ENVRI Common Operations of Environmental Research Infrastructures

ENVRI

Services for the Environmental Community

Assessment of the State of the Art

Document identifier:	D3.1: Assessment of the State of the Art
Date:	29/06/2012
Activity:	WP3
Lead Partner:	UEDIN
Document Status:	FINAL
Dissemination Level:	PUBLIC
Document Link:	http://envri.eu/group/envri/documents/10995/41057/Deliverable3-1.pdf

ABSTRACT

The objectives of ENVRI can only be achieved with an in-depth comprehension and appreciation of the status and requirements of the ESFRI environmental research infrastructures currently in various states of readiness. This document surveys six of these infrastructures (EISCAT_3D, EMSO, EPOS, Euro-Argo, ICOS and LifeWatch) as well as three pivotal ICT infrastructures (EGI, EUDAT and D4Science) which may be able to provide services to them. In doing so, we hope to appraise the state-of-the-art for deployed data, tools and services within European research infrastructures today, and so begin to formally identify some of the common challenges which all such infrastructures must face, as well as some common solutions to those challenges.

SEVENTH FRAMEWORK PROGRAMME

1. COPYRIGHT NOTICE

2. DELIVERY SLIP

	Name	Partner/Activity	Date
From	Paul Martin	UEDIN/WP3	28/6/12
Reviewed by	Alex Hardisty, Mikka Dal Maso, Leonardo Candela	CU/WP3, UHEL/WP2, CNR/WP4	28/6/12
Approved by	Wouter Los	UvA/WP1	29/06/12

3. DOCUMENT LOG

Issue	Date	Comment	Author/Partner
1.0	Nov. 2011 to May 2012	Acquisition of RI information in ENVRI wiki.	UEDIN, CEA, CSC, CU, EAA, ESA, INGV
2.0	15/5/12	Compilation of collected data begins.	Paul Martin (UEDIN)
3.0	31/5/12	Release of internal review draft.	Paul Martin
4.0	28/6/12	Release of final draft.	Paul Martin

4. APPLICATION AREA

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

5. DOCUMENT AMENDMENT PROCEDURE

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

SEVENTH FRAMEWORK PROGRAMME

6. TERMINOLOGY

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

7. PROJECT SUMMARY

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

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

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

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

8. EXECUTIVE SUMMARY

The purpose of the ENVRI project is to find common ICT solutions to the problems faced by environmental research infrastructures in Europe, partly by modelling their common operations in order to provide insight into the construction of current and future infrastructures. This deliverable represents one part of the preparation towards achieving that purpose.

This report describes the state-of-the-art (as identified so far) of six ESFRI environmental research infrastructures (RIs) chosen as representative of the broad range of ESFRI infrastructures active currently: *EISCAT_3D* (geo-space monitoring), *EMSO* (seafloor monitoring), *EPOS* (solid Earth science), *Euro-Argo* (ocean monitoring), *ICOS* (carbon observation) and *LifeWatch* (biodiversity). For each RI, a description of its mission and status is given, followed by an overview of current activities, data products and technical

SEVENTH FRAMEWORK PROGRAMME

architecture. This report can therefore be seen as a snapshot of state of thinking about those infrastructures as well as of the on-going dialogue between ENVRI and the participating RIs.

This report also describes the state-of-the-art of three more infrastructures providing ICT services which are potentially of use to many of the ESFRI infrastructures: *EGI* (focused primarily on Grid computation), *EUDAT* (focused primarily on data management) and D4Science (focused on realising Virtual Research Environments).

From these surveys, a initial set of observations can be drawn to inform future efforts within ENVRI regarding both the commonalities and contrasts between the surveyed infrastructures. In addition, an initial discussion can begin into the formulation of the ENVRI Reference Model, a major part of the ENVRI contribution which can begin now in earnest.

SEVENTH FRAMEWORK PROGRAMME

TABLE OF CONTENTS

1 Introduction 8
1.1 Background 8
1.2 Content 9

2 EISCAT_3D 10
2.1 Activity 10
2.1.1 Incoherent scatter 10
2.1.2 Beamforming 11
2.1.3 Interferometry 11
2.1.4 Modularity 11
2.2 Data 11
2.2.1 Data products and metadata 12
2.2.2 Data preservation 12
2.3 Infrastructure 13
2.3.1 Multistatic configuration 13
2.3.2 Modular design 13

3 EMSO 15
3.1 Activity 15
3.1.1 IT activities 15
3.1.2 Users and stakeholders 16
3.2 Data 17
3.2.1 Data products and metadata 17
3.2.2 Data acquisition and archival 18
3.2.3 Data delivery 19
3.2.4 Data availability 20
3.3 Infrastructure 21

4 EPOS 22
4.1 Activity 22
4.2 Data 23
4.2.1 Data products and metadata 23
4.2.2 Data acquisition and archival 24
4.3 Infrastructure 24
4.3.1 ORFEUS Data Center 25
4.3.2 European Integrated Data Archive 25
4.3.3 Computing resources 25

5 Euro-Argo 27
5.1 Activity 27
5.1.1 Float operation 28
5.1.2 Satellite transmission 28
5.1.3 Use-cases 28
5.1.4 Contributions 29
5.2 Data 29
5.2.1 Data products and metadata 29
5.2.2 Data acquisition and archival 30
5.2.3 Quality control 30
5.2.4 Monitoring 31

SEVENTH FRAMEWORK PROGRAMME

5.3 Infrastructure 31

6 ICOS 33
6.1 Activity 33
6.2 Data 34
6.2.1 Data products and metadata 34
6.2.2 Data acquisition and archival 36
6.2.3 Data processing 37
6.3 Infrastructure 38
6.3.1 Computing resources 38
6.3.2 E-science capabilities 39
6.3.3 Common user services 39
6.3.4 Operational services 40

7 LifeWatch 41
7.1 Activity 41
7.2 Data 42
7.2.1 Data products and metadata 42
7.2.2 Data preservation 43
7.2.3 Data processing 43
7.2.4 Data publication 43
7.3 Infrastructure 44
7.3.1 Interaction 44
7.3.2 Data transport 45

8 Observations 46
8.1 Services 46
8.2 Architecture 47
8.3 Data 47
8.4 Computation 48
8.5 Working practice 49
8.6 ENVRI's contribution 50

9 EGI 51
9.1 Activity 51
9.1.1 Key IT activities 51
9.2 Data 52
9.2.1 Data acquisition 52
9.2.2 Data storage 52
9.2.3 Metadata 53
9.2.4 Data transfer 53
9.3 Infrastructure 54
9.3.1 Architecture 54
9.3.2 Security 54
9.3.3 Connectivity 54
9.4 Computation 55
9.5 Working practices 55
9.5.1 Security 56
9.5.2 Legality 56
9.5.3 Service 56
9.5.4 Ethicality 56

SEVENTH FRAMEWORK PROGRAMME

9.5.5 Authentication *56*
9.5.6 Authorization *57*
9.5.7 Accounting *57*

10 EUDAT **58**
10.1 Activity 58
10.1.1 Community *59*
10.2 Data 59
10.2.1 Metadata *59*
10.3 Infrastructure 60
10.3.1 Safe data replication *60*
10.3.2 Dynamic data replication *61*

11 D4Science **62**
11.1 Activity 62
11.2 Data 63
11.3 Infrastructure 63

12 Discussion **64**
12.1 ENVRI Reference Model 64
12.1.1 Enterprise viewpoint *65*
12.1.2 Information viewpoint *65*
12.1.3 Computational viewpoint *66*
12.1.4 Engineering viewpoint *66*
12.1.5 Technology viewpoint *66*

1 INTRODUCTION

This deliverable is a report on the state-of-the-art of existing and deployed data, tools and services within six ESFRI research infrastructures, along with an overview of two ICT research infrastructures (EGI, EUDAT and D4Science) which may be able to provide services to them or to future environmental infrastructures. The purpose of this deliverable is to provide information useful for identifying the common problems shared by RIs (Task 3.2) as well as for identifying focal points for further work on the ENVRI Reference Model (Task 3.3).

1.1 Background

The European Strategy Forum on Research Infrastructures (ESFRI) is an instrument for promoting the integration of European scientific research and strengthening its outreach capabilities. Each ESFRI infrastructure serves (or intends to serve) a broad community of researchers working in a particular field of research by providing large-scale data archival, integration or computational services which has not been previously available to that community.

Each research infrastructure has its own particular set of problems which it must solve to achieve its objectives; however every research infrastructure must also address certain common problems regardless of its particular field of interest. These common problems address issues of data collection, preservation, quality control, integration and availability, as well as providing the capability to move data to the kinds of computational context required to perform the analyses of interest to researchers (or *vice versa*). Moreover, whilst each RI is separately concerned with the integration of data within its domain of interest, it is also imperative to find robust yet lightweight means to integrate data and computation *across* RIs to serve an increasingly multidisciplinary scientific community.

In this report we survey six ESFRI research infrastructures (RIs). These are:

- **EISCAT_3D**, a project led by EISCAT **(European Incoherent Scatter)** which seeks to construct a three-dimensional imaging radar to make continuous measurements of the geo-space environment and its coupling to the Earth's atmosphere.
- **EMSO (European Multidisciplinary Seafloor Observatory)**, a European network of fixed-point, deep-seafloor and water column observatories.
- **EPOS (European Plate Observing System)**, which seeks to integrate existing European facilities for solid Earth science into one distributed and coherent multidisciplinary infrastructure.
- **Euro-Argo**, the European contingent of the **Argo** project, a global ocean observing system comprised of a large network of robotic floats distributed across the world's oceans.
- **ICOS (Integrated Carbon Observation System)**, dedicated to the monitoring of greenhouse gases (GHG) through its atmospheric, ecosystem and ocean networks.
- **LifeWatch**, a research infrastructure for studying biodiversity and the Earth's ecosystems.

The ESFRI RIs should not develop their data and computing infrastructures in isolation. Rather, they should rely on other European projects and initiatives which develop general services for data management or provide large-scale platforms for computation and storage. In this report we also survey three such research infrastructures:

- **EGI (European Grid Infrastructure)**, a sustainable combined infrastructure for European Grid computing resources.
- **EUDAT (European Data Infrastructure)**, which seeks to produce a Collaborative Data Infrastructure for European scientific data products.
- **D4Science**, which realises a hybrid data infrastructure supporting the creation of Virtual Research Infrastructures.

These projects provide services (or intend to provide services) of interest to environmental RIs including the normalisation and interoperation of data products and services, and the provision of a distributed, secure computational platform. These projects also have the potential to be integral in the integration of services across all the RIs.

1.2 Content

In this report we dedicate a section to each RI surveyed, providing an overview of its purpose, core activities, principal data products and architecture. Where known, information is also given on data handling, access and control policies, typical computations, and specific technologies deployed. For many of the RIs however, no decision has yet been made about many technical aspects of the RI; in this case most attention has been given to the services which the RI intends to provide. Naturally, ENVRI will continue to liaise with those RIs as they evolve over the lifespan of the ENVRI project.

After the six environmental RIs, a section is dedicated both to making common observations and drawing contrasts between the surveyed RIs in the spheres of data, computation, architecture and services. An initial analysis of requirements (in anticipation of later deliverables) is made, setting the scene for the surveys of the three generic infrastructures.

Finally the report is concluded with an initial discussion on how the state-of-the-art of the environmental RIs can be translated into a single ENVRI reference model. Since the intent is to use the RM-ODP (Open Distributed Processing)[1] approach to modelling, this entails a *preliminary* discussion of how different RI requirements decompose into the five ODP viewpoints of enterprise, information, computation, engineering and technology.

[1] http://en.wikipedia.org/wiki/RM-ODP

2 EISCAT_3D

EISCAT_3D (http://www.eiscat3d.se) is an infrastructure project led by the EISCAT (European Incoherent Scatter) Scientific Association, an association involving Norway, Sweden, Finland, the UK, Japan and China. EISCAT operates three incoherent scatter radars in Tromsø and on the Svalbard archipelago in Norway. The UHF radar is the only tristatic incoherent scatter radar in the world with additional receiver sites in Kiruna (Sweden) and Sodankylä (Finland). Tristatic operation interferes with mobile communications however, and so operation will come to an end soon, motivating the construction of a new facility.

EISCAT_3D is a next generation incoherent scatter radar, providing three-dimensional monitoring of the atmosphere and ionosphere above the northern Fenno-Scandinavian region (measuring such things as solar variability and coupling of atmospheric layers). The EISCAT_3D radar system will update existing EISCAT facilities with state-of-the-art technologies, outstripping the current capabilities of any equivalent radar system in the world.

ESFRI selected EISCAT_3D for the Roadmap 2008 for Large-Scale European Research Infrastructures for the next 20-30 years. The facility will be built as a modular system with construction starting by 2015.

2.1 Activity

EISCAT_3D will be a volumetric radar capable of imaging an extended spatial area with simultaneous full-vector drift velocities, having continuous operation modes, short baseline interferometry capability for imaging sub-beamwidth scales, real-time data access for applications and extensive data archiving facilities.

EISCAT_3D has three-dimensional volumetric imaging capability throughout its field of view. This allows the study of the variability, coupling and energy dissipation between the solar wind, magnetosphere and atmosphere (being a function of the direction relative to the Earth's magnetic field) – no current radar system has the same level of flexibility and autonomy as planned for EISCAT_3D.

2.1.1 Incoherent scatter

Incoherent scatter is a sophisticated radio method for remotely sensing conditions in the atmosphere and in near-Earth space. ISR (Incoherent Scatter Radar) is a high power, large aperture radar system which exploits a combination of incoherent scatter theory, measurement theory and standard inversion mathematics.

The basic scanning mechanism is to transmit a high power (~2 Megawatt) radio pulse into a target atmospheric volume; electrons within the volume scatter the radio wave, and this scattering is detected on the surface by an extremely sensitive (~10^{-18} Watt) receiver array.

