future users and owners of the results of the ENVRI project. The SAB is the forum to discuss common visions, strategies, to promote collaboration and striving toward interoperability as well as sharing of the best practices, taking into account different development levels and interests of participating ESFRI research infrastructures. Thus, a highly significant outcome of the ENVRI project is the establishment of a collaboration forum with a vision on sustainable collaboration of Environmental Research Infrastructures beyond the ENVRI project's duration.

In addition to the SAB, the science policy results are one of the key products of the ENVRI project. They are characterised by the building of trust and collaboration network within the environmental research infrastructures, emerging new observation networks, and representatives of major other players in the Earth System observations. As a specific product, within the project the SAB helped directly to build a common Environmental Research Infrastructure Strategy (ERIS) for the European research infrastructures and Earth System sciences in general.

3.2 ENVRI Reference Model

ENVRI advances system-level interoperability of research infrastructures through the development of a common Reference Model. The ENVRI Reference Model hosted at www.envri.eu/rm provides the ESFRI Environmental Research Infrastructures with a common ontological framework for the description and characterisation of ICT infrastructures, and provides them a community standard to help achieve greater levels of interoperability between their heterogeneous resources.

The ENVRI RM itself together with the associated documentation on how to use and comply with the model form another main result of the ENVRI project.

The ENVRI RM defines a conceptual model that captures computational requirements and state-of-the-art design experiences. In a sense, the model reveals a snapshot of the existing landscape of the ESFRI environmental science research infrastructures at a high level of abstraction and in language terms that are common to and can be recognised by every RI.

The ENVRI RM is built using the Open Distributed Processing (ODP) framework, an international standard for distributed system specification published by ISO/IEC (ISO/IEC 10746-1, 1998). ODP provides an overall conceptual framework for specifying large or

complex distributed computing systems. It adopts the object modelling approach, and defines five specific viewpoints – abstractions that yield specifications of the whole system related to particular sets of concerns. The five viewpoints are: *Enterprise Viewpoint*, *Information Viewpoint*, *Computational Viewpoint*, *Engineering Viewpoint*, and *Technology Viewpoint*. The ODP model, and hence the ENVRI RM, addresses the risk of technical divergence in large distributed ICT projects, a common cause of failure or extra cost– particularly so when such endeavours involve many different partner organisations. The ENVRI RM facilitates the recognition of commonalities that enables subsystem costs to be more widely amortised and improves interoperability.

By examining the environmental research infrastructures and their computational characteristics, five common *subsystems* have been identified: *Data Acquisition*, *Data Curation*, *Data Access*, *Data Processing* and *Community Support*. The primary reason for the division into five subsystems is that it reflects the observation that all applications, services and software tools are designed and implemented around five categories of physical resource: the sensor networks, the digital storage provision, the (Internet) communication networks, the computational servers and the client interaction devices.

3.3 Technical results

The focus of the technical development activities in ENVRI is in data discovery and data access across communities. The output of the project in this area is mainly technologies that enable building platforms, solutions and services to match the Research Infrastructures' needs and to advance integrated data-driven science. ENVRI has also produced prototypes and pilots of a few services that showcase the potential of the underlying technologies and tools. The main technical results of the project are briefly introduced in the following subsections.

3.3.1 Data Discovery Tools

Data Discovery capability produced in ENVRI is mainly focusing to enable OpenSearch based data discovery. OpenSearch is a collection of simple formats for sharing search results, which can easily be extended to retrieve metadata needed by specific community domains to search results, as a mass market technology that can be easily implemented by existing catalogue instances.

We hereby expose ENVRI existing examples of such technology that can be helpful for developing ESFRIs related scenarios to enable easy discovery of communities' data and related metadata.

ENVRI ESA OpenSearch Catalogue Access Service

During the ENVRI project, ESA has deployed a catalogue instance hosting the metadata related to several heterogeneous ESFRI product samples in order to demonstrate and validate the OpenSearch approach to address the requirements of Environmental RIs in terms of data discovery.

The catalogue instance deployed in ESA is OpenSearch based and its metadata are ingested and provided in rdf/xml format. A Central Site aggregator node using the same

technology is used to aggregate external OpenSearch based catalogues, and ingest external data sets.

The catalogue is able to work with different kinds of information and is extensible. For the sake of simplicity we can organize such information in two categories:

- General information, containing elements that define general properties of the Series and Datasets to be published and searched in the catalogue.
- Domain specific information, containing metadata related to the specific domain.

With respect to the general information the catalogue enables data providers to insert and publish a set of metadata elements that comply with the INSPIRE directive. Even if it is strongly suggested to be compliant with this set of elements, the Catalogue Access Service can work also with a subset of such elements.

In addition to generic metadata, the Catalogue Access Service enables the specification of domain specific metadata in order to enable publication of domain metadata. Besides including them at the series and/or dataset level, the Catalogue Access Service database shall be updated (by mean of a configuration file) to include such new elements.

Publishing metadata

In the context of the afore-mentioned demonstrating activity, in order to enable discovery of metadata by means of ESA provided services, the interested community shall follow these steps:

- In case of existing OpenSearch catalogues the community shall provide for each series some general information (such as the title, the identifier, the abstract, the temporal range and the geographic area and the URL of the series). This information will be published on the aggregator Node registries and will enable searches by the central catalogue.
- In case of need for publication metadata in xml/rdf format shall be provided to load the catalogue. This can be done easily with scripts and XSLT from the available data; these extraction procedures strongly depend on the original metadata/data format.

The Catalogue Access Service has been deployed in ESA to demonstrate that the OpenSearch approach is viable and well suited to address the multidisciplinary discovery needs of the heterogeneous ENVRI communities. It should be clear, however, that the current deployment is for demonstration purposes only and a more sustainable federated approach is needed in the future where each RI rather becomes responsible for the operation and maintenance of its data catalogues.

This means that:

- Each ESFRI should expose an OpenSearch Catalogue Access Service for its data;
- Clusters of ESFRIs, or clusters of sub communities within different ESFRIs, which share specific scientific goals can collaborate developing:

- Dedicated aggregator nodes used to aggregate or collect the ensemble of OpenSearch based catalogues of interest for the specific scientific goal;
- Specific web clients designed for their specific needs.

As an example:

> *ESA will continue offering discovery of its past and future satellite data via an OpenSearch Catalogue Access Service. ICOS Ocean Thematic Center and EuroArgo may decide to develop their own Catalogue Access Services and to set up a specific web client for discovering data related to Ocean physical parameters thus providing discovery capabilities of ICOS OTC, EuroArgo, and ESA data.*

3.3.2 Data Access Tools

ENVRI has developed a number of services for Spatial Data Infrastructures offering facilities to publish, discover, access, visualize, and process large geospatial datasets. These components contribute to the gCube open source technology, a comprehensive software system supporting the creation and management of a Hybrid Data Infrastructure (HDI).