EISCAT_3D will be divided amongst a number of sites, each possessing a cluster of antennae which collectively act as programmable receiver arrays suitable for the detection of

SEVENTH FRAMEWORK PROGRAMME

incoherent scatter signals. Adaptive (phased array) beamforming allows directional signal reception for multiple simultaneous beams according to much the same principles as used in radio astronomy.

2.1.2 Beamforming

Beamforming technology allows beam direction to be switched in milliseconds without the need for mechanical repositioning as for traditional dish-based radars. Very wide spatial coverage can be obtained by constructive interleaving of multiple beam directions for nigh-simultaneous volumetric imaging. Interferometry can be performed with multiple baseline angles and lengths, a technique demonstrated by the EISCAT Svalbard Radar; this will assist in the measurement of micro-physical processes such as NEIALs, black auroras, meteor head echoes and PMSE. At passive sites, the design allows for up to five simultaneous beams at full bandwidth, or twenty if bandwidth is limited to the ion line. This allows the whole range of a transmitted beam to be imaged using holographic radar techniques, which also allows satellites and space debris to be tracked across the sky – something that cannot currently be done by existing EISCAT radar systems.

The capability to receive scatter signals from multiple directions can also be used to resolve outstanding issues of spatial-temporal ambiguity (*e.g.* understanding the dynamics of dusty plasmas in the mesopause region and tracking space debris and meteors).

2.1.3 Interferometry

Active arrays can be split into smaller elements to be used for aperture synthesis imaging, resulting in a data product consisting of range-dependent images of small sub-beamwidth scale structures (with sizes down to 20 metres). All EISCAT_3D data will be tested continuously for the presence of such structures so that specialised processing can be enlisted to determine their shape and location when they emerge.

2.1.4 Modularity

Phased array radars are inherently modular; the system can, in principle, be scaled to whatever size is needed, either by increasing the size of individual sites or by enlisting multiple sites in order to extend system coverage. Geophysical conditions will also be monitored so as to allow experimental setups to be interchanged autonomously in response to changes in geophysical conditions.

Ultimately, it is a goal of EISCAT to set up multiple active sites capable of imaging the whole region of the ionosphere above northern Scandinavia and surrounding territories. Thus the design of EISCAT_3D should be inherently extensible.

2.2 Data

The EISCAT_3D radar system will consist of multiple phased arrays, using the latest digital signal processing to achieve ten times higher temporal and spatial resolution than the present radars. Parameters measured by EISCAT_3D include election density, electron temperature, ion temperature and plasma velocity. The fitting of models to the raw data permits the derivation of additional parameters of the target volume.

The data extracted from the operation of EISCAT_3D could prove vital in validating various forecasting models for deriving the state of the Earth or its geo-space environment.

2.2.1 Data products and metadata

The following low level data can be obtained by the EISCAT_3D system:

- **Voltage data (lowest level)**: 80 MHz sampling at 16 bits. This amounts to 2.56 Gigabits per second per element; elements are combined by group to measure up to 10 beams which results in a daily data intake of ~25 Terabytes.
- **Beamformed data**: produces approximately 1 Terabyte per hour per site.

Adaptive beamforming is highly computationally intensive. Beamforming is to be performed at two levels; the first level is done locally (within groups of adjacent antennae; each level 1 unit is connected to 6 neighbours by 10Gb/s Ethernet lines), whilst level 2 beamforming can be calculated within a computational cluster:

- 3 independent streams of full speed data can be routed for beamforming sums, totalling (theoretically) 23 simultaneous beams (10 Gb/s line capacity, 80 MHz sampling at 16 bit, 3 independent streams).
- With band-limited data, many more beams possible (*e.g.* for 5 MHz, 184 simultaneous beams).

Significant compute power is necessary to parameterise parallel beamforming operations and then distinguish the resultant beam readings.

Potentially, the following higher-level data can be obtained:

- **Interferometry data**: 19 modules in use (202 MB/s), but only 5% of samples above threshold are kept. Lead-in and follow-on data (in the order of tens of GBs) is also obtained.
- **Supporting instruments** (including optical instruments, other radars and diagnostic data) are estimated to produce 150 GB/day at the central site and 30 GB/day for each remote station.
- **Highest-level data products**: whilst much derived data is of insignificant size (next to the raw data), the results of correlation functions can accumulate in the order of ~200 TB/year.

2.2.2 Data preservation

Three types of data preservation have been identified as being required by EISCAT_3D:

- **Ring buffer**: of high volume (~100 TB), but short duration (hours/days). Data accumulates constantly, with oldest data continuously overwritten. Records interferometry when events detected and provides latent archive data in the event of a network outage.
- **Interferometry system**: a small area (~100 GB) storing a few minutes of data. Data accumulates constantly and is threshold tested; upon identification of an event, data flow is diverted for specialised processing – otherwise data is deleted.
- **Permanent archival**: initially must be of large capacity (~1 PB). Permanent storage will be only for mid- and high-level data, which is expected to accumulate at ~200 TB/year. Tiered storage connected to multi-user computing facilities is required. Over

SEVENTH FRAMEWORK PROGRAMME

the longer term, this capacity must be extended by at least 200 TB/year in order to maintain the desired data availability.

Low level data should be preserved as long as is feasible, retaining all metadata and provenance information available so as to permit multiple analyses to be processed on the same data volume in order to extract the most scientific value.

Data is available via the EISCAT Madrigal database[2] in the form of sampled time series. The Madrigal database is a federated system which distributes metadata between the Open Madrigal site and a number of individual servers holding specific data products. The Madrigal system is accessible via a number of Virtual Organisations. The construction of an additional archival site has also been proposed.

2.3 Infrastructure

EISCAT_3D is designed for continuous operation, with passive sites operating unattended; automated systems will be used to monitor and control radar operations and manage data input. This continuous operation, limited only by power consumption and data storage space, permits EISCAT_3D to respond to sudden and unexpected events which might occur above it.

2.3.1 Multistatic configuration

EISCAT_3D is able to take advantage of advances in networking, high performance computing and data storage in the Nordic region to handle greater data throughput as well as perform beamforming calculations and schedule multiple parallel experiments. EISCAT_3D will be the first phased array incoherent scatter radar to use a multistatic configuration. Envisaged are five distinct radar sites, consisting of two pairs located around 120 km and 250 km from the active site respectively, on baselines running East and South from the active core. This provides an optimal geometry for calculation of vector velocities in the middle and upper atmosphere.

The gain of the EISCAT_3D antennae and the large size of the active site arrays will deliver an enormous increase in both the sensitivity and confidence level of measurements made. An active site of 5,000 elements would already exceed the performance of the current EISCAT VHF system, while an active site comprising 16,000 elements will exceed the sensitivity of the present VHF radar by an order of magnitude.

2.3.2 Modular design

The design of the antenna arrays will be modular at different scales allowing for mass-production of the components. Some arrays will be very large, in the scale of 32,000 individual antenna elements. The receiver arrays will be located at 50-150 km distance from

[2] http://www.eiscat.se/madrigal/

the illuminators, and some smaller arrays closer by to support continuous interferometric observations. The total system will comprise ~100,000 elements. The actual radar sites have to be carefully chosen.

Each transmitter unit will have its own signal generator, allowing the generation and transmission of arbitrary waveforms, limited only by the available transmission bandwidth and spectrum mask as allocated by the respective frequency management authorities. This allows the implementation of all currently used and envisaged modulation schemes and antenna codings (such as polyphase alternating codes, array tapering, orbital angular momentum beams) and also provides the possibility to adopt any kind of future code. In addition, it will allow advanced clutter mitigation strategies such as adaptive null steering and null shaping.

3 EMSO

The processes that occur in the oceans have a direct impact on human societies, and it is therefore crucial to improve our understanding of how they operate and interact. The European Multidisciplinary Seafloor Observatory (EMSO; http://www.emso-eu.org) is a European network of fixed-point, deep-seafloor and water column observatories. Its main objective is to obtain real-time information relevant to scientific research and environmental sustainable management. This is achieved through long-term monitoring of environmental processes related to the interaction between the geosphere, biosphere and hydrosphere.

EMSO will encompass the breadth of these major processes by means of sustained and integrated observations and an appreciation of the interconnectedness of atmospheric, surface ocean, biological pump, deep sea, and solid-Earth dynamics. EMSO will address:

- natural and anthropogenic change;
- interactions between ecosystem services, biodiversity, biogeochemistry, physics, and climate;
- impacts of exploration and extraction of energy, minerals, and living resources;
- geo-hazard early warning capability for earthquakes, tsunamis, gas-hydrate release, and slope instability and failure;
- connecting scientific outcomes to stakeholders and policy makers.

Long-term, continuous data sets from a variety of fields are necessary to build a comprehensive picture of the earth-ocean system. These fields include geosciences, physical oceanography, biogeochemistry and marine ecology.

EMSO has been in its preparatory phase since 2008 and is now entering the implementation phase (2012-2016).

3.1 Activity

EMSO is geographically distributed in key sites of European waters, spanning from the Arctic, through the Atlantic and Mediterranean Sea to the Black Sea. It presently consists of twelve sites which have been identified by the scientific community according to their importance with respect to marine ecosystems, climate changes and marine geo-hazards.

3.1.1 IT activities

Most IT activities have been performed in EMSO's sister project ESONET[3]. A major focus is on standardisation and interoperability within the distributed EMSO architecture. Both observatory data and data archiving services are already provided by several observatory nodes and data centres; therefore the main challenge for the ESONET/EMSO data

[3] http://www.esonet-emso.org/

SEVENTH FRAMEWORK PROGRAMME

infrastructure is to provide a technical architecture based on international standards for implementing data management policies and workflows. Beneath common standards for metadata description and exchange such as OAI-PMH[4] and ISO 19139[5], ESONET has chosen to implement core standards of the Open Geospatial Consortium (OGC) Sensor Web Enablement (SWE) suite of standards[6]: namely the OGC standards SensorML, Sensor Registry, Catalogue Service for Web (CS-W), Sensor Observation Service (SOS) and Observations and Measurements (O&M).

However, in order to be useful for as many applications as possible, each of the above mentioned SWE standards represents a generic and abstract framework rather than detailed implementation rules, intentionally allowing much interpretative freedom and many different implementation approaches. To ensure internal compatibility so-called 'application profiles' have to be defined and accepted by OGC in cooperation with EUROSITES[7]. These standards are not yet implemented in parts of the EMSO infrastructure and the standardisation process is still an on-going process – in general, the maturity of the ESONET/EMSO IT infrastructure still varies amongst its components.

3.1.2 Users and stakeholders

EMSO is critical to provide information on the global environmental state, climate change, seasonal forecasting, safety at sea, developing applications for the offshore industry and fisheries, responding to accidents and pollution, and to defense requirements. Although one can easily envisage that the larger community is science-oriented, EMSO does not exclusively address scientists alone, but extends its services to a larger community interested in ocean physical processes, both natural and man-induced, over different time and spatial scales. EMSO aims at integrating European science by linking geographically scattered complementary research as well as industrial and governmental elements using data collected by deep-sea observatories.

EMSO users can be identified according to the following factors:

- **Sociological**: recognition of climate change and fears about environmental damage; fears about tsunamis and earthquakes; security fears; public fascination with the ocean.
- **Technological**: more smart sensors on the market; improved bandwidth; better power systems for remote subsea areas; increased access to broadband in homes; reduced costs for fibre optic cabling; more off-the-shelf solutions.
- **Economic**: increasing private sector spending on environmental monitoring; public sector investment in innovation; increasing natural hazard insurance claims; investment in bio-prospecting by the pharmaceutical industry; new markets in BRIC economies.

[4] http://www.openarchives.org/pmh/
[5] http://www.iso.org/iso/catalogue_detail.htm?csnumber=32557
[6] http://www.ogcnetwork.net/SWE
[7] http://www.eurosites.info/

SEVENTH FRAMEWORK PROGRAMME

- **Political and legal**: demand for environmental security; need for marine spatial planning; drive for sustainable development of marine resources; UNCLOS (United Nations Convention on the Law of the Sea) and need for international oversight of deep ocean resources.

These factors identify the following user groups:

- science;
- governmental institutions;
- industry;
- society (non-governmental entities);
- general public (*e.g.* aquaria, museums, exhibitions visitors);
- trainees (undergraduate and graduate).

Each user group must be adequately catered for by the interfaces by which EMSO presents its data to the world.

3.2 Data

The vision of EMSO is to allow scientists all over the world to access observatory data using an open access model. EMSO will deliver multi-parametric, long-term (spanning years) time series data addressing the seabed and water column.

3.2.1 Data products and metadata

EMSO observatories will be equipped with a common set of sensors for basic measurements and further sensors for specific purposes defined by the users. A non-exhaustive list of measurements / sensors that can be made / attached to seafloor observatories is listed below:

- seismic ground motion;
- gravity;
- magnetism;
- geodesy and seafloor deformation;
- fluid related processes monitoring;
- Chemical and Aqueous Transport (CAT);
- pore pressure;
- gas hydrate monitoring;
- dissolved *Fe*, *Mn* and sulfide species;
- acoustic tomography;
- CTD equipment for hydrothermal vents;
- methane;
- carbon dioxide;
- heat flow;
- nutrient analysers;
- *pH*, *Eh* and alkalinity;
- hydrocarbon fluorescence;
- in situ mass spectrometer;
- particle flux trap;

SEVENTH FRAMEWORK PROGRAMME

- image based particle flux;
- pigment fluorescence;
- deep biosphere sensors;
- time-lapse cameras;
- holographic imaging;
- videos;
- passive and active acoustics;
- zooplankton sampling;
- in situ sample processors with molecular/genetic probes;
- in situ respiration.

Although each EMSO node will have a suite of sensors in accordance with the scientific interest of the site, EMSO observatories will include a common set of sensors for basic measurements and further sensors for specific purposes defined by the users. The common set of instruments comprises seismometers, hydrophones for geophysics, magnetometers, gravity meters, CTD, current meters, chemical sensors, pressure sensors and hydrophones for bio-acoustic monitoring. Additionally, laboratory studies are performed on material collected at these sites by sampling devices (*e.g*. water samplers, sediment cores, traps *etc*.).

EMSO collects metadata on both the physical sensors and observatories as well as on the data. Observatories are intended to be described by SensorML. Metadata on archived data sets is compatible to ISO 19115, DIF or the NetCDF (CF) specification.

Ownership of data is indicated within the metadata for each dataset, being typically the principal investigator responsible for a given experiment (and thus the person holding the IPR). Though not yet fully realized, interoperability is intended to be supported by (for example) continuing to implement GEOSS registered standards.

One focus of ESONET was the development of a Sensor Registry which was intended to provide a catalogue on sensor and instrumentation data in SensorML format via a CS-W interface. The tool as well as content are still in a beta phase, but it is hoped that further development will occur.

3.2.2 Data archival

EMSO distinguishes between real time data and archived data. The latter is offered in two major formats, ASCII and NetCDF. For real time data, EMSO promotes the use of the OGC O&M format.

Three major institutions / data centres are currently offering access to EMSO data: UniHB (PANGAEA[8]), INGV (MOIST[9]) and IFREMER (EUROSITES):

[8] http://www.pangaea.de/
[9] http://moist.rm.ingv.it/

- PANGAEA offers access to metadata via OAI-PMH and delivers metadata in various formats including ISO 19136[10]. Data is delivered via a web interface or web services in ASCII format. The major observatory, the 'hausgarten' is not cabled; once this is done an SOS interface will be provided.
- MOIST offers access to data via a web interface in JSON[11] and NetCDF[12] format and offers metadata in DIF format. OAI-PMH is not yet implemented but planned. SOS interfaces are also planned.
- EUROSITES offers access to data via a FTP site in NetCDF format. Additionally a SOS server offers access to data in O&M format. Metadata is included within the NetCDF files.

The data is archived in three major data archives as mentioned above. Data is processed according to *e.g.* the OCEANSITES data quality assurance routines and transferred into the respective digest format. Metadata is added and completed (*e.g.* by data curators at PANGAEA).

PANGAEA is a WDS certified long term archive and offers the data via a web interface. Data sets are quality flagged. Data is published via the data archives web pages and via the ESONET data portal[13]; however a common EMSO data portal is planned. Data can be published as an article or via PANGAEA. PANGAEA uses DOIs to identify data and has agreements with scientific publishers such as Elsevier to link data via these identifiers.

3.2.3 Data delivery

The data infrastructure for EMSO is being designed as a distributed system. Presently, EMSO data collected in experiments at each site are locally stored and organized in catalogues or relational database and run by the institutions involved. As an example, observatory data, as well as data archiving services are already provided by IFREMER, UniHB and INGV. A common data infrastructure was outlined during the ESONET Network of Excellence project work-plan but has not yet been implemented. A central archive hosting a web-served access to all the databases is planned for the near future.

Data delivery of both small and larger data sets is done via FTP or HTTP. No cloud or grid services have been deployed yet but initial experiments have been performed. Data is not systematically replicated at different physical locations, but data archives perform the usual backup procedures.

The EMSO data infrastructure has been conceived to utilize the existing distributed network of data infrastructures in Europe and use INSPIRE and GEOSS data sharing principles. A number of standards have been set forth which will allow for state-of-the-art transmission and

[10] http://www.isotc211.org/Outreach/Overview/Factsheet_19136.pdf
[11] http://www.json.org/
[12] http://www.unidata.ucar.edu/software/netcdf/
[13] http://dataportals.pangaea.de/esonet/

archiving of data with the kind of metadata recording and interoperability that allows for more straightforward use and communication of data. OGC SensorML is an eXtensible Markup Language (XML) for describing sensor systems and processes. Following on progress from EUROSITES and others, a SensorML profile is being created that can be stored in a so-called Sensor Registry that will act as a catalogue of each EMSO sensor. This dynamic framework can accommodate the diverse array of data and formats used in EMSO, including the addition of delayed mode data.

3.2.4 Data availability

A common data catalogue is under construction for the future use of EMSO. Observatory data, as well as data archives that are already provided by IFREMER, UniHB and INGV. The amount of data produced per dataset depends on the instrumentation and configuration of the observatory, but is typically between several megabytes to several gigabytes, demonstrating notable variability of data products. EMSO is participating to the EC project SCIDIP-ES[14], which has the aim of realizing and deploying services for long term data preservation. EMSO, as a use-case in this project, will adopt the SCIDIP-ES results.

EMSO has a distributed e-infrastructure, so it could use local e-infrastructure provided by each partner and adopt existing European distributed e-infrastructure for storage and computation (*e.g.* Grid or cloud resources).

Although not performed systematically at every EMSO site, online web-pages (*e.g.*, `http://moist.rm.ingv.it/`) allow the user to visualize data from different instruments in the same area, with some data interpolation and correlation analysis capabilities. It will be possible to use software tools realized *ad hoc* to interpolate data, correlate different data series, visualize results.

Services partially in operation and/or under implementation:

1. **PANGAEA OAI-PMH** for ESONET data in EMSO sites: harvesting test, integration into ENVRI metadata catalogue *etc*..
2. **PANGAEA GeoRSS** for embedding GeoRSS feed.
3. **Ifremer SOS** for EUROSITES oceanographic data in EMSO sites.
4. **PANGAEA SOS** for INGV data in EMSO sites (via MOIST).
5. **MOIST OpenSearch** for INGV data and metadata in EMSO sites: data and metadata search according to time, space or other parameter.
6. **Common NetCDF metadata extraction / transformation service** for metadata extraction.
7. **MOIST OAI-PMH** for harvesting INGV data and metadata in EMSO sites: data and metadata harvesting.

[14] `http://www.scidip-es.eu`

3.3 Infrastructure

Although some EMSO sites are already equipped with operational monitoring infrastructure, EMSO has not yet undergone a design phase. The main challenge for the EMSO data infrastructure is to provide a technical architecture based on international standards to implement data management policies and workflows.

According to the EMSO statutes drawn up during the Preparatory Phase project, the technical architecture of each node pertains to the institutions involved in the node development. The EMSO-ERIC Scientific and Technological Advisory Board and the Executive Board will establish that general and detailed requirements and standards are fulfilled in order to integrate the node into a unique research Infrastructure.

As a general indication the possible node models are:

- **Cabled node**: the system is powered continuously by cable; data are transmitted in real-time to an land-based station. In this case monitoring activities can last from years to decades.
- **Stand-alone node**: the system is powered by batteries; acoustic and satellite communications guarantee the recovery in quasi-real time of segments of data or summary data messages from land. The average monitoring period varies from weeks to several months.

4 EPOS

The European Plate Observing System (EPOS; http://www.epos-eu.org) seeks to provide a research infrastructure for solid Earth science (seismology, volcanology, geodesy, tectonics, *etc.*). The RI is intended to integrate existing European facilities into one distributed and coherent multidisciplinary infrastructure allowing sustainable long-term Earth science research strategies and an effective coordinated European-scale facility for solid Earth data. The EPOS mission is to increase the accessibility and usability of multidisciplinary data from seismic and geodetic monitoring networks, volcano observatories, laboratory experiments and computational simulations. Specifically, EPOS will provide:

- A distributed infrastructure consisting of existing seismic and geodetic monitoring networks.
- Dedicated observatories for data acquisition (from volcanoes, fault-zones, deep drilling experiments, *etc*.).
- A network of laboratories presented as a single infrastructure for research into rock and mineral properties and analogue tectonic monitoring.
- Facilities for data repositories permitting the analysis, integration, visualisation and archiving of various datasets.
- Facilities for distributed storage and high-performance computation.

Currently EPOS is in its preparatory phase, aiming to bring the project to an adequate level of legal, financial, administrative and technical maturity. The present objective is to integrate (and, where necessary, create) the EPOS data centres, requiring a coherent architecture and implementation plan guaranteeing long-term sustainability. This entails extracting agreements from national bodies for support and funding; proper legal, governance and financial requirements must be met by each participating body. A prototype for the EPOS e-infrastructure must be created in time for the construction phase beginning in 2015.

4.1 Activity

EPOS is a federated distributed infrastructure, integrating the data and observations produced by a collection of national (and international) volcano observatories, data archives, GPS networks, seismic networks and experimental laboratories. EPOS intends to encapsulate these sub-systems within a number of EPOS Data Centres, representing community specific services with their own computational resources. These Data Centres are further integrated by the EPOS Core Services, which represent an infrastructure layer of common data services.

The design and establishment of the EPOS Data Centres is intended to be carried out during the EPOS preparatory phase in response to the needs and desires of the various contributing Earth science communities; likewise the design and composition of the EPOS Core Services.

At present, the most mature data infrastructure within EPOS is that of the seismology community, and their requirements and existing infrastructure dominate the current EPOS state-of-the-art. Within seismology there are a number of international data centres (such as

ORFEUS Data Centre[15] and the European Integrated Data Archive EIDA[16]) as well as participation in infrastructures for global integration of data in GEO[17]. There is also a well established policy for data exchange with national seismic networks as well as a long lasting tradition in data standardization, archiving and mining. Therefore, there already exist operational data centers and discussions regarding new implementations of metadata to guarantee interoperability of those data centres with other EPOS data centres.

4.2 Data

Scientists within the seismology community mostly use raw data and continuous waveform data. Accelerometric waveforms are also of interest to the engineering community. Data products (such as magnitude, earthquake size and earthquake locations) are of interest to policy and decision makers. In some cases there is interest from industry (mostly monitoring of induced seismicity).

4.2.1 Data products and metadata

EPOS members handle continuous waveform data in Standard for the Exchange of Earthquake Data (SEED) formats (along with associated metadata); aggregated data volume across all data centres is estimated to be in the order of one to several hundred terabytes. Accelerometric waveform data is typically handled in ASCII format. Earthquake catalogues are represented in many different formats, usually either text-based or XML-based.

Each data centre stores and preserve its data in accordance with their preferences; for example the portal at `http://www.seismicportal.eu` preserves user composed collections of datasets extracted from the EIDA networks and their extended metadata. Unique and persistent identifiers (URIs) are assigned to these datasets and the metadata are stored in a RDF Triple Store at ORFEUS. Currently the following distinction is made between data levels:

- **Level 0**: raw data.
- **Level 1**: quality control data.
- **Level 2**: filtered data.
- **Level 3**: research level pre-processed data.
- **Level 4**: research product.

Portals should explicitly assign identifiers (URIS) that clearly specify the origin of the data (up to the data centre storing and providing that specific copy of the dataset).

Metadata definitions are currently under discussion. Within seismology a task force will be established to define and store the concepts and the vocabulary terms for metadata items. Dataless SEED is the current format to describe instrument characteristics (derivative XML

[15] `http://www.orfeus-eu.org/`
[16] `http://eidawiki.orfeus-eu.com`
[17] `http://www.earthobservations.org/index.shtml`

SEVENTH FRAMEWORK PROGRAMME

formats are also in use, but common agreement has not been reached yet). The main requirements of the next phase of the metadata definition are based on standards being developed by EUDAT (see Section 10).

4.2.2 Data acquisition and archival

Currently the EIDA network is providing access to continuous raw data coming from over 1000 stations recording around 40MB per day – more than 15 TB per year. EMSC (European-Mediterranean Seismological Centre[18]) stores a database of 1.85 GB of earthquake parameters, which is constantly growing and updated with refined information, which at last measure included 222705 events, 632327 origins and 642555 magnitudes.

Accelerometric data are stored at LGIT (ISTerrre[19]). The data are received at data centres in real time, through dedicated TCP/UDP connections to the sensors, adopting the widely known application level protocol SeedLink.

Quality control processing and data transformation are already present in many data centres. These techniques will need though to be refined and standardized. Post processing techniques such as cross correlation and synthetic seismograms generation are the subject of use cases proposed within the VERCE initiative[20].

Typical Quality Control data for seismograms include:

- **Power Spectral Density**: PSD of the background noise as function of time, for selected frequencies.
- **Magnitude**: histograms for magnitude differences between station magnitude and VEBSN magnitude.
- **Time residuals**: time residuals distribution for each station.

Data catalogues and inventories are currently available under the EIDA initiative. Several websites and web services currently give access to the data stored within the EIDA network; one of these online services can be found at `http://www.seismicportal.eu`.

4.3 Infrastructure

The national seismic networks of nearly 60 countries are federated in ORFEUS (Observatories and Research Facilities for European Seismology) to exchange seismological data. Instruments and detectors are deployed at national or regional level and data are collected and archived in national data centers. Waveforms are also collected at the ORFEUS Data Centre (ODC) and real time seismograms (waveforms) can be downloaded by the EIDA web service. Therefore in seismology there is a centralized data center in

[18] `http://www.emsc-csem.org/`
[19] `http://isterre.fr/`
[20] `http://www.verce.eu/`

Utrecht and distributed repositories of real time waveforms. Services for users are already in place and their implementation is in the design phase within EPOS.

4.3.1 ORFEUS Data Center

The ODC collects in real-time seismic waveform data from more than 500 broadband stations in Europe, known as the Virtual European Broadband Seismograph Network (VEBSN[21]), using Antelope and SeisComP. Complementary data from the VEBSN, EIDA and temporary deployments are collected and archived using archive protocols (*e.g.* ArcLink and `mseed2dmc`). Currently, the archive contains 15 TB of continuous SEED data, starting from 2002, and pre-packed earthquake waveform data starting from 1988 (for European earthquakes of magnitude greater than or equal to 4.5, and global earthquakes of magnitude greater than or equal to 5.5). All data is openly available to the research community through a variety of services, such as web services, direct access and interactive tools. ODC operations are built around open software systems, like Linux and MySQL.