ENVRI Contributions to gCube include the following elements that will be used to sustain services that were developed by the project:

- ***Geospatial Data Access service*** dedicated to provide users with facilities for having access to geospatial data, relying on THREDDS Data Server to realize a data service that gives access to geospatial data though protocols including OPeNDAP, OGC WMS and WCS. In this context, ENVRI has developed a software library named *GIS Interface* to simplify the interaction with a cluster of GeoNetwork and GeoServer instances to offer methods for storing and retrieving data and metadata from the instances of such technologies by abstracting over data distribution and replication details. To do that, it exploits the GeoNetwork Software Library, i.e. a Java library dedicated to interact with GeoNetwork.
- ***Geospatial Data Processing*** software library enabling *data mining* on geospatial data. This Software library is named Ecological Engine Geo Spatial Extension.
- ***Geospatial Data Publishing, Visualization and Explorer*** and ***GIS Viewer***, components dedicated to support the browsing and visualization of geospatial data.

3.3.3 Data Portal Pilot

One of the prototype services developed within the context of ENVRI project is a data access and visualisation portal hosted at CSC. The portal is currently connected to atmospheric data collections from measurement stations in Finland, and could be extended to host data from other infrastructures as well. Using the currently available applications on the platform, the user can e.g. download atmospheric data, study spreading of pollution clouds on a map, or use the OpenStreetMap through the interface service. The prototype portal URL is http://avaa.tdata.fi/web/smart/smear

Figure 1 shows the main window of the prototype data portal. The portal shows currently visualizations of real-time parameters such as air temperature, humidity, CO2 concentrations, and Ozone concentrations. Parameters and flows can be visualized on a map as well. The data sets can be searched by date, and the data can also be downloaded. The data in the portal is openly accessible for anyone.

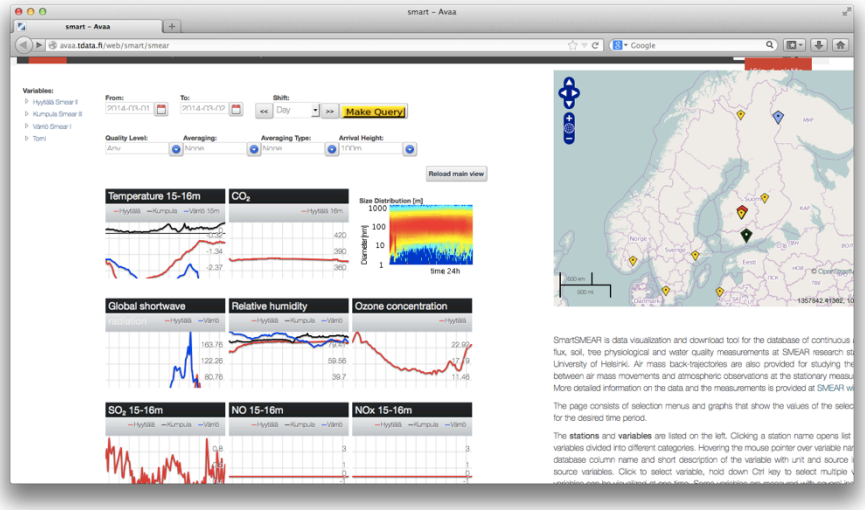

Figure 1. The main view of the prototype data portal at CSC.

The portal originates from a collaboration between CSC and UHEL. In the ENVRI project the portal has been further developed to match the needs of the Environmental Research Infrastructures with ICOS as the main contact. Connections to the services introduced in the earlier subsections 3.3.1 and 3.3.2 is under way, OpenSearch descriptions of sample datasets have been produced for the purpose of having data visible in the ENVRI data discovery demo portal.

The prototype data portal is built on the Liferay portal platform, which is a common open source portal platform. The same technology is used in the ENVRI data discovery service (subsection 3.3.1).

4 BENEFITS OF ENVRI RESULTS

4.1 Evaluation of benefits: ENVRI policy level collaboration

In addition to ENVRI reference model and technologies, a major benefit of the ENVRI work is the strategic level collaboration between the Research Infrastructures. This is a crucial matter for the purpose of Earth System sciences in general, as the Earth observations are generally fragmented and discipline-oriented. Lack of co-operation in long-term goals has seen to be a major drawback, hindering use of the RI products and their development. Even more importantly, the lack of organised high-level collaboration in Earth System sciences have lead to difficulties in answering key societal challenges. In fact, one of the crucial outcomes of the ENVRI project is the increased level of dialog and institutional co-operation between Research Infrastructures. This dialogue is further improved by the ENVRI RM, which provides and develops a language and framework for agreeing consistent approaches and target standards.

By forming the ENVRI SAB, the project has brought together European Research Infrastructures and other key participants to communicate and plan long-term solutions for better collaboration and integration. One of the key outcomes of this strategic level cooperation is the development of the Environmental Research Infrastructure Strategy (ERIS), which describes the vision of the whole Earth System sciences, aiming for the year 2030. This plan is based on the vision of far more integrated Earth System sciences, full understanding of the processes, flows and feedbacks within the Earth System, and furthering the capability of the Earth System sciences to answer societal and environmental challenges.

It is also true that the products developed by ENVRI are not only for the participating RI's, or even the current ESFRIs or I3 projects. The overall interoperability work brings a great benefit for the whole environmental sciences and GEOSS. This integrative aspect will only come more important in the future when new environmental RIs are developed. As an indication of the expansion in this field, EC Assessment report[x] based on the consultation on possible topics for future activities of research infrastructures, include 31 potential environmental RI's to be considered for future development in Europe.

The actors on the expanding field will clearly have benefit from deep co-operation. The collaboration is also a key towards interoperability between the multitudes of RIs. This interoperability will also create significant societal benefit, making use of environmental datasets in policy, business and citizen science far more accessible. Without a clear and deep collaboration between the RI's, the work towards these goals will also be far more challenging. If the first steps towards the interoperability developed in ENVRI are not nurtured and curated, the environmental cluster will not have the strategic critical mass to develop such Earth system science, leading to fractured field with much lost opportunity.

Together all these activities have several benefits which can be hard to quantify monetarily, but are of extremely high importance for the overall Environmental Research. In short, they enable among other things

- Efficient strategic collaboration between the research infrastructures, including long-term planning.
- Reduction of overlap and duplication of work.
- Provision of research infrastructures with a common voice of the needs and aims of the Environmental sector.
- Common approach to the user base, making Environmental sector more approachable for the stakeholders.
- Shared vision to all research infrastructures and stakeholders on the future needs roles of the Earth System observations.
- Coordinate the training and requirements of the technical and IT specialist.

The costs of these coordinating actions are minor. Organisation, dissemination, travel and practical involvement of the strategic collaboration board are in monetary terms only a small part of the ENVRI work, but they are shown to have major impact.

4.2 ENVRI Reference Model Cost-Benefit Analysis

A basic Cost Benefit Analysis (CBA) is a commonly used model when determining whether or not it is financially viable to implement a change or adopt a certain solution. In its t simplest form the value of the benefits of a certain action are added up and the costs associated with the action are subtracted. The effect of time is also built into the analysis by calculating a certain payback period. The CBA becomes more complex when trying to put a financial value on intangible costs and benefits, which can be highly subjective.

In the ENVRI mid-project review, CBA was suggested as an appropriate tool to clearly demonstrate the benefits of the ENVRI Reference Model. The ENVRI RM defines common characteristics of environmental science research infrastructures in order to provide a common language and understanding, to promote technology and solution sharing and improve interoperability. When determining the benefits of the Reference Model, actual translation into monetary terms is a delicate matter and may be best understood by trying to imagine what costs would be incurred if the determined benefits were **not** in place. Even so, the costs and benefits remain very difficult to quantify. What are the costs for doing unnecessary replication of technologies and solutions, trying to communicate without a common language or doing computational planning with less clearly defined roles and processes? One of the strongest assets of the ENVRI RM is its flexibility and wide range of usage, depending on the specific needs of the research infrastructures. This broad spectra of the ENVRI RM is however what makes Cost-Benefit Analysis more intractable, since the variety of combinations of benefits makes it extremely complex when determining the financial values.

As an alternative to a traditional CBA, the benefits and costs of the ENVRI RM is not put into monetary terms, but grouped from a time perspective. Time is a crucial aspect to take into account, since benefits and costs often occur at different points in time – perhaps years apart. The return on investment is long-term, whereas many of the costs are up front, so a

five-year time scale was taken as the basis for discussion, as this matches project time scales, even though collaboration and research infrastructures are anticipated to last thirty years or more in many cases. The benefits are grouped in two levels:

- Operational benefits: immediate (1-5 years)
- Development benefits: intermediate (5-10 years)

Many of the conclusions made in this preliminary analysis can bring insight also on evaluating the benefits of other results of the project. The costs are both those encountered during the ENVRI project and those needed to sustain sufficient continuation. The scope of benefits could potentially be extended beyond the environmental cluster, but that is not attempted in this *preliminary* CBA.

4.2.1 Determined benefits of the ENVRI Reference Model

The time aspect is also important when determining the target group for the benefits obtained. Whereas most of the immediate benefits are targeted at the RIs themselves, the intermediate benefits will also be evident to the end-users of the products and the cluster as a whole. To secure benefits for the end-users is also beneficial for the RIs, since the end-users are the core reasons for the existence of the RIs. Another dimension of the intermediate benefits are emerging RIs, where initial costs can be cut already at the planning stage by re-using an already established framework.

Immediate benefits

- The ENVRI RM provides predefined professional framework to clearly define roles and processes in RI operations. This makes it far easier to design RI in the construction phase, and helps to evaluate current research infrastructures for division of tasks and finding missing or duplicated actions within the RI work.
- Easier definition of requirements of IT components, enabling more modular approach for the RI IT solutions, and making possible to use external suppliers (e.g. international IT co-operation projects) for the component development.
- More precise definition of the planned distributed ICT e-Infrastructure, its architecture and planned partitioning into subsystems, making the allocation of tasks and their products more precise, thereby reducing the risks of wasted work, inadequate components or incompatible subsystems.
- A framework that facilitates the oversight and management of all phases of e-Infrastructure development, operation and refresh, so that control can be more effective and more likely to achieve success.

Intermediate benefits

- Development of a common language (taxonomy of terms, concepts and definitions) and a common understanding which also facilitates communication with external communities. This enables the research infrastructures to understand each other's

operations, makes the roles of individuals working in the different RIs clear, and facilitates far better usage of the RI products, due common language used in the metadata solutions.

- Avoiding duplication of work via identifying common missing components or components needing improvement. Common boundaries between different aspects of research infrastructure work can make some of the solutions, especially in the IT parts, interchangeable.
- Enabling re-use of components, solutions and policies. This enables easier planning for emerging research infrastructures, providing them with a “cookbook” of working solutions and facilitates direct connection to other infrastructures already from the beginning.
- The use of planned standard modular approach provides research infrastructures with scalable design solutions, where the parts of the RM needed can be further detailed for each RI needs, still keeping the overall benefits of common approach valid.
- Better risk management of research infrastructure development, due possibility of changing individual modules and operations of the infrastructures, without needing to completely redesign the systems due some ad-hoc solutions.
- Improving the trustworthiness of the research infrastructure products due clearly defined and standardised ways to present workflows.
- Greater level of interoperability between research infrastructures through the use of common standards, which makes data usage and communication between the RIs commonplace.
- Support of cross-disciplinary perspectives and products and enablement of systems science approach.
- Larger potential user base due easier use of the products, which increases the impact and return of investment of research infrastructures.
- Greater chance of using widely adopted solutions thereby amortising operational and maintenance costs with many other organisations and communities.
- A well-defined description of the operational e-Infrastructure so that the inevitable extensions and revisions can be discussed, planned and executed effectively.

4.2.2 Determined costs of the ENVRI Reference Model

Similarly to the benefits of the ENVRI Reference Model, the determined costs are heavily depending on the specific needs of the particular research infrastructure. There are however a few basic cost elements which are more generic. These costs are described and shown in project Person Months (PMs) below, based on experience gained in developing and using the model. The list clearly shows that compared to the benefits many of the larger cost posts are up front:

- Development of the ENVRI Reference Model to date: 120 PMs
- Further development of the ENVRI RM, for example to include Engineering and Technology viewpoints, and tools to support operational usage: 72 PMs
- Set up of Open-Source ENVRI RM sharing & contribution site: 12 PMs
- Costs for Existing research infrastructures who adopt the ENVRI RM: 6 PM as a one-time cost
- Marginal Costs for new research infrastructures who adopt the ENVRI RM (in comparison to use some model or mechanism to describe their actions): 0 PM
- Maintenance & extension of the model to meet new requirements: 12PMs/year
- Coordination, oversight and ICT support of Open-Source ENVRI RM sharing & contribution site and processes: 6PMs/year
- Outreach and training: 6PMs/year for just Environmental ESFRI cluster, more if user community grows
- Operational costs managing the above development and services: 4 PMs/year

The operational costs for sustaining the Reference Model amount to are approximated to be 28 PMs/year. This is substantially smaller than the investment in the development of the model, which is about 200 PMs. The cost for a Research Infrastructure to adopt the model is still smaller; according ENVRI project experience it is about 6 PMs. For new emerging RIs, the marginal cost is negligible as they infrastructure need to plan and model their operations in some method in any case.