4.3.2 European Integrated Data Archive

EIDA is a consortium of waveform data centers that share a common agreement on issues related to data formats, metadata, transfer protocols and interfaces within the consortium.

The technical architecture consists of the ArcLink middleware, which is installed at each node of the consortium. It takes care to synchronize the station inventory amongst all nodes, allowing the exchange of waveform data within the consortium through a dedicated TCP/IP application protocol. End users can have access to the data preserved within EIDA by using several web pages, web services and standalone tools that are connected to EIDA via ArcLink. The variety of services available can be freely used by researchers depending on their need and they all guarantee access to the whole EIDA archive.

The nodes of the EIDA network exist in independent data centres that share a common format and protocols. Typically data is stored in file systems while metadata get saved within relational databases.

EIDA data and Earthquake parameters are generally open and free to use. Few restrictions are applied on a few seismic networks and access is regulated depending on email based authentication/authorization.

4.3.3 Computing resources

No general agreement has been currently formalized regarding computing resources. On the other hand, other initiatives referring to EPOS, such as VERCE and EUDAT, include HPC partners evaluating the opportunity of providing computational access through PRACE[22].

[21] `http://www.orfeus-eu.org/Data-info/vebsn.html`
[22] `http://www.prace-project.eu/`

Currently a massive investigation is on-going with respect to the adoption of new strategies for data ingestion and replication. This approach, implemented within EUDAT, is based on data centres hosting iRODS[23] servers which store and replicate the data, providing also unique and persistent ID (PID) to data granules through a federated handle system.

The VERCE and NERA[24] initiatives are working on the deployment within the community of those strategies and technologies aiming to achieve data aggregation and integration tasks, through the execution of distributed workflows.

Catalogues of intermediate datasets and users' annotation are under the attention of VERCE. Persistent identification will be implemented through the handle system provided by the EUDAT project. VERCE will lay the basis for a transformative development in data exploitation and modelling capabilities of the earthquake and seismology research community in Europe and consequently have a significant impact on data mining and analysis in research; this extends the work already started within the EC project ADMIRE[25]. A joint effort of EUDAT and VERCE will consider the interface of their technologies with PRACE.

[23] https://www.irods.org/
[24] http://www.nera-eu.org/
[25] http://www.admire-project.eu/

5 EURO-ARGO

Euro-Argo (http://www.euro-argo.eu/) is the European contingent of the Argo project. Its purpose is to ensure a sustained European contribution to the Argo system.

The science of climate dynamics and change is needed to understand changes to the atmosphere and oceans and shape policy and strategy. 80% of global warming has occurred in the last 50 years; it is to the oceans we must look in order to understand the redistribution of heat. Needed are global datasets – to this end, Argo is a global ocean observing system comprised of a large network of robotic floats distributed across the world's oceans and supporting infrastructure.

In a few years, Argo has become an essential part of the global ocean observing system. Continuous maintenance is required to preserve the system however, in the order of 800 to 900 new float deployments per year. Euro-Argo started in 2008, its purpose to procure and deploy about 250 floats a year, to monitor those floats and to ensure that all the data can be processed and delivered to users in both real-time and delayed-mode. Euro-Argo supports about 1/4 of the global Argo array, enhancing coverage (by about 50 floats) in European and marginal seas (the Nordic seas, the Mediterranean and the Black seas). Euro-Argo allows national bodies to pool their resources, establishing a high level of operational cooperation. Operation at sea, array monitoring and evolution, technological and scientific developments, improving data access for research and operational oceanography, all link to international management of the Argo programme.

The RI involves 15 organisations from 12 countries. Since July 2011, the RI has been in a transition phase before setting up the ERIC (European Research Infrastructure Consortium) planned for 2012. The ERIC is the starting point of a new European contribution to the long-term (greater than 10 years) operation of Euro-Argo.

5.1 Activity

Given increasing concern about global change and regional impact, sea level rises at a rate of 3mm/year (and rising), shrinking Arctic ice and warming of high latitude areas, a lack of sustained observations hinders development and validation of climate models.

Concerns about lack of observations of key factors led governments to form the Global Earth Observation System of Systems (GEOSS[26]) in 2003 and in Europe, the initiative on Global Monitoring for Environment and Security (GMES[27]). The climate and ocean components of GEOSS are delivered by the Global Climate Observing System (GCOS[28]) and the Global

[26] http://www.earthobservations.org/geoss.shtml
[27] http://www.gmes.info/
[28] http://www.wmo.int/

Ocean Observing System (GOOS[29]). The Argo network is a global array of autonomous floats, deployed over the world's oceans, reporting subsurface ocean properties via satellite transmission. It is now the major systematic source of information about the ocean's interior.

5.1.1 Float operation

After being released, floats dive to a programmable depth (currently 1000 metres), drifting freely in currents. Every 10 days, a float dives to 2000 metres, then rises to the surface to send data by satellite link. More than 200 cycles can be performed during the float's 4 year lifespan. A typical float possesses a salinity sensor, a temperature sensor, a pressure gauge and a satellite transmission antenna. Further development is being performed in order to include more sophisticated sensor packages, including oxygen and biosensors, as well as permit longer float life. Currently 3000 robotic samplers, each uploading a few KB of data approximately every 10 days. This amounts to around 100,000 uploads/year.

5.1.2 Satellite transmission

Data is collected from floats via satellite transmission. Argos has the ability to geographically locate the source of data anywhere on Earth using the Doppler effect.

- **Argos-2** has 5 steps: platforms send regular messages to the satellite; polar orbiting satellites receive their messages; antennas receive information from satellites; treatment centres receive data and distribute them to users; users receive data via different media to be shared with the scientific community or government.
- The first Argos-2 system has only one-way communication; floats spent 6-12 hours at the surface to transmit around 100 levels of only CTD (Conductivity, Temperature and Depth) sensor data.
- **Argos-3 and Iridium** reduce amount of time on surface, delay beaching and delay bio-fouling; also desired is the ability to modify mission parameters to monitor specific events.
- **Argos-3** spends less than 40 minutes at the surface, permits the installation of extra sensors (oxygen, chlorophyll, *etc*.) for the transmission of around 1000 CTD levels. Argo-3 also offers the ability to change float mission settings.
- **Iridium** uses the Iridium satellite constellation, spends only 3 minutes at surface; ability to observe surface currents by tracking movements of floats is lost, but the trajectories of floats becomes more representative of flow at their parking depth.

Array monitoring required to identify over- and under-sampled areas and to plan deployments accordingly.

5.1.3 Use-cases

Euro-Argo supports a number of use-cases:

[29] http://www.ioc-goos.org/

- **Research users**: study of global climate change; understanding the ability of the ocean to absorb excess CO2 from the atmosphere; understanding temperature increases and growing acidification of the oceans due to anthropogenic increase in greenhouse gas concentration; understanding how this modifies the geographic spread of marine species.
- **Operational users**: observations for oceanography and the GMES Marine Core Service (MCS). Euro-Argo has a legal framework within which contributions can co-exist with contributions from GMES. Argo and satellite data are assimilated into MCS models used to deliver regular and systematic reference information (including forecasts) on the state of the ocean for marine transport, industry, safety-at-sea (search and rescue) and fisheries.
- **Educational applications for the general public**: Argo data is available on the web and has been linked with the Google Earth GIS tool[30].

5.1.4 Contributions

Floats are deployed, often using ships of opportunity; Euro-Argo currently provides 1/3 of new contributions to global Argo array. The aim is to reach progressively 250 new floats a year. The Nordic sea area provides specific challenges: meso- and sub-mesoscale variability cannot be resolved by any reasonable float programme, so variability has to be treated as noise; at least 6 floats per basin have to be present in order to reduce hydrographic uncertainties below signal level.

5.2 Data

Euro-Argo is involved in oversight of all procedures related to data acquisition, technological developments, quality control sampling strategy, network design, product development, *etc*. in the European portion of the Argo system.

5.2.1 Data products and metadata

Argo measures: heat, salt transport / storage, ocean circulation and global overturning changes in order to understand (amongst other things) the ocean's absorption of excess carbon dioxide. Projects such as MyOcean[31] and SeaDataNet[32] use the Argo data to monitor the ocean as well as investigate ways to combine the available data to provide products to various types of user and to assimilate data into any model provided.

There is a need to know what data is available in order to improve observation as well as to provide easy access to all sources of data. Needed is a common catalogue with common categories and tools (6 metadata standards catalogues currently exist: EDMO, CSR, EDMED, EDMERP, EDIOS and CDI).

[30] http://www.euro-argo.eu/Outreach/Educational-Web-Site
[31] http://www.myocean.eu/
[32] http://www.seadatanet.org/

5.2.2 Data acquisition and archival

Data should be delivered with the shortest delay possible whilst still having extensive quality control. Three modes; real time (under 24 hours), near real time (a few days) and delayed mode. QC is highest for delayed mode data. Data management is geared to handle different versions of data in a consistent and coherent manner.

The International Data System is based on 2 Global Data Assembly Centres, 11 national Data Assembly Centres and several Argo Regional Centres:

- GDACs located at Coriolis (France) and USGODAE (USA)[33] are in charge of collecting data from the 11 DACs and provide access to the best version of an Argo profile. Data available in NetCDF format over FTP and WWW. The 2 GDACs synchronise every day.
- DACs receives data from satellite operators, decode and perform quality control on the data according to a set of 19 real-time automatic tests. Erroneous data are flagged, corrected if possible and then passed to the 2 GDACS and to the World Meteorological Office Global Telecommunications System (GTS). The GTS data stream does not include quality flags and bad data and grey-listed data are not transmitted on the GTS.
- ARCs provide expertise on specific geographical ocean regions to provide comprehensive data sets (including non-Argo data). ARCs act as the delayed mode operator for 'orphan' floats (deployed by institutes which do not have the ability to perform delayed mode QC); gather the recent complementary in situ ship-based data needed for delayed mode validation; check the overall consistency of the Argo dataset in an area.

For Euro-Argo specifically, Coriolis hosts one of the two GDACs; there are also 2 DACs operated by Coriolis and the British Oceanographic Data Centre and there are 2 ARCs - the Atlantic ARC and the Southern Ocean ARC.

5.2.3 Quality control

Quality Control (QC) must be applied on all three levels of data use:

- **Level 1** is the real-time system which performs a set of agreed checks on all float measurements. Real-time data with quality flags are available to users within 24-48 hours.
- **Level 2** is the delayed-mode system.
- **Level 3** is regional scientific analyses of all float data with other available data.

Profiles from floats are qualified individually; thus individual profiles may look good whilst not being coherent with neighbours. Consistency check measures have been developed within Euro-Argo to assess quality of bad data not flagged by real-time QC procedures (in collaboration with MyOcean project). Statistical methods which look at Argo data as a whole

[33] http://www.usgodae.org/

are being worked on, including an objective / residual analysis based on optimal estimation methods, an anomaly method wherein profiles are compared against different climatologies such as ERIVO and World Ocean Atlas 2005, as well as a gridded filed of all Coriolis data obtained 3 months before a given profile's measurement[34].

5.2.4 Monitoring

At sea monitoring relies on a rigorous record of exhaustive metadata at all steps of a float's life which allows:

* Traceability of the components integrated by the manufacturer.
* Information about the integration and acceptance tests done by the manufacturer and the customer.
* Information on float programming.
* Information on float deployment environment.

It relies on a monitoring systems which computes automatically some statistics and detects some events such as float death, grounding and transmission rate and is able to raise alerts automatically. It also relies on a person in charge of periodically studying alerts and coordinating actions.

At Coriolis a monthly analysis for quick problems detection requiring rapid action is performed; an annual analysis to provide an overview of float behaviour and report on evolution made during the year is likewise performed.

5.3 Infrastructure

Euro-Argo's organisation is distributed between a Central RI (C-RI) and various national facilities. The C-RI is responsible for overall coordination, participates in procurement and deployment of floats, has expertise in all aspects of the programme and acts as a resource centre for all participants. Specifically the C-RI:

- receives and manage funds from members, observers and from the European Commission (e.g. GMES);
- orders floats, receives them and tests them as necessary;
- organizes their deployment and ships them to the appropriate lab or port of call for deployment;
- monitors their performance in order to detect any dysfunction;
- contributes to technological developments, field trials, and implementation;
- engages in data management activities in relation with the appropriate Argo Data Centre;
- conducts R&D activities at European level;
- hosts scientists engaged in Argo related research;

[34] http://www.euro-argo.eu/Main-Achievements/European-Contributions/Data-Processing/Consistency-check-methods-on-T-S-from-Argo

SEVENTH FRAMEWORK PROGRAMME

The distributed RI agrees to commitment of resources and coordinates through the C-RI.

The Argo data management system links with other data management infrastructures such as MyOcean (via its Thematic Assembling Centre for *in situ* data) and the SeaDataNet data infrastructure (a “Pan-European Infrastructure for Ocean & Marine Data Management”, started October 2011 for 4 years).

6 ICOS

ICOS is the Integrated Carbon Observation System (http://www.icos-infrastructure.eu/), a distributed infrastructure dedicated to the monitoring of greenhouse gases (GHG) through its atmospheric, ecosystem and ocean networks. As such, ICOS's essential product is measurement data over a long-term period. Three Thematic Centres are defined to process in a standardized way data from all the stations from each network, and provide access to these data. At this time (beginning of 2012), only the Atmospheric and Ecosystem thematic centre (ATC and ETC) are under construction. The Ocean thematic centre construction will begin in 2013.