4.2.3 Already materialised benefits for ESFRIs

ENVRI deliverable D3.5 presented some of the already materialised examples of the ENVRI RM's use by ESFRIs. It is presented below to give concrete examples of realistic short-term benefits of the ENVRI RM.

Using the Reference Model as a Research Tool to Guide Research Activities (Modelled system: EISCAT 3D)

It was shown that the Reference Model could be used to conduct various system analysis tasks. Using the Reference Model for EISCAT 3D ESFRI

- Clarified the boundary of the EISCAT 3D data infrastructure and identified the missing functionalities in the design;
- Provided a solution to integrate the EGI services into EISCAT 3D data infrastructure;
- Identified gaps between the EGI generic service infrastructure with the requirements from the domain specific research infrastructure, EISCAT 3D;
- Provided a knowledge base containing useful information could be referred in various system analysis and design activities;

- Provided a uniform platform into which computational elements of different infrastructures could be fitted, enabling comparison and analysis;
- Provided a way of thinking of constructions of plausible system architectures.

Using the Reference Model as an Analysis Tool (Modelled system EISCAT 3D and EUDAT services)

The principal potential benefit of using the Reference Model in general is the ability to precisely identify components required by an environmental RI and then identify how (if at all) the RI implements those components. In this example, the D3.5 used the EUDAT services as analysis targets. EUDAT provides a number of services that implement certain components (primarily in Data curation) it should therefore be possible to identify the equivalent services in a modelled RI and determine whether or not there is a benefit to delegating those services to EUDAT. This decision may be based on cost (particularly related to economies of scale) and development time (in cases where the RI has not yet implemented the service, but may be able to use the EUDAT service instead).

Using the Reference Model to Describe the Architectural Features of an Infrastructure (Modelled system: EMSO and EPOS research infrastructures together with Research Data Alliance (RDA))

The EMSO example demonstrates how to use the common language defined by the Reference Model in documentation to communicate with the RDA community.

It has been recognised there is a common challenge when communicating with external organisations or communities -- "*your 'model' is not my 'model', your 'data' is not my 'data'*". With a public accessible reference base, an external community that has little domain knowledge, such as RDA, is able to understand the specific descriptions of EMSO by looking up the terminology in the Reference Model. In a way, using the Reference Model, the communication efficiency can be improved. The intention is to show that in practice, to pursue the fitness, significance or beauty of the writing, an author may use different vocabularies to express a single concept. However, one can link them to the related concepts and definitions in the Reference Model to indicate their precise meanings. In this sense, using the Reference Model is different from using a dictionary – referring to the Reference Model places more emphasis on conceptual relationships.

Using the Reference Model as Design Reference (Modelled system: EPOS RI)

The ENVRI Reference Model captures common requirements of a collection of representative environmental research infrastructures, providing a projection of Europe-wide requirements they have, which potentially can be served as a technology roadmap to position and orchestrate collaborations in their design and development. It provides well-defined subsystems of components specified from different complementary viewpoints (Science, Information and Computation), which can help break down the complexity and simplify the design problems, enabling designers to deliver a practical architecture that leads to concrete implementations. It offers a descriptive framework for specifying uniform distributed systems, allowing designers from different organisations to carry out design activities in parallel.

Using the Reference Model as a Guide to Provide Common Services

To demonstrate the feasibility of the design specifications of the reference model, instances of selected model components have been developed to implement a data access subsystem that supports integrated data discovery and access. Data products from different environmental research infrastructures including, measurements of deep sea, upper space, volcano and seismology, open sea, atmosphere, and biodiversity, can now be pulled out through a single data access interface[xi]. Scientists are using this newly-available data resource to study environmental problems previously unachievable including, the study of the climate impact caused by the eruptions of the Eyjafjallajökull volcano in 2010, and the alien-species invasion phenomena around Sicily.

4.3 Evaluation of benefits: ENVRI technical results

The guiding principle in developing technologies in ENVRI has been to tackle issues that are common to the Environmental RI cluster. The data discovery, access and visualization tools that the project has produced enhance the possibility of multi-disciplinary environmental science. This supports in a concrete way the vision for Environmental RIs created by the ENVRI policy activities.

Being able to integrate research data from various Research Infrastructures from different domains of environmental science gives new kind of opportunities for research. Just as a simple example, combining marine measurement data with atmospheric data can bring novel understanding on the mass exchange between the atmosphere and the oceans. To reach the integration of data sets requires many steps to be taken in advance. This starts from discovery of the data. ENVRI outcomes can facilitate a federated system of data catalogues, providing centralized information on what types of data is available from certain locations or time intervals.

The advantages deriving from the adoption of well-known technologies in ENVRI – the OpenSearch and the gCube technologies – are numerous. They enable cross-disciplinary data discovery, both within an RI and between different RIs. They also facilitate a federative approach, which is more sustainable than a purely centralized system. As the technologies are based on open source codes, open interfaces and standards, they provide flexibility in customizing solutions.

The ENVRI project has demonstrated the benefits of applying selected technologies to integrate data from different domains by providing different prototype and pilot services and portals to the ENVRI community. The simple implementation and extension potential of the services allows their customization and adoption at the Environmental RI communities.

5 SUSTAINABILITY OPTIONS

5.1 Sustainability of policy level results

The sustainability of the policy advances from the ENVRI cluster activities are much dependent on the overall strategy chosen by the participating RIs. ENVRI is working to achieve a strong collaboration and overall policy of interoperability for the benefit of all Environmental Research Infrastructures. The collaborative cluster activity can be sustained only if most of the research infrastructures find the collaboration beneficial. According to experiences in the ENVRI project and within the SAB this seems currently to be the case. Based on this it is conceivable that the ENVRI cluster can be maintained as a collaborative cluster activity on the basis of RIs own initiative.

However, to really facilitate active collaboration would require that the collaboration forum has a role and clear function – as is the case with ENVRI SAB evaluating and advising the ENVRI project. Defining such a role with some level of funding will boost the sustainability of the collaboration significantly. The collaboration body will also need a function that could be called a secretariat, taking care of administrative functions that are currently handled within ENVRI project. A crucial part of the work is to sustain close co-operation between the RIs, especially between the project leaders (in the preparatory phase) followed by the RI directors (in their construction and operational phase), to make sure that their commitment to collaboration is maintained.