Currently, ICOS is in its Preparatory Phase (2008-2013), initiating the construction of facilities, ensuring the commitment of funding from parent institutions and governments, and providing a demonstration of use on a reduced scale.

6.1 Activity

The main service of ICOS is to provide data. The Thematic Centres offer software services to external sites for visualization and expert validation of their data. ICOS is also developing a Carbon Portal, where data discovery and access will be offered, as well as hosting of end-products elaborated from ICOS data by external users.

Typical ICOS (data) users include:

- National and international scientific programmes and environmental agencies (Global Carbon Project, WMO-Global Atmosphere Watch, NOAA collaborative network, *etc*.) that monitor greenhouse gas or climate data.
- Operational and pre-operational service providers for carbon fluxes (GMES projects MACC-II, Geoland2).
- Regional authorities.
- Protocol verification bodies.
- Scientific communities including remote sensing communities.
- The private sector (several companies invest in carbon monitoring and associated products).
- Educational organizations.
- The media and the general public ('citizen science').

The core use-case is that of a scientist researching carbon and other greenhouse gas concentrations, including their reactions and exchange processes with the Earth surface, requiring data from the ICOS station network. The scientist needs information on, among other factors, the availability, temporal and spatial coverage, quality of the data, and the processing algorithms used to produce higher-level data. Due to the nature of the research, which requires comparisons of temporally and spatially distributed observations, researchers must be able to rely that data is produced using comparable methods.

To provide this, the ICOS data infrastructure must collect observations from stations equipped with standardized measuring equipment, and process all the data in a similar manner.

6.2 Data

ICOS collects observations of GHG concentrations and fluxes, meteorological data and boundary layer height. ICOS data consists of the measurements done at the hundred stations of the 3 networks, and processed by the central facilities. Each of the 3 networks defines standard, commercial instruments for data collection. For ecosystem and atmospheric stations, level 1 and 2 stations are defined; Level 1 stations collect a full range of data, whereas level 2 stations only measure CO_2 and CH_4 concentrations or local fluxes, respectively.

6.2.1 Data products and metadata

Different steps of processing lead to a natural data hierarchy. The data hierarchy in the ICOS Atmospheric Thematic Center is divided into 4 levels:

- **Level 0**: raw data (e.g. current, voltages) produced by each instrument.
- **Level 1**: parameters expressed in geophysical units. For example it can be GHG concentrations (*e.g.* ppm-CO_2). Level 1 is also divided into two levels:
  - **Level 1.a**: rapid delivery data (near real-time, within 24 hours);
  - **Level 1.b**: long term validated data.
- **Level 2**: elaborated products. For GHG concentrations it can be *e.g.* gap-filling or selection.
- **Level 3**: added value products derived from lower level data.

The precise nature of level 3 data products is still under consideration with principal investigators, but will include dataset resulting from aggregation of multiple lower levels ICOS products. All atmospheric stations make continuous measurements as follows:

- CO_2 concentration;
- CH_4 concentration;
- N_2O concentration (optional);
- SF_6 concentration (optional);
- CO concentration;
- O_2/N_2 (optional);
- ^{13}C in CO_2;
- ^{18}O in CO_2;
- ^{14}C in CO_2;
- wind speed & direction;
- atmospheric pressure;
- atmospheric temperature;
- relative humidity;
- PBL (Planetary Boundary Layer) height;
- CO_2 flux (optional);
- Radon-222 (optional).

Stations can also make certain daily to monthly measurements (flasks measurements):

- N_2O concentration (optional);
- SF_6 concentration (optional);
- O_2/N_2 (optional);
- ^{13}C in CO_2;
- ^{18}O in CO_2;
- ^{14}C in CO_2.

Conversely, the data hierarchy in the ICOS Ecosystem Thematic Center is divided into 5 levels:

- **Level 0**: Raw data,
- **Level 1**: First set of corrections applied to the raw data,
- **Level 2**: Consolidated half-hourly fluxes,
- **Level 3**: Standardized QAQC and filtering applied to the half hourly data,
- **Level 4**: Data gap-filled and aggregated at different resolutions,
- **Level 5**: Derived variables calculated) data products.

The data collected at the ecosystem sites are raw data at 10 Hz time resolution. These data need a first processing step to calculate greenhouse gas fluxes with typical time resolution of 30 min. These fluxes are further corrected, filtered, gap-filled where necessary, and processed to retrieve additional variables. For the ecosystem data, the following continuous measurements are made:

- CO_2, H_2O and energy fluxes;
- soil heat flux;
- high (L1) / normal (L2) precision CO_2 concentration vertical profile;
- net radiation;
- incoming/reflected/diffuse global radiation;
- incoming/outgoing long-wave radiation;
- albedo;
- incoming/reflected PAR (Photosynthetically Active Radiation);
- spectral reflectance in selected wavelength (L1);
- relative humidity;
- temperature vertical profile;
- soil temperature and water content profile;
- wind speed & direction;
- air pressure;
- canopy temperature (L1);
- precipitation, through-fall, ground water level, snow depth;
- sap flow (L1).

Daily to monthly measurements:

- Leaf Area Index (LAI);
- soil respiration (L1);
- CH_4, N_2O by automatic (L1) / manual (L2) chambers;
- plant respiration (L1);
- phenology.

Yearly measurements:

- biomass (above ground);
- soil carbon;
- stem diameter;
- above-ground NPP (Net Primary Production);
- litter fall;
- carbon and nitrogen import / export on managed sites;
- bulk nitrogen deposition (L1);
- leaf nitrogen content (L1);
- soil water nitrogen content (L1);
- land-use history;
- managements and natural disasters.

For all types of site, metadata information includes;

- site information, spatial coordinates;
- measurement equipment (*e.g.* Tank) information;
- data processing information;
- data version information.

For atmospheric data, metadata are provided by principal investigators via graphical applications developed at the ATC. Raw data are transferred daily to the ATC where data are automatically processed. Those raw datasets are mainly ASCII files, depending on the source instrument. A specialized processing chain is dedicated to each type of instrument deployed in an ICOS Atmospheric Station. The process involves the transformation of raw data (Level 0) to upper level products. A level 1 ICOS Atmospheric Station is continuously measuring 18 parameters (among them greenhouse gases, meteorological parameters and planetary boundary layer height). Most data are continuous measurements. About 400 MB from level 1 ICOS Atmospheric Stations are daily uploaded to the ATC. Considering that the ICOS Atmospheric network will comprise about 50 atmospheric observatories, the amount of data produced is estimated to be around 20 GB/day, *i.e.* 7.3 TB/year. Note that that is an upper bound value since not all stations are going to be labeled as level 1 ICOS atmospheric stations. A data catalog of produced datasets is not yet automatically available but is intended to be in future.

Metadata is expected to be handled similarly by the ICOS ETC.

6.2.2 Data acquisition and archival

The data processing philosophy is based on answering the following needs:

- near real time data collection & processing;
- communication with station PIs immediately in case of problems, annually in case of no problems;
- archival of data, ensuring traceability (metadata);
- dealing with data from associated sites (if they meet ICOS requirements);
- provision of data products and ‘quick-look’ tools.

Currently, data is collected at the thematic centre servers where it is periodically sent by FTP from the stations. Data and metadata are stored on a dedicated MySQL server (at the ATC). The data processing follows the hierarchy described above.

Data are daily uploaded from ICOS atmospheric stations onto a dedicated FTP server at ICOS ATC. Note that data can be exceptionally provided via attached document in an email. ATC raw data are automatically ingested into a MySQL database, and then processed. ICOS ecosystem sites submit their raw data monthly to the ICOS ETC. In addition, preliminary half hourly fluxes and meteorological data will be transferred automatically to the ETC in near real time (around 24 hours after the event).

Data and metadata from ICOS ATC are stored in a dedicated MySQL server at ICOS ATC in France. The metadata include system information such as site, instrument, tank information and processing information such as when and how data were processed. There are no distributed repositories. Backups are made daily by an automated backup robot. Offsite backup is also planned. ICOS ETC will ensure a database structure able to store all the data, both fluxes and ancillary information, track all the changes in the data and, for processing using versioning systems, provide all the meta information about the sites, the variables, the instruments and the processing methods applied. A backup system and mirror servers are planned for the ETC and the ETC database will probably be based on SQL and .NET.

For the ATC, processing and maintenance are managed by the IT team at ICOS ATC. It is planned to make annual or semi-annual global dataset releases. Data management tools are also under development to give to the principal investigators a means to access and perform quality control on their data. For most applications, a secured graphical interface will allow to principal investigators to check and screen their data. ETC maintenance is managed in Italy, but the ETC will also receive quality checked and standardized data from the French and Belgian ETC sections. Every month, all raw data will be processed and each step fully documented.

Interoperability with other systems is under consideration via the deployment of an OpenDAP based solution for data distribution.

6.2.3 Data processing

For the ATC, a global automated filtering process is running when raw data are ingested into database. It is a filtering on external instrumental parameters provided along with raw measurements. More elaborated techniques (*e.g.* spectral analysis) are under consideration. ICOS ETC is responsible for the development and upgrade of the open source software that will be used for ETC data acquisition at the different sites and for technical assistance on the software use and configuration.

Near real time ATC products made available on internet in a 24 hour delay. These are low precision/un-flagged data in comparison to final products. On other hand, these products are quickly available. Different forms of data processing are under consideration, and a list of data products provided to users is under elaboration. As for the data processing part, the ETC is responsible for data post-processing. The ETC will implement automatic tests, *e.g.*

cross-checking of correlated variables within and across sites and calculating budgets, in order to spot possible problems.

6.3 Infrastructure

The ICOS technical architecture reflects the progressive integration of sensors into stations, of site data to central facilities, and from central facilities to the end users through the Carbon portal. The Thematic centres and their servers will act as interfaces for different atmospheric stations, while the Carbon Portal acts as the interface between the TC and the outside world.

The ICOS technical architecture reflects the progressive integration of sensors into the stations, of site data to central facilities, and from central facilities to the end users through the Carbon portal (see Figure 1 - ICOS distributed architecture below):

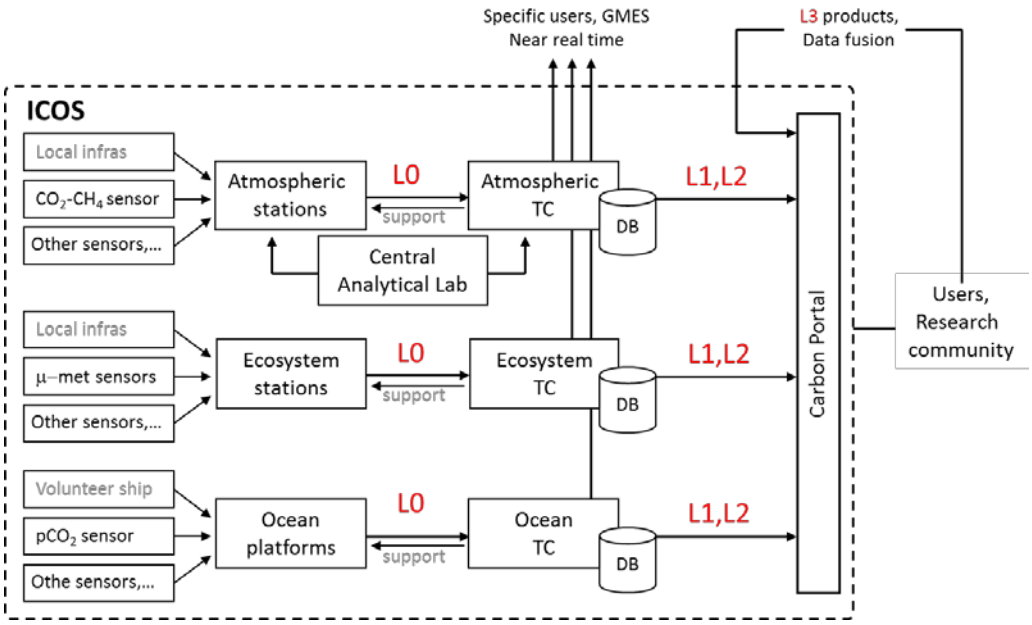

Figure 1 - ICOS distributed architecture

6.3.1 Computing resources

ICOS ATC data center is already deployed. Computing resources are composed of three dedicated servers: one for the data distribution, one for secured access and graphical applications and one for the database and computations. ICOS ETC is under construction – hardware selection is not yet fixed.

Data will be accessible through a license with full and open access. No particular restriction in the access and eventual use of the data is anticipated, excepting the inability to redistribute the data. Acknowledgement of ICOS and traceability of the data will be sought in a specific, way (*e.g.* via DOI of datasets).

The data will be accessed either through the relevant Thematic center for advanced users of specific data or through the Carbon portal where data discovery and access systems will be

provided online. A single login system will be put in place to monitor access to the data and ensure acknowledgment of the data policy.

6.3.2 E-science capabilities

E-science capabilities are put into place for meta-data collection via the development of an ICOS specific graphical application. Data integration is essentially done via centralized databases located at the various ICOS thematic centers which receive ASCII files from the stations. Limited integration is done at the station level to merge regular data acquisition with calibration sequences.

ICOS is looking at the use of the digital object identifier (DOI) mechanism for proper data attribution.

The carbon portal of ICOS will allow for data search across the different ICOS databases. Data search restricted to geographical areas or time periods, of similar nature than the one developed on the INSPIRE GEOPORTAL, will be implemented. The Carbon Portal will also act as a platform to offer access to higher level data product and fluxes.

6.3.3 Common user services

Level '1.b' data products such as near real-time GHG measurements are available to users via the ATC web portal[35]. Based on Google Chart Tools, an interactive time series line chart with optional annotations allows user to scroll and zoom inside a time series of CO2 or CH4 measurement at a ICOS Atmospheric station. The chart is rendered within the browser using Flash.

Some level 2 products are also available to ensure instrument monitoring by PIs. It is mainly instrumental and comparison data plots automatically generated (using the R language and the Python `Matplotlib` 2D plotting library) and pushed daily onto ICOS web server.

Level 3 data products such as gridded GHG fluxes derived from ICOS observations increase the scientific impact of ICOS. For this purpose ICOS supports its community of users. The Carbon portal is expected to act as a platform that will offer visualization of the flux products that incorporate ICOS data. Example of candidate Level 3 products from future ICOS GHG concentration data are for instance maps of European high-resolution CO2 or CH4 fluxes obtained by atmospheric inversion modelers in Europe. Visual tools for comparisons between products will be developed by the Carbon Portal (see early efforts in Carboscope[36]). Contributions will be open to any product of high scientific quality.

[35] `https://icos-atc-demo.lsce.ipsl.fr`
[36] `http://www.carboscope.eu`

6.3.4 Operational services

Operational data stream provision is in construction between ICOS and GMES (Global Monitoring for Environment and Security[37]). ICOS produces both Near Real Time Data (daily frequency) with fully automated data processing and high precision data (6 month frequency) where human expertise is implicated.

Quality control in ICOS implies both automated mechanisms and manual expert control. Early warning systems are put into place. Expert control is done via visual inspection of quick looks. For atmospheric measurements, the use of target gas and measurement duplication with flasks contribute to the quality assessment of the data.

ICOS also participate in international inter-comparison program sometimes with collocated measurements to assess inter-comparability of the measurements.

A mobile lab will circulate among the ICOS network to duplicate measurements as part of the QA/QC ICOS plan. ICOS is in close collaboration with WMO to determine best practices for making measurements.

[37] http://www.gmes.info/

7 LIFEWATCH

LifeWatch (http://www.lifewatch.eu) is a research infrastructure for studying biodiversity and Earth's ecosystems. The LifeWatch infrastructure will allow scientists to tackle big basic questions in biodiversity research, as well to address urgent societal and fundamental scientific challenges concerning the planet. Advanced systems oriented research on the complex biodiversity system will be supported through dedicated virtual environments, enabling integrated access to data, analytical and modelling workflows and computational capacity. In LifeWatch, policy makers will be able to work on these problems directly or with assistance from local researchers in platforms actively investigating a policy-science interface. Environmental policy can move from uncertainty to increased confidence and safety.

The present intended countries of the LifeWatch ERIC are: Belgium, Finland, Greece, Hungary, Italy, Netherlands, Romania, Spain, Sweden. The LifeWatch preparatory project finished in January 2011. LifeWatch is currently in a "start-up phase" funded by 3 countries, with full construction starting late 2012.

7.1 Activity

LifeWatch is a distributed facility infrastructure. The LifeWatch ERIC will operate its Common Facilities: the operations that have to be coordinated and managed at a central, European level. These are the infrastructure scientific, technical and service activities that require a common and integrated effort at an European scale. The LifeWatch Common Facilities are located in three countries. Spain is hosting the Statutory Seat and the coordination of the ICT core functionalities; the Netherlands is hosting the Virtual Lab developments and the Innovation Lab; Italy is hosting the Service Centre to support end-users. LifeWatch Centres, established as independent entities in the member countries of the LifeWatch ERIC or as thematic services for scientific networks, contribute to the infrastructure construction and operation with parts of the core functionalities or with specialized capabilities (e-Labs). These contributions are part of the LifeWatch research infrastructure and have to meet the standards and protocols according to the LifeWatch Reference Model. Construction contracts and service-level-agreements secure effective operations and provide the basis for calculating national in-kind contributions.

LifeWatch will construct and bring into operation the facilities, hardware, software and governance structures in an e-Infrastructure for research on the protection, management and sustainable use of biodiversity. The infrastructure includes enabling facilities for data generation and processing; a network of observatories and sensors; facilities for data integration and interoperability; capabilities to create work flows of analytical and modelling tools; and a Service Centre providing special services for scientific and policy users, including training and research opportunities for young scientists. Its Grid-enabled Service-oriented design (Service Network) supports access to and the integration of external resources such as data from associated infrastructures and distributed computational capacity from high performance clusters. It can be characterised as a Collaborative Spatial Data Infrastructure sitting on top of available distributed computing infrastructures. User groups may create their own e-laboratories or e-services within the common architecture of the infrastructure. Any new data will be channelled to the appropriate external resources

such as GBIF (Global Biodiversity Information Facility[38]). LifeWatch enables distributed large scale and collaborative research on complex and multidisciplinary problems.

Key challenges which must be tackled by LifeWatch include:

- Data discovery, access and visualization.
- Developing a generic ICT environment, allowing users to build their preferred virtual labs (including their preferred data and software selections), and allowing them to create easily workflows.
- Annotation, provenance, and citation tracking.

LifeWatch caters to a range of kinds of user, including basic and applied scientists, private companies (*e.g.* environmental engineering companies), policy makers, agricultural engineers, epidemics analysts and citizen scientists. Scientists are served with data and software tool discovery, analytical and modelling capabilities, collaborative tools, associated computing power and various publishing options.

Key use-cases which LifeWatch seeks to cater to include:

- Population biology of migrating birds.
- Understanding the role of marine wetlands on biodiversity migration patterns.
- Impacts of invading alien species.
- Assessment of regional biodiversity.
- Preservation of ecosystem services (curing habitat destruction).
- ICT support for human observations of biodiversity.

Formal specifications of use-cases can be found at `http://www.uva-bits.nl`, or in the work of the BioVeL (Biodiversity Virtual e-Laboratory[39]).

7.2 Data

LifeWatch is not a research infrastructure generating data, but an environment to discover, process, model and collaborate with data (including software tools). Primary data originates from various sources. Contacted are: EBONE, GEOBON, GBIF, BioCASE, PESI, LTER-Europe, MARS, EBI-ELIXIR, ESA, EVOLTREE, Species2000, TWReferenceNET, Lagunet, ALTERnet, EMBRC, EUMON, DAISY, ALARM, BALLOON, ELNET, BIOFRESH and EUROCEANS. Data can also be generated by individual scientists who want to share their data.

7.2.1 Data products and metadata

Principal data products include: species names lists (integrating various concepts, life history and ecological attributes), specimen data, field observations (all these with associated spatial

[38] `http://www.gbif.org/`
[39] `http://www.biovel.eu/`

coordinates, date, time, size, environmental data as salinity, collection or observer name), monitor data (temp, height, precipitation, humidity, plant physiological responses, CO2, genetic data (sequences, genes), habitat and landscape data (aggregated data) and earth observation data. Given the range of data products, the size of a 'typical' data product is highly variable. All processed and published data should be automatically annotated with metadata.

Secondary data (*e.g.* model results and provenance data) can be generated as a result of processing and analysing primary data within the infrastructure. Likely volumes of data to be handled by LifeWatch (including primary data imported from integrated sources will be in the region of hundreds of terabytes.

Currently a diverse set of ontologies describes various aspects of the data. Further work to link such ontologies (*e.g.* through a thesaurus and associated semantic network) is presently being considered.

7.2.2 Data preservation

Time series data are very relevant for environmental research. As such, data preservation is very relevant. Since input data originates from external data resources, those resources store the data, though there may be circumstances where LifeWatch is requested to provide storage services when no external service is available. Processed (secondary) and published data can be stored by LifeWatch. Storage will be in delegated to facilities of third parties.

7.2.3 Data processing

LifeWatch specialises in various forms of data integration (including schema and formatting services) and a variety of analytical and modelling services and publishing, as predicated by its use-cases. LifeWatch is offering users a variety of processes, analyses and models.

There are quite a few analytical tools and methods offered by various institutes, but which today are not offered through a single portal together with necessary data and sufficient computational power. The provision of such a portal, along with new processes, analyses and models created specifically for LifeWatch, is a principal contribution of LifeWatch. LifeWatch aims to bring these together through its catalogue of services and to make them available for composing in workflows and virtual laboratories.

7.2.4 Data publication

LifeWatch will provide visualization tools, such as mapping tools. It is intended for there to be a proper process for publishing data to the outside world.

Data is shared with all external data resources. For all LifeWatch users, deploying data from external data resources requires agreement with their conditions. Agreements of LifeWatch with these resources will facilitate this. There exist various ethical and legal concerns. Data resources have various policies for sharing data. Location data of for example endangered species are not open for everyone. Data on deceased or quarantined organisms also have various restrictions.

7.3 Infrastructure

The choice of technologies for various parts of the infrastructure are not yet fully resolved. But in principle, frameworks for Service Oriented Architectures, based on SOAP and REST models for Web Services are being used.

The core resources within LifeWatch include:

- The services catalogue;
- Datasets catalogue;
- Annotations repository;
- Applications servers;
- Virtual collaborative environments;
- Provenance and citation tracking repository;
- Security (AAA) services (possibly outsourced);
- Access to computational resources;
- Portal framework.

The RI also uses resources affiliated with other RIs such as EGI, PRACE, GEANT, ICOS, GBIF, EBI-ELIXIR and ESA.

7.3.1 Interaction

Users are expected to interact with LifeWatch via the LifeWatch portal, offering various pre-constructed virtual labs, and with services to build new dedicated virtual labs. LifeWatch defines an access policy (still in draft form at the time of writing) which aims to offer open access to its resources and capabilities to all users, but which at the same time controls access to data and resources.

Authentication within the RI has not yet been resolved. Two parallel mechanisms are required: firstly, an approach to support access by individuals with an institutional affiliation *e.g.* eduGAIN federation[40] (using Shibboleth); secondly another mechanism for individuals without institutional affiliation or without eduGAIN credentials – OpenID is being considered. In this aspect, LifeWatch hopes to benefit from external developments (particularly in EUDAT).

The architecture of LifeWatch is layered. The lowest layer contains the (external) data, software and computational resources. The e-infrastructure layer integrates these, while the composition layer allows for the composition of preferred workflows. Users can access these services through specialized e-laboratories, or e-services, and can build their own ones. Distributed construction is 'controlled' with a LifeWatch Reference Model, including standards, procedures *etc*.

[40] http://www.edugain.org/

7.3.2 Data transport

In principle, data should not be moved where possible – computation would be performed remotely (*e.g.* in e-labs and clouds). All data assets acquire a persistent identifier on first entry to the infrastructure. This will be used for tracking purposes. Most intense computations are for modelling complex systems and for scenario analyses. Some of these may require access to high-performance computing facilities. Data integration is automatically performed upon selection of constituent data products. Analyses and modelling is scheduled by users.

Data has to be dealt with in (near) real-time, and be archived and processed later in accordance with specific use-cases. Near real-time data originates from sensor networks belonging to external projects. Reproducibility of operations (experiments, transactions, data integration and access) is important to the RI. Correct attribution of data to the RI by its users is crucial.

Reconstruction of data products by repeating previous analyses (rather than archival of all integrated data products and queries) is preferred to avoid storage of multiple end-products. However, it depends on computation time for reconstruction, which may justify the preservation of certain secondary products in certain cases.

8 OBSERVATIONS

The first thing to note about the six ESFRI projects which ENVRI chooses to survey is that fundamental differences exist between many of them. For example, EPOS and LifeWatch are principally concerned with integrating existing national and international research infrastructures to create a federated super-infrastructure that offers access to a broad range of multidisciplinary data products and computational services. Euro-Argo and EISCAT_3D have much more narrowly defined scopes, being principally concerned with being a source for a particular (albeit sophisticated) class of data product. In question then is whether generic conclusions can be drawn that are applicable to all projects equally.

Another difference is in the maturity of some of the RIs. To illustrate:

- Euro-Argo is already a mature project which has been in operation for some years. Whilst ENVRI can learn from the experiences of Euro-Argo, any reciprocal benefit will likely have to be directed to affiliate projects of Euro-Argo such as SeaDataNet, which faces much the same challenges as many of the other RIs surveyed in this report.
- EISCAT_3D is yet to build much of its infrastructure, but there is considerable prior experience (spanning decades) within the EISCAT Scientific Association of the kind of facility being built.
- EPOS has only just started, but builds upon existing standards and data archives within the seismological community specifically. The federation over these existing data archives, whilst maintaining such aspects as attribution of data to source, is not a solved problem.
- Similar issues affect LifeWatch, though that infrastructure is more evolved, with an existing reference model (similar to that being planned by ENVRI) already in existence.

Another consideration is the logistics of maintaining the various research apparatus. For example, the maintenance of robotic floats used in Euro-Argo, or the maintenance of seismographs. For Euro-Argo, this is an essential part of the RI's operations, whilst for EPOS this might be considered to be devolved to individual institutions, and not of direct concern to the overarching RI itself. Therefore, it is unlikely that ENVRI has anything to contribute here, particularly when contrasted to more generic problems involving federation, (near) real-time data handling and attribution.

8.1 Services

Fundamentally, each RI exists in order to provide a service to one or more constituencies of users. Aside from professional scientists, there may be more casual users (‘citizen scientists’), as well as data analysts and policy makers. The services offered by the RI may be unified for all classes of user, or may be differentiated (possibly by the interface by which each class of user interacts with the RI).

Fundamental to the uptake of services offered by RIs are the tools and interfaces exposed to users. If only a well-informed cabal of users is able to make easy use of the provided services, then there will not be any dramatic benefit to the wider user community, since they will not be using those services.

SEVENTH FRAMEWORK PROGRAMME

Services exist on a spectrum from those that are completely delivered through ICT means to, on the other end, those that are completely delivered by human action. In between there is a broad range of mixed services having more or less ICT and human elements. It is unlikely to be consistent from one RI to another. An apparently similar service existing in two or more RIs may be delivered anywhere along that spectrum in each RI.