An additional concern on the sustainability of the policy activities is the integration with new RIs that are currently in development. The ENVRI cluster must directly approach these infrastructures, and keep direct and open discussion with the new players. The current ENVRI RIs will form only a minority of the potential new environmental RIs, and thus it is critical that not only the ENVRI RIs maintain cohesion and discussion, but also that new views are directly included in the processes. As the number of participating infrastructures grows, the requirements for the management and institutional development of the cluster will increase.

The main critical factors recognised within the sustainability of the policy advances are

- Maintaining the current discussion and collaboration level. This includes maintaining the contact information and regular meetings, including common portals for the data sharing and dissemination to stakeholder groups
- Dissemination of the policy developments to new RIs, stakeholder groups and funding agencies. This ensures that new RIs will participate in the development and collaboration, and to make sure that the policy advances will lead to more unified voice from the environmental RI cluster. If the policy development will be shown to produce results in this field (e.g. via improved visibility, usability and funding opportunities), the cluster will ensure the necessary critical mass for sustainable growth;

- Openness of the collaboration with respect to the emerging Environmental Research Infrastructures to achieve an inclusive and truly representative collaboration
- The sustainability of the SAB as a collaboration forum will benefit from more permanent institutional form. This is due two critical needs: The cluster needs to maintain cohesion and long-term respectability, and the results need to be consistently and strongly disseminated and communicated outside of the cluster. This could take a form of RIs expressed commitment to maintain a joint and sustainable collaboration board based on the ENVRI SAB model
- Funding for continuing ENV cluster collaboration started in ENVRI project for the critical intermediate period before more permanent structures can be adopted.

5.2 Sustainability of the Reference Model

The ENVRI Reference Model is published on the wikisite, www.envri.eu/rm, which is accessible to all ESFRI ENV RIs and general public. The access and adoption of the ERM are fully open. The edition right is currently restricted to the ENVRI RM authors and contributors from the ENVRI community.

The wikisite represents a work-in-progress development of the Reference Model. From time-to-time new versions of the RM will be released allowing downloading, e.g., in PDF and Word Doc formats, which are the snapshots of the continuing evolvement of the understanding. The latest version is 1.1, and a PDF copy of this version is deposited at ORCA online Research @ Cardiff[xii], an institutional repository. Besides bug fixation, new development of the Reference Model has been ceased since September 2013, in order to provide a stable version for evaluation and early adoption. A second version of the Reference Model, in responding to additional requirements from ESFRI ENV RIs, would be produced once future funding become available.

The existing ENVRI RM is a substantial body of work developed by IT specialists based on intensive and effective discussions with RI specialists to identify requirements, choose options and select appropriate vocabulary. The Open Process Development model on which the ENVRI RM is based was developed to help large distributed IT projects avoid pitfalls and to efficiently achieve their targets. This will apply well to most RIs as they develop their ICT services and data handling systems. It will apply even more to multiple RIs, and collaborating e-Infrastructure providers taken together. But to achieve the anticipated benefits, the ENVRI RM has to be a dynamic and active framework that is also completed with sufficient detail wherever the need arises.

This requires a transition in the mode of development of the ERM, which leads to a self-sustaining open community with proper support and governance for the future versions of the ERM and its continued use. This may be reached via the following three phases:

1. Completion of key details. The process already shown effective in ENVRI1, would continue by developing aspects of the ENVRI RM that the ESFRI communities urgently need to plan their developing ICT services and to coordinate their ICT experts and e-Infrastructure providers. This would include the two viewpoints: *Engineering* and *Technology*, in aspects of

the total system that are being planned or implemented. This phase would prepare for the next two phases, including training key ESFRI ICT staff and community building.

2. Transition to Expanded Contributing Community. The expertise of centralized ICT specialists cannot encompass all of the ESFRI technical issues efficiently. In the transition phase, leading ICT staff in the ESFRIs and e-Infrastructure providers would take on part of the ENVRI RM design and maintenance role. This has two advantages: (i) it avoids overheads in communication with generic IT specialists, and (ii) it gives the actual ESFRI ICT practitioners ownership of the model. During this phase, existing ENVRI RM experts would be available to help in RM design and specification steps and would also maintain oversight of quality and consistency issues. This phase would continue the community building and training processes. It would also initiate a forum where technical interactions between ICT development and operations can be coordinated across multiple RIs and service providers using the ENVRI RM.

3. Open Community Process. This is the long-term sustainable goal towards which the two preceding phases build. They must have achieved a critical mass in the ENVRI RM definitions and in active community members that the larger ESFRI community, suppliers and other similar R&D activities in the environmental context and beyond recognize as worth sustaining. The community now takes over governance and steers future development of the ENVRI RM. In particular it accommodates the many changes in the digital context that the RIs can exploit. It encourages wider use to increase the gains, conformity to common standards and the critical mass of "customers" that e-Infrastructure providers and commercial vendors recognize. This community now takes responsibility for its own quality control and trains its new members. It may choose to retain some ICT ENVRI RM specialists to help in this, at least initially.

5.3 Sustainability of technologies and tools

5.3.1 Sustainability of Data Discovery tools

The selected approach for data discovery is based on an approved OGC standard, namely OpenSearch, and is federative in nature. These two aspects contribute to its sustainability as detailed below.

Catalogue services have been installed in ESA ESRIN and offered to Research Infrastructures. Although it is of primary interest for ESA to maintain these services, this is to be considered more as an interim solution rather than as the operational one. RIs in ENVRI should develop similar interfaces, following the documented OGC standard approach, providing OpenSearch interfaces at their premises. Sustainability of each single catalogue will then depend on the data provider, i.e., the individual RI hosting the data. As an alternative, EGI is providing appliances for catalogue services and data storage. As the approach is federative, in case one of the RIs dismisses its catalogues, this won't affect the rest of the federation and it will still be possible to discover and access data from the other providers.

An additional central service is needed for the federated data discovery service to function. This can be called the Central catalogue. There needs to exist a central catalogue that acts

like a registry for the federated resources. For the time being this service is hosted at ESA. This catalogue is a simple registry designed in a way that it can be easily reproduced and additional copies may coexist. It is recommended that each RI project identifies clusters between themselves and deploys such registries at its premises, so improving sustainability.

As outlined before (subsection 3.3.1) an ENVRI portal has been implemented to let the user familiarize with the adopted approach and demonstrating its colocation capabilities. ESA cannot guarantee the maintenance of this portal, but this won't affect the sustainability of the federation: the data will be still discoverable using the underlying OpenSearch interfaces, for instance by the D4Science infrastructure, by the GEO Web Portal in GEOSS, or by community or cross-community clients that the RIs may decide to produce.