8.2 Architecture

The ICT infrastructure (the architecture) of an RI is considered to be the technology platform underlying the computational science being performed. It is the network of interconnected ICT resources (physical and virtual facilities for the management of data and delivery of services, as well as for computation in general) provided to permit the RI to function.

Most of the RIs are geographically distributed, with a significant network of data collectors associated with a (usually) smaller number of dedicated archival or processing sites. In a few cases considerable processing is done at point of collection (EISCAT_3D), and indeed considerable processing is required to gather the data (EISCAT_3D again, for beamforming). Many RIs are reliant on partnerships with independent institutions which provide archival or processing facilities (e.g. EMSO, EPOS, LifeWatch). This requires both the adoption (or creation) of standards for data formatting and exchange, and contracts of service (Service Level Agreement, SLA). Some RIs are dependent on advanced telecommunications infrastructure– for example Euro-Argo requires regular satellite communication to occur between satellite constellations and its portion of the Argo robotic float network.

The technical infrastructure of the RIs is in many cases still under construction or even only being planned. For example in ICOS, the Atmospheric and Ecological Thematic Centres are well on their way to being built, but the Oceanic Thematic Centre will not emerge until much later in the project.

ENVRI will not have influence on the actual architectures used by RIs, but it will potentially be able to make concrete observations about how different architectures deal with or exacerbate specific data and computational issues. ENVRI will also be able to make suggestions about the components needed to fulfil particular functions.

8.3 Data

Whilst computation and the provision of services such as portals is of great importance to all RIs, the management and curation of data remains the overriding concern.

It is difficult to separate data from the underlying RI architecture; data format, transport, storage and replication are all constrained by the composition of the underlying ICT infrastructure. Many kinds of data might exist within an RI: primary (raw) data, secondary (processed) data, semantic (referential) data, metadata and provenance data. The same data may be considered to be of a different kind by different observers. For each data product, there may exist one or more data schemata describing it. Each RI has its own data lifecycle; typically:

- *Acquisition* of data from sensors, experiments and simulations, as well as human annotation. Data may be cleaned, classified, catalogued and annotated automatically.
- *Analysis* of data, including periodic quality control, which may lead to additional discovery of data from data mining.
- *Integration* and *harmonisation* of data, statically or dynamically, leading to derived data products.
- *Preservation* of data, entailing storage with guarantees of fidelity. Also *archival* of data, concerned with long-term preservation, possibly at the cost of easy accessibility.
- *Disposal* of data, concerned with the deletion of data to free resources and prevent the interception of sensitive information.
- *Publication* of data, providing access to data whilst ensuring correct attribution; this also covers *visualisation* of data.

The precise nature and understanding of this lifecycle varies between RIs, but all RIs have demonstrated that they take dealing with their domain's data lifecycle seriously.

Consideration must also be given to the sharing of data between jurisdictions via affiliated resources. Data may be accessed remotely from within its own jurisdiction, or may be transferred into the jurisdiction of another RI. This may require the reassignment or further assignment of persistent identifiers. This relates to the general problem of data movement. Data transport in general requires redundancy of data and common protocols. Specific issues arise based on the data involved:

- Movement of *scientific* data requires mechanisms for moving large volumes of data directly and indirectly within and between resources and jurisdictions, accounting for bandwidth and robustness.
- Movement of *event* data requires mechanisms for moving small packets of data between resources and services in order to invoke or prepare for specific activities.
- Data *replication* requires automatic distribution of copies of scientific data amongst RI resources. *Coherent* replication requires update propagation as well.
- Propagation of referential data is also necessary to ensure that the recipient of scientific data has access to its semantics.

All RIs have data-flows critical to their construction. These are generally single-directional from sensors / data collectors onto archival sites which then must be able to handle substantial numbers of data requests for data products of varying size. Many RIs work with sampled time series (*e.g.* EPOS and Euro-Argo), which offer further complications for data management, since data must be either pre-sampled, or services must permit efficient retrieval of specific time samples from larger products. Retrieval of specific samples must work within the framework of any persistent identifier system constructed.

8.4 Computation

Each RI exists to facilitate a particular form of computational science, either as a data provider, a service provider or both. As such there exists a body of algorithms, processes and models intended to be applied to the data held by each RI. Each RI must be able to either serve the data to a computational platform, or provide one alongside its data. This necessitates simple interfaces for retrieving data or scheduling computations. There is also

SEVENTH FRAMEWORK PROGRAMME

need for technologies to maintain the data products held by each RI and perform standard analyses. Such technologies may include operating systems, program libraries, distributed file systems, database management systems, distributed query systems, resource portals and workflow engines.

One question is whether the RIs provide sufficient computational resources to actually perform the computations desired by their users. Certainly, not all the RIs consider providing computational services to its users as part of their missions, except perhaps where analyses of raw data are standard (*e.g.* EMSO and ICOS). Conversely, other RIs consider such services to be integral (*e.g.* EPOS and, to a lesser extent, LifeWatch). Some advanced analyses require access to high-performance / high-throughput computing facilities, which if not provided by the RI, must be enlisted from elsewhere. In either case, some thought needs to be given as to how to stage data onto such facilities and how then to curate the results.

8.5 Working practice

Working practice is concerned with the interaction of organisations with technology, particularly with the constraints imposed by the nature of the research being performed. Constraints come in two forms:

- *Soft* constraints concern the working practices of domain experts and relate to how any new technological platform can be integrated into those working practices in a way that those experts would deem acceptable.
- *Hard* constraints concern concrete issues of ownership, security, legality and ethics.

Violations of soft constraints reduces interest in a technology; violations of hard constraints create liabilities for users interacting with a technology.

Soft constraints include considerations of how a given body of work is performed, the nature of that work, who does the work and *why* the work is done as it is. Attempting to integrate the scientific process with new technology in ways which actively interferes with preferred working practice is difficult.

Hard constraints include legal constraints, formally-agreed policies, international agreements (*e.g.* regarding interchange data or metadata standards) and commitments to data suppliers and data users to provide a given quality of service or a given API. Hard constraints must address privacy and ethics. These concerns require that the norms concerning the use of any aspect of the system or its data are documented, understood and complied with. A system must also correctly account for the people, data and resources in the system as well as identify ownership, rights, credit and blame. This relates to the 'AAA' model of Authentication, Authorization and Accounting.

Considerable effort is being put (or has been put) into the fulfilment of hard constraints by the ESFRI infrastructures – a significant part of the current phase of EPOS for example is the formalisation of agreements between different national bodies which are expected to contribute resources to the infrastructure. The experience of all the RIs in this area should certainly be fed into ENVRI's contribution. Significant consideration should also be given to soft constraints however. In this case whilst there is clear evidence of such consideration on the part of the RIs already, there is also clear opportunity for ENVRI to articulate these constraints more completely and feed these back to the RIs.

8.6 ENVRI's contribution

One thing that is clear is that ENVRI cannot ignore the various other European projects which seek to provide common solutions to problems affecting the various ESFRI infrastructures. Whilst ENVRI is notable for attempting to provide a high-level model for describing various aspects of the research infrastructure 'problem', other projects exist focused on data management, data transport or providing a platform for computation under various frameworks. Such projects include GEOSS, PRACE and GÉANT. In the following sections however, we shall survey three generic infrastructures, one principally (though not solely) focused on computation (EGI), one nascent project focused on data (EUDAT) and one supporting hybrid data infrastructures and virtual research environments (D4Science).

9 EGI

The European Grid Infrastructure (EGI; http://www.egi.eu/) is a European e-Infrastructure supporting open science in Europe and across the world, interoperating with integrated infrastructures in Canada, Latin America and the Asia-Pacific region. EGI is centrally coordinated by EGI.eu and governed by the EGI Council with members from 34 European countries as well as 4 associated members. It is supported by the EC through the EGI-InSPIRE project, whose consortium includes 50 partners in Europe and the Asia-Pacific area, including 37 National Grid Initiatives and two European International Research Organizations (CERN and EMBL)[41].

9.1 Activity

EGI is a production infrastructure operationally running and delivering IT services to multiple scientific disciplines including: the earth sciences, physics, the life sciences, computational chemistry, the humanities, computer science, astronomy and astrophysics. The grid computing resources provided by the EGI federation are being used by scientists and researchers across Europe and beyond[42]. The Virtual Research Communities that are partners of EGI are: CLARIN, DARIAH, HMRC, LSGC, WeNMR and WLCG. EGI has a partnership with a number of EC projects, including MAPPER, Mantychore, ScalaLife, and SIENA[43].

9.1.1 Key IT activities

Key IT activities include:

- **User and resource centre requirement collection**: this is an on-going process established in EGI to ensure that services meet the needs of research communities and evolve with them. Requirements are publicly accessible[44].
- **Production deployment of IT services provided by a number of technology providers**: EGI does not develop software, but facilitates the deployment and operation of platforms needed by research communities. The main current technology providers are the European Middleware Initiative (EMI) providing ARC, dCache, gLite and UNICORE middleware, and the Initiative for Globus in Europe (IGE), providing a customized release of GLOBUS. However, EGI is open, and can liaise with other technology providers according to the needs of the user and operations community. Software platforms are internally validated and tested before being rolled to production.

[41] http://www.egi.eu/about/egi-inspire/EGI-InSPIRE_partners.html
[42] http://www.egi.eu/case-studies/
[43] http://www.egi.eu/community/collaborations/
[44] https://wiki.egi.eu/wiki/Middleware

SEVENTH FRAMEWORK PROGRAMME

- **Daily Operations of the infrastructure**: the EGI infrastructure comprises more than 350 resource centres. Operations is a community effort involving Resource Centres, National Grid Initiatives, and EGI.eu for central coordination.
- **Support to user and resource centre administrations**.
- **Community outreach**.

Key use-cases which the RI are focusing attention on:

- **Functional services**: including user authentication and authorization, information discovery, storage management, data and metadata catalogues, file transfer, sequential and parallel computing and workflow management.
- **Operational services**: including monitoring, maintaining the operations portal and the infrastructure topology repository, and accounting.
- **Federated** cloud provisioning of IT services and resources (currently in its pilot phase).

EGI tries to serve each constituent with particular kinds of services that can simultaneously be both generic and customized. This requires constant evolution of the deployed platforms whilst sharing a common set of generic operational and infrastructural services such as user authentication and authorization, information discovery and accounting.

9.2 Data

With regards to data, EGI is discipline agnostic and supports the modality of choice of the user community. Currently most of the data held within the EGI infrastructure is accessed in file format. Different access interfaces are supported: POSIX, WebDAV, WS Interface and SRM (Storage Resource Management interface).

9.2.1 Data acquisition

Data is collected from data sources (sensors, instruments, archives, *etc*.) and is persistently or temporarily stored in Storage Resource Management systems. Persistency is a choice of the data owner. Data can then be located using logical and physical file names (defining a human-readable name) and the physical location of the file (or its replicas). To handle this EGI offers a file location service which provides:

- access to distributed files and distributed file replicas via physical location;
- mapping of human-readable file names to storage URLs;
- organization of files into a hierarchical name space;
- definition of user-metadata for each directory entry;
- implementation in C with a high-performance multi-threaded frontend, integrated with grid user authentication/authorization.

These specifications are supported by the Logical File Catalogue (LFC) released by the EMI project, and by the Globus Replica Location Service (RLS) as supported by IGE.

9.2.2 Data storage

Data is stored via file transfer according to the following specifications:

- files are stored disk-only or using combined disk and tape storage systems;
- there are multiple disk server nodes;
- different space types or multiple file replicas are used;
- relevant standards (different implementations, different standard sets) are applied;
- access is via POSIX or GridFTP protocols;
- OGF Storage Resource Management interface (SRM) v2.2 is used to provide a uniform interface to heterogeneous storage systems;
- for web access, web service interface and HTTP-based WebDAV standards are used;
- the backend system uses NFS 4.1, General Parallel File System (GPFS) and Lustre.

Various technical implementations exist: dCache (Java), DPM, StoRM (Java) and HYDRA for storage of encrypted files (all released by the EMI project). These support different back-ends such as MySQL and ORACLE.

The Storage Resource Management interface exposed by heterogeneous storage management systems allows to perform data cleaning and maintenance (allocation of storage space, creation/relocation/removal of files, blocks of files and storage areas). Location of files and annotation of files is implemented with data distributed catalogue solutions:

- key/value pairs describe research data in a human readable fashion;
- users can annotate and search files based on their descriptions;
- hierarchical metadata schemas describe similar data objects.

9.2.3 Metadata

EGI offers services for metadata management in the form of AMGA (Distributed Grid Metadata Service), an EMI project, which supports:

- replication (full/partial) to implement a federation of catalogues with high availability and performance;
- WS SOAP interface;
- client APIs for Java, Python, Perl, PHP and C++;
- access control through ACLs.

9.2.4 Data transfer

A single service is deployed to provide a solution to this supporting:

- transfer of files across distributed storage systems;
- transfer of a single file or of block of files across a mesh of storage-to-storage network paths;
- control of a number of simultaneous transfers, of a number of streams, as well as transfer across multiple groups of users;
- network bandwidth tuning.

Relevant standards: Storage Resource Management interface (SRM V2.2) and GridFTP. These specifications are implemented by the File Transfer Service (FTS) – an EMI project.

9.3 Infrastructure

EGI is a federation of national and domain specific resource providers. EGI enables researchers within Virtual Research Communities (VRCs) to collaborate, communicate and share resources across international boundaries by offering benefits such as the integration of community resources into EGI through user support and assistance, training and technology specialists and representation.

9.3.1 Architecture

EGI technical services are lead by the principles of openness and inclusiveness; in addition to a number of functional services such as user authentication and authorization, storage management, file access transfer and cataloguing, computing, and compute workload management, user-specific services can be integrated into a seamless operations environment where services are monitored and accounted for use.