Sustainability of the OpenSearch technology is guaranteed by the mass market (Google) supporting it, and from the adoption from several bodies from the Geospatial community like OGC.

Several approaches for sustaining a catalogue service are possible. The main scenarios are to contract a service from a service provider or to develop own solutions based on the open source technology. As a specific case of the service provider option is the EGI OpenSearch Catalogue Access Service.

The EGI solution is based on OpenSource Geospatial Catalogue (OSGC) software that is an Open Source OpenSearch Catalogue developed in ENVRI. The software code has been released on Sourceforge (https://sourceforge.net/projects/osgcat/). It has many desirable features, including support for various data formats, integration to cloud storage services, and web interfaces. Input from the EISCAT infrastructure has been used in designing the service, but it can be customized for other RIs as well. The models to utilize the OSGC software are to access the EGI service as a cloud Platform-as-a-Service model based on a service contract, or to develop and implement own solution based on the open source code.

Finally, we can estimate the following costs for sustaining OpenSearch federation based services:

- 5-6 PMs for developing a catalogue from the scratch.
- 0.5-1PMs per data series to added in the catalogue depending on the complexity of the case, including metadata and automated scripts.
- 2-3 PMs for extending an existing catalogue to support OpenSearch.

Summarizing different options are possible for each Research Infrastructure:

- To develop a new OpenSearch catalogue from the scratch (5-6PMs).
- To download, customize and use the EGI OpenSource Geospatial Catalogue (cost for support to be defined by EGI + 1.5PM for learning to use it + customization efforts).
- To use the EGI OpenSource Geospatial Catalogue as a service (cost to be defined by EGI + 1PM for learning to use it).

- To extend an existing data catalogue to support OpenSearch (effort 2-3PMs).

In addition, the catalogue population requires an approximate effort of 0.5-1PMs per series depending on its complexity. This effort is slightly lower in the case of extending an existing catalogue.

5.3.2 Sustainability of Data Access tools

In the context of ENVRI, a number of components conceived to enable the realization of a scalable and reliable Spatial Data Infrastructure (SDI) have been developed. In particular, these components are oriented to provide users with facilities to publish, discover, access, visualize, and process large geospatial datasets. For the implementation of these components, state-of-the-art technologies and standards have been largely used. These components contribute to the gCube open source technology, gCube being a comprehensive software system supporting the creation and management of a Hybrid Data Infrastructure.

Environmental RIs can exploit the gCube facilities according to two major exploitation models: ***as a Service*** offered by the D4Science.org infrastructure, which reveals the advantage to have access to a maintained infrastructure. In this case the management of the software and infrastructure is outsourced to expert operators that can leverage economies of scale to keep the costs low and use elastic approaches to scale; and ***Open Source*** where any Research Infrastructure can take this technology and deploy it (or part of it) on their own premises, gCube being accessible for free.

In the context of ENVRI, the technology is also offered “as-a-Service” via the D4Science infrastructure, operated by the D4Science.org initiative. The services are currently offered without costs to ENVRI consortium members by relying on CNR resources and support. The exploitation of these services in production-quality scenarios requires the establishment of an agreement regulating the expected quality of service for availability, performance, storage space, data preservation, etc. This agreement and the related cost is negotiated on case by case basis with D4Science.org.

5.3.3 Sustainability of Data Portal Pilot

CSC plans to maintain the Prototype Data portal (c.f. Section 2.3.3) continuously after the current ENVRI project. The maintenance and development of the portal will be done mainly within a national project on improving open access to research data, ATT (acronym from Finnish for Support to Open Science). This project has funding until the end of 2014, but the platform together with the data portal is planned to be maintained also beyond the project funding However, no formal guarantee is made to continue the service or keep the data collections available, unless otherwise agreed contractually and with the understanding that such a contract may include cost for the customer community. However, truly European usage of the data portal or its underlying technologies will require commitment from the Environmental research infrastructures, for example as a further collaborative project to realize the potential of the demonstrated features in full scale.

The current instance of the portal is seen as a prototype or pilot service for ENVRI communities introducing the capabilities of the technologies and the connectivity to ENVRI data-discovery portal. This same platform can be used to setup similar data access and

download services connected to some other data collections. In this way the portal is a template for a data visualisation and download service to be used as a reference for those RIs that would like to setup such a service for their data and for their users. The portal software stack will be made openly available. The interfaces to data sources are open, so that the portal can be customised and connected to different data sources.

5.4 E-Infrastructure components

Many of the technological developments of ENVRI are such that they can benefit from European e-Infrastructure initiatives, as mentioned in the Chapter 2 of the document. The two e-Infrastructures that ENVRI has had most frequent interaction with are EGI and EUDAT, which is natural as the coordinating entities for both are presented in the ENVRI Consortium (EGI.eu and CSC, respectively). The following subsections present concrete examples and views how Research Infrastructures can benefit from services and capabilities provided by the e-Infrastructures. The role of an e-Infrastructure is usually to be a trusted partner to the RI – providing reliable service according to a mutual contract or as part of a collaborative project. The ideas can be extended for additional ICT service providers as well.

5.4.1 EGI Federated cloud service

ENVRI and EGI.eu shared effort on the integration of the ENVRI data catalogue and dissemination tools into the EGI Federated Cloud. This activity produced an Open Source Software-as-a-Service solution for data cataloguing and dissemination that has been integrated with the EGI cloud storage and is being tested by the EISCAT-3D community.

To identify services and solutions from state-of-the-art e-science infrastructure providers that can address the data pre-processing, post-processing and publishing needs of EISCAT_3D, together the EISCAT_3D project, EGI and EUDAT established a ‘study case’ partnership in February 2013 within ENVRI. The study case aimed at identifying and allocating solutions that directly benefit EISCAT_3D and that is expected to be reusable in other ESFRI projects of ENVRI as well.

Within this partnership, data, applications and other types of requirements from the EISCAT community have been collected for the proof of concept system. Identification and evaluation of technologies from EGI, EUDAT (and possibly from other e-infrastructures) that could satisfy the requirements and that can work together within a single architecture have been performed.

Solutions that are capable of migrating and registering large number of files and metadata on EGI resources have been the first area of focus. Other areas of interest are the portal framework, the applications repository and the data visualisation and mining tools.

A prototype system based on technologies and resources available for the architecture at that time has been setup.

Lessons learnt and recommendations for EISCAT towards the setup of the 'Off site' component of the EISCAT_3D system, and for ENVRI concerning the setup of ESFRI infrastructures in the environmental sciences have been drawn. EGI will continue to work in collaboration with EUDAT to provide joint production quality solution to EISCAT_3D.

The table below includes identified ENVRI use cases for deployment onto the EGI Federated Cloud:

Platform	Description	Key services	Use cases
Data dissemination/Open data	Cloud storage provides an infrastructure to collect and disseminate scientific data. Data intake is curated, data access is inherently open (anonymous) or tracked (social identity?)	•SaaS catalogues tailored to user communities (ex. EO, Biology, etc…) •Custom Data ACL •Federated AAI for data access •Monitoring of data usage	ENVRI GeoCatalogue EISCAT 3D
Virtual Laboratories	Tools to customise and manage virtual laboratories for different communities. Laboratories have shared tools to access data from different sources. User communities manage the instruments, EGI operates the underlying infrastructure and provides the generic tools to access storage (cloud storage, etc...)	•Common shared tools to access data from different resources •Tools to ease laboratories setup •Underlying IaaS and SaaS to support running of the laboratories and store user data storage (with easy scalability) •Simple user interface to request laboratories access	LifeWatch virtual laboratory
Data preservation	Data long-term preservation. To provide consolidation, persistency, integrity, redundancy, and usability over long periods of time.	•Data consolidation (ensure all the data is harmonised in terms of format, nomenclature, access, etc…) •Data integrity and redundancy (ensure no loss of data) •Data access preservation (ensuring software to read and analyse the data is maintained)	EISCAT 3D

EGI is working on a pilot pay-for-use access scheme and will officially launch the EGI Federated Cloud in production in May 2014, during the EGI Community Forum.

Other services EGI may offer to ENVRI include support for SAB in terms of user forum, collaboration tools (e.g. svn, liferay, wiki, applications database repository, etc.) and dissemination activity (mainly through EGI website and service catalogue[xiii]) including possible hosting for ENVRI meetings.

Concerning the ENVRI Reference Model, EGI may act as a technical consultancy for use of the EGI Federated Cloud and be a brokering actor for ENVRI. Support for the user communities can be organised utilising EGI's GGUS support system through EGI Support Unit or Distributed Competence Centre.

Last but not least, the involvement of Environmental Research Infrastructures in the EGI.eu governance is also an option to be further investigated by ENVRI, as soon as the new EGI Governance model will be available.

5.4.2 EUDAT services

The EUDAT project operates in a European landscape of developing or already existing data infrastructures. These Research Infrastructures already have developed solutions and tools for managing their data. The goal of EUDAT is not to replace these infrastructures but to support and enrich them by proving strong data infrastructure component and generic services on which they can rely to build up their data strategy. The EUDAT project established the European Collaborative Data Infrastructure (CDI) as a network of collaborating, cooperating centres, combining the richness of numerous community-specific data repositories with the permanence and persistence of some of Europe's largest scientific data centres. According to the CDI model, two categories of users can be established: "Internal users" are those concerned with the management of community-specific data repositories. Internal users can join their repositories formally with the CDI network, instantly benefitting from the persistence and resilience offered by the EUDAT partner network. Internal users are interested in archiving, replicating, processing and cataloguing data on behalf of the research community they support; "External users" are those wishing to share data with colleagues or collaborators, or those wishing to discover, and re-use data as part of their ongoing research. External users are anybody – researchers (from academia and industry), citizen scientists, policy makers, members of the public – anyone wanting to share or re-use European research data in simple, powerful ways.

EUDAT is currently funded by the European Commission to establish and support the organisational framework which allows developing and operating these services on a pan-European level.

EUDAT provides currently four services aimed at supporting research data management: B2Find for finding research data, B2Safe to safely replicate research data, B2Share for storing and sharing of research data, and B2Stage to transfer data to computation. Other services in the area of semantics and dynamic data are also being developed. Several

research communities within ENVRI have already expressed interest in using and deploying these services which could be of interest for the ENVRI community as a whole.

The EUDAT services aim to be sustainable beyond the project. in particular the storage services which are guaranteed by the EUDAT hosting sites through specific service level agreements. The sustainability of the EUDAT CDI is addressed from three main angles:

- organisational, through strong connections between data centres and community repositories formalised in service and operational level agreements. EUDAT leverages the stability of a geographically distributed network of established legal entities with reliable long-term organizational, national and/or European funding.
- technological, avoiding any technology lock-in to ensure flexibility and smooth evolution, and,
- financial, through the adoption of suitable business models for adopting, supporting and sustaining common services developed for, and partly operated by, the different research communities.

5.5 Summary

The different outputs of the ENVRI project have, as we have seen from the previous sections, different outlooks for sustainability. However, all of them depend ultimately on the interest of the Research Infrastructures to adopt, use, develop, support, and/or endorse these results.

Sustaining the policy level results, i.e. the cluster activity and added value of increased and continuous communication requires commitment from the management of the research infrastructures. The commitment needed is mainly in the recognition of the benefit of such an activity and the willingness to spend a limited amount of time attending relevant meetings rather than a commitment to make financial contributions. However, it is argued that the sustained cluster activity would be insecure in absence of continued funding at least for the critical intermediate period until a more permanent structure can be designed and agreed on.

The ENVRI Reference Model is key to maintaining interoperability between the research infrastructures on the system level. Additional benefits of the model were analysed in the Cost-Benefit analysis described earlier. Sustainability of the Reference Model can be achieved through adoption of the model at the infrastructures, which has already progressed within ENVRI. However, some group of people or an organisation needs to take responsibility on maintaining the model, updating releases and correcting bugs. It is conceivable that this could be done as a community effort, possibly supported by an interest group established at Research Data Alliance (RDA). The developers of the ENVRI Reference Model (cf. section 5.2) are willing to continue working on the model but without support from research infrastructures or EC this is not likely to be possible.

The technical tools and service prototypes of ENVRI are developed by ICT partners of the project, in collaboration with ESFRI projects. As the statements in sections 5.3.1-5.3.3 outline, the partners have desire to keep the outputs operational. However, in most cases the

continued operation cannot be guaranteed without a paying customer or a jointly funded development or support action. The technologies are offered for the Research Infrastructures to utilise in their own operations, with possibility for further development. This is supported by the general use of open source components in the services. The ICT partners of ENVRI are all committed to support the Environmental research cluster and so encourage Research Infrastructure communities to enter into discussions on further utilising the tools and technologies.

Many of the technologies of ENVRI can benefit from European e-Infrastructure initiatives. The role of an e-Infrastructure is usually to be a trusted partner to the RI – providing reliable service according to a mutual contract or as part of a collaborative project. The ideas can be extended for additional ICT service providers as well.

6 WAY FORWARD

6.1 Conclusions

The main aim of sustainability planning in ENVRI is to make sure that the project outputs are taken to the next level and put in a bigger context for further development and operational use within the Research Infrastructures. The success of the project is very much dependent on including IT partners who are not directly part of these RIs. On the other hand this requires even more effort in the sustainability planning stage.

The analysis made for this document shows that the current partners of ENVRI are committed to continue collaboration within the Environmental Research Infrastructure cluster. This includes also the statement that the partners involved in development of the ENVRI technical results – the ENVRI Reference Model as well as technology and tools – have interest in continuing the development activity. The ENVRI RM is dynamic by its nature and would additionally benefit greatly from user involvement in its further development. This naturally requires commitment from the customer community side supported preferably with public funding. Particularly, to provide operational services based on pilots presented at the ENVRI project will require a service contract with associated funding. It is important to underline that the ENVRI RM is not limited to the environmental field, but applicable to Research Infrastructures in general. Investments to the ENVRI RM development thus have the potential for an even wider impact on European research.

On the other hand, the sustainability of the project outcomes is strongly supported by the fact that these are based on open source technologies, open standards, and open interfaces (the technical tools); and distributed under open licensing (the ENVRI RM). The main developers of the technical outcomes plan to continue with the technologies, meaning that there is a strong competence center available ensuring continuity of the technologies. The maintenance of the ENVRI Reference Model may also benefit from community efforts, possibly also through RDA.

Three main methods to utilize and adopt Reference model, tools and technologies developed in ENVRI project can be identified, namely

1. *'As a service model'*: The RIs acquire the functionality they require from an external service provider
2. *Open source model*: The RIs take the technology and deploy it within their operations
3. *Collaborative model*: The RIs and ICT providers further develop and setup services in a collaborative effort

The three ENVRI RM models are summarized in the table on the next page, describing the main features, pros, cons and the optimal suitability

MODELS TO USE ENVRI OUTPUTS	DESCRIPTION	PROS	CONS	OPTIMAL SUITABILITY
AS A SERVICE	Client/provider relation between the RI and the service provider (ICT center, e-infrastructure, SME, etc.). Service provided according to a defined service level and description for a defined cost to the RI.	Well-defined service. RI does not need implementation or operation personnel for the service in question. Can be cost-effective if provider can utilize economics of scale.	Changes to the service may require additional fee. Flexibility to change configurations or implement quick changes may be partially lost. Interoperability may suffer if provider is not following common standards and interfaces.	Well-defined operational services.
OPEN SOURCE	RIs adopt the open source technology and implement solutions that suit their own needs.	Flexibility in implementing changes and total control of configuration of the service or tailoring of software. Possible to test new ideas quickly.	Requires investment to expert personnel in software development, system operation and administration. Requires appropriate processes to manage software and service development. Interoperability may suffer if many RIs develop own solutions.	Preparatory phase services where the actual requirements are partially unknown and changes are often made.
COLLABORATIVE	RIs and ICT providers (including e-infrastructures) develop models and tools as a collaborative effort within a funded project.	Collaboration between provider and user bring better understanding of requirements and alternatives. Development investment and effort can be shared with a number of partners. Facilitates inter-RI interoperability.	Development in collaboration tends to focus on common demands meaning that certain areas important to individual RIs may be left with smaller attention.	Services, which are common to more than one RI. Integrating services and services enhancing interoperability.

European e-infrastructures such as EGI and EUDAT provide many services that are suitable for the current and emerging Environmental Research Infrastructures. Involving these in providing platforms and services for RIs will in many cases be very beneficial, especially what comes to data discovery and access across research infrastructures. The e-infrastructures are in a position to provide platforms across user communities and operate them in a cost-effective manner.

In summary, The RIs are now in key position to decide how to benefit from the project outcomes. The three main recognised models to adopt the outcomes were listed in the table. It should be mentioned that the three models are not exclusive. Instead, the approach at each RI is probably a combination of two or three models. Particularly, the collaborative model can be said to be crucial for the future success of the Environmental Research Infrastructure cluster, because in absence of collaboration the activities towards interoperability and integration of data will be more difficult which will slow down the progress of Environmental science. Example of such integrating development is the ENVRI Reference Model in particular, but also tools for data discovery and access.

The continued Environmental cluster activity is also extremely important to sustain the fruitful policy level coordination that has been implemented through the ENVRI Stakeholder Advisory Board. One of the key outcomes of this strategic level cooperation is the development of the Environmental Research Infrastructure Strategy (ERIS), which describes the vision of the whole Earth System sciences. This can be in fact a guiding document for the continued strategic collaboration board.

6.2 Recommendations

Based on the conclusions and the conducted analysis a number of recommendations can be made towards facilitating the sustainability of the project results:

1. Collaboration of ESFRIs on Environmental cluster should be extended in long-term beyond the current ENVRI project in order to continue policy level discussion between the partners and to form a joint interface towards major European and international initiatives.

2. The ENVRI project is an excellent basis for the cooperation of environmental research infrastructures in support of advanced interdisciplinary research addressing the grand challenges of our changing environment. Next steps are required to benefit from the added value as developed in ENVRI.

3. To facilitate the creation of the above, and to allow further development of technical results and the ENVRI reference model, as well as their transfer to production use, an EC funded follow-up project to ENVRI from 2014 onwards is needed.

4. The ENVRI Reference Model is seen as highly beneficial tool for the RIs to plan, manage and develop their operations. The maintenance and further development of the ENVRI RM should be secured by collaborative efforts preferably under public funding.

5. The ICT developers are willing to continue the technical development for Environmental cluster, but the adoption of ENVRI outcomes at the Research Infrastructures requires action from the RIs themselves. To support the adoption and sustainability of ENVRI technical results, three alternative models are presented (as a service, open source, and collaborative). Based on this study, these models are feasible both for RIs and for ICT partners, and thus can form the basis for discussion on the form of future collaboration.

6. The capabilities for inter-disciplinary platforms and cost-effective solutions from European e-infrastructures should be used and exploited when possible.

7 REFERENCES

[i] http://ec.europa.eu/research/infrastructures/pdf/jd-final-aegreport-23sept13.pdf

[ii] http://tinyurl.com/oxs4z95

[iii] http://www.earthobservations.org/geoss.shtml

[iv] http://cf2014.egi.eu

[v] http://eudat.eu/

[vi] http://eudat.eu/dynamic-data

[vii] http://www.helix-nebula.eu

[viii] https://europe.rd-alliance.org/

[ix] https://rd-alliance.org/

[x] "Consultation on possible topics for future activities for integrating and opening existing national research infrastructures", EC Assessment Report, February 2013, http://ec.europa.eu/research/infrastructures/pdf/final-report-CEI-2013.pdf

[xi] ENVRI Data Portal: http://portal.envri.eu

[xii] http://orca.cf.ac.uk/50846/

[xiii] http://www.egi.eu/services/catalogue/