Resources and IT services are distributed across Resource Centres. These are federated nationally within national infrastructures that in turn are federated at a European level. Core resources are computing (270,000 CPU cores), data (130 PB on disk and 130 PB on tape), and storage resources, the networking infrastructure being used as communication channels.

Any external resource can potentially be integrated. Access is managed so that only authenticated and authorized users can access them. Authorization normally depends on the affiliation of a user to one or more research groups. The following specifications detail how user authentication and authorization work across the whole infrastructure.

9.3.2 Security

Security is very important. User authentication and authorization (see the authentication and authorization sections below for specifications of these two functions) are used to securely access resources and authorize users in accordance with their affiliation to Virtual Research Communities. Data can be encrypted in case of data privacy issues.

EGI has an operational structure for incident response in case of security incidents, revocation of user credentials in case of stolen user identities and procedures for suspensions in case critical vulnerabilities are found in production. The software deployed is also assessed, and has security monitoring tools that assist daily security operations.

9.3.3 Connectivity

Available bandwidth between resources depends on the resource centres hosting resources and their data movement requirements. Typically it ranges from a few tens of Mb/s to 10 Gb/s. Connectivity is offered by the National Research Networks and GEANT at a European level.

For data access it is SRM which provides a homogeneous interface on top of heterogeneous storage management systems.

SEVENTH FRAMEWORK PROGRAMME

9.4 Computation

Models, algorithms, *etc*. are chosen by each Virtual Research Community independently. EGI supports both serial and parallel computing, and supports user workflow management systems embedded in user portals, that can use a harmonized interface (under standardization at OGF) on top of different computing infrastructures.

EGI is also investigating the support – in collaboration with the PRACE supercomputing infrastructure – of tightly coupled and loosely coupled multi-scale simulations. The use cases supported are in the fields of fusion, hydrology, physiology, nano-science materials and computational biology[45].

Access to compute resources for execution of user compute workflows is provided through a standard (or standard *de facto*) interface that provides access to heterogeneous batch system compute systems. Compute Job Execution is supported by EGI to:

- Execute an authorized compute job, serial or parallel, in an arbitrary number of computing batch systems – exposing a harmonized interface which hides the heterogeneity of the batch system backend.
- Access computing farms as authorized depending on the identify of the submitting user.
- Share resources across multiple user communities.

Different batch systems are supported depending on the type of CE deployed, *e.g.* LSF, Torque, GE or SLRUM. Various implementations of the compute job execution exist: CREAM (C++, AXIS WS interface), ARC-CE, UNICORE (EMI Project) and GLOBUS GRAM5 (IGE Project).

Different user workflow engine systems (such as Taverna) can be deployed on EGI assuming that they can interact with the job execution interfaces of EGI. EGI currently has more than 350 centres in production exposing a job execution interface.

The job execution workload of the user can be distributed across multiple compute centres through a workload management system which:

- Distributes jobs across multiple compute centres according to the load and status of local batch systems.
- Redirects workload in case of failure or too high queuing time.
- Supports opportunistic usage of other user community resources when idle.

9.5 Working practices

Given that EGI is a highly distributed infrastructure, interworking relies on a set of common operational procedures and policies, including policies for secure access, for correct handling

[45] http://www.mapper-project.eu/web/guest/applications

of Intellectual Property rights, resource usage policies, *etc*. The whole set of EGI policies and procedures can be consulted on-line [46].

9.5.1 Security

EGI security concerns include: provision of single sign on, user authentication, federated identity support for seamless access to services, acceptance of Acceptable Usage Policies by users, provisioning of a secure IT infrastructure to the users, confidentiality and integrity of data, presence of a Computer Security Response Team capable of handling incidents, security risk assessment, assessment of software security, and vulnerability assessment of deployed software.

9.5.2 Legality

Available are escalation procedures in case of controversy between EGI-InSPIRE project partners.

9.5.3 Service

Resource Centres are requested to provide a minimum level of service availability. Services provided by Resource Centres have to comply to a basic set of service level targets. These are defined in the EGI Resource Centre Operational Level Agreement[47], agreed between a Resource Centre operations manager and the respective national infrastructure representative. Services that are centrally provisioned by a national Resource infrastructure Provider have to comply to a set of service level targets defined in the EGI Resource Infrastructure Provider Operational Level Agreement[48]. Central EGI services (shared among all user and operations communities) also have to comply to quality parameters. These will be defined a EGI.eu Operational Level Agreement (which at the time of writing is still being drafted).

9.5.4 Ethicality

Access to user groups from nations that do not comply to UN treaties can be banned in some countries. This is defined by national policies as applicable.

9.5.5 Authentication

Access to a service is based on user authentication. Federated identity provisioning guarantees that user identity is seamlessly understood by services. Identity is provided through X.509 certificates. Mapping of existing user credentials (username/password) into short-lived X.509 certificates is also possible. Features include:

[46] http://www.egi.eu/policy/policies_procedures.html
[47] https://documents.egi.eu/document/31
[48] https://documents.egi.eu/document/463

- Users grouped into Virtual Organizations (VOs).
- Release of security attributes based on VOMS (VO Membership Service) and UVOS (UNICORE).
- Attribute Authority server to obtain signed security credentials with attributes of end-users (*e.g.* role possession, group/project membership) used during authorization
- User identities are transparently mapped to local identities/accounts (provided that authorization is granted).

Relevant standards: Security Assertion Markup Language (SAML) 2.0, SOAP-based Web service Interfaces and REST (Representational State Transfer) interface. The VOMS service is released by the European Middleware Initiative (EMI) project. There are currently 64 instances deployed in production, operated to provide 99% availability and reliability.

9.5.6 Authorization

Access to services in distributed infrastructures rests on user authorization; the user is authorized depending on membership of participating research infrastructures. EGI uses the ARGUS service (developed by the EMI project) to derive user authorization decisions across multiple services (computing, data, portals, *etc*.). ARGUS consists of a:

- Policy Enforcement Point (PEP) and Policy Decision Point (PDP);
- Policy Administration Point (PAP).

Relevant standards: Extensible Access Control Markup language (XCAML) standard via SOAP Web service interface. C/C++ and Java applications can use the PEP client APIs to request an authorization from a remote ARGUS server acting as a PEP server. ARGUS internally forwards this request encoded in XACML to the PDP to evaluate the request. The ARGUS PDP retrieves policies (in XACML format) from the PAP and evaluates them. This service is operated to provide 99% availability and reliability.

9.5.7 Accounting

A fully distributed accounting infrastructure is in production to account usage of compute resources. This infrastructure will be extended to account for storage resource usage. The accounting infrastructure is extensible; accounting of new services can be easily supported once the related usage record schema is agreed between the partners, and possibly standardized. Accounting records can be published directly by services into a central accounting DB, or can be collected within a given domain of the infrastructure, and be published centrally in a summarized form. This means that an arbitrary hierarchical tree of accounting infrastructures with a central collection point is possible. Specifications:

- Relevant standards: OGF Usage Record schema.
- Messaging: StoMP interface (Streaming Text-Oriented Message Protocol) JMS 1.1.
- Usage Records collected locally and published through messaging.
- Central DB consuming Usage Records from the message broker network.
- Encryption of user information when publishing accounting data. Different accounting information views are supported - depending on the identity of the consumer, to ensure privacy of data is enforced.

10 EUDAT

The European Data Infrastructure (EUDAT; http://www.eudat.eu).

EUDAT's goal is to deliver a cost-efficient and high quality Collaborative Data Infrastructure (CDI) which can meet researchers' needs in a flexible and sustainable way, across geographical and disciplinary boundaries.

The EUDAT consortium comprises 25 European partners, including data centres, technology providers, research communities and funding agencies from 13 countries. It includes key representatives from research communities in linguistics (CLARIN), earth sciences (EPOS), climate sciences (ENES), environmental sciences (LIFEWATCH), and biological and medical sciences (VPH), all of which have been allocated project resources to help specify their requirements and co-design related services.

The project started in October 2011, and is currently carrying out a comprehensive review of research communities' current approaches to data infrastructure and their resulting requirements. EUDAT is investigating and designing the appropriate services and technologies to match these requirements, to be operated as part of its distributed infrastructure.

10.1 Activity

To build a sustainable data infrastructure upon which common services can be deployed for use by diverse communities, a comprehensive approach is required, including several activity strands:

- EUDAT is currently investigating user requirements, starting with its five core research communities. This investigation will be extended to additional communities in 2012.
- A second activity strand concerns the appraisal of technologies and service candidates, which involves identifying, designing and constructing appropriate services, using existing solutions where possible. This will also help in identifying the gaps that should be addressed by EUDAT.
- The third activity strand involves primarily the data centres and deals with the operation of the collaborative infrastructure, particularly the provisioning of secure, reliable (generic) services in a production environment, with interfaces for cross-site and cross-community operation. The operation of the infrastructure should provide full life cycle data management services, ensuring the authenticity, integrity, retention and preservation of data, especially those marked for long-term archiving.
- Other activities focus on sustainability and funding models, dissemination, outreach and training.

EUDAT has shortlisted six service cases identified by user communities as priorities: safe replication of data from site to site; data staging to (HPC) compute facilities; easy storage; metadata; AAI (Authentication and Authorisation Infrastructure); and PIDs (Persistent Identifiers). In each case, a multi-disciplinary taskforce involving representatives from

communities and data centres has been set up to plan for the deployment of these services on the EUDAT infrastructure.

10.1.1 Community

The key constituents of the EUDAT user community are the participating research communities, which collectively represent a wide array of research interests. EUDAT also targets disciplines and communities beyond this initial set of communities. For this purpose, User Forums are held on a regular basis, and are open to all communities interested in adapting their solutions or contributing to the design of the CDI.

The idea is to serve each constituent with particular kinds of services that can simultaneously be both generic and customized. This requires the construction of federated data grids capable of encompassing all of these generic and customised services.

Archetypal users cover usage patterns which are taken as a basis for the type services that can be provided. The principal security concern for EUDAT is the confidentiality, integrity and availability of data. Legally speaking, the EUDAT project complies with privacy and security related with EU directives and national laws and regulations of the participating sites.

It is expected that the first EUDAT services will be available in 2012.

10.2 Data

All kinds of research data that adhere to given data policies are to be collected. An aggregated metadata model is being defined which allows both the definition of universal data attributes and the definition of domain-specific attributes. Data collection is by replication of existing data collections outwith the EUDAT ecosystem. The replicated data is stored within data grids (iRODS is currently being considered to manage this).

10.2.1 Metadata

A three-level metadata model is being proposed:

- A simple ‘flat’ metadata standard for *discovery*; flat metadata means it is a single record with attributes rather than a group of linked records each with attributes and with relationships between the records.
- A structured (linked entity) standard for *context* (relating the dataset to provenance, purpose, environment in which generated *etc*.).
- Detailed metadata standards for each kind of data to be co-processed.

The following standards may be appropriate to support such a model:

- **Discovery**: Dublin Core.

- **Contextual**: CERIF (Common European Research Information Format) or ISO 19115.
- **Detailed**: Individual standards depending on type of dataset; for research datasets from large-scale facilities CSMD[49] and PaNData[50], for geospatial datasets INSPIRE 1[51].

For the Metadata collection and delivery EUDAT is proposing the adoption of the OAI-PMH protocol[52]. Once the metadata structure is defined, searching capabilities will be implemented adopting standard interfaces, *e.g.* OGC web services.

10.3 Infrastructure

The EUDAT infrastructure is comprised of the resources provided by each participating site in accordance with a common plan and set of schemata. Core resources include networks, servers, software, maintenance and help desks.

EUDAT favours open source software and aims for platform independency due to the multiple type of technologies in use at the participating sites.

10.3.1 Safe data replication

Relevant data needs to be replicated for bit-stream preservation, optimal data curation and accessibility. Managers of data centres must be able to state that *X* replications are required for *Y* years, with *Z* centres excluded from the replication scheme. All replications of the same dataset should be identified via a single PID --- therefore all locations of a given dataset should be associated with a given PID. Other requirements:

- It must be possible to tell whether or not replications are identical to the original source. Regular checks are necessary.
- Replicas should be exactly as accessible as their sources.
- All centres are being audited, so all activities need to be explicitly specified by policies.
- Transmission of replicas must be secure.
- Access logging will be performed and aggregate statistics produced.

iRODS is being considered as a starting point. Handle Services are to be used for giving identity to data objects. Federated AAI is required; Shibboleth is currently favoured. A central registry describing all participating data centres is also required.

[49] http://www.ijdc.net/index.php/ijdc/article/view/149
[50] http://www.pan-data.eu/PANDATA_-_Photon_and_Neutron_Data_Infrastructure
[51] http://inspire.jrc.ec.europa.eu/
[52] http://www.openarchives.org/OAI/openarchivesprotocol.html

A key requirement is that data needs to be replicated from user community centres to a number of trusted data centres and vice versa. Replicated data objects are identified by a single PID. It has not been decided whether PID records can be updated by replication sites, or only be the PID administrator. Safe replication must ensure bit-stream preservation; desirable also is optimal data curation and accessibility. Replication services must be available through a single interface. Technologies for long-term archives will likely be institution-specific, with policy-based replication handled by iRODS and persistent identifiers produced using the EPIC/Handle system.

Safe replication will run in parallel with authentication and authorisation. A central centre registry would seem to be required.

10.3.2 Dynamic data replication

Dynamic replication is concerned with the replication of data between storage resources (as might be used by EUDAT) and HPC staging areas. This form of replication is driven by community users as they need the data for their experiments. Users can specify which data centres to use, and how long data should be kept close to a HPC system.

- If the HPC system and data collection are found on the same EUDAT node, then low level data transfer (`cp`, `scp`, `rsync`) can be used.
- If the HPC system is not on a EUDAT node (*e.g.* PRACE or EGI), then remote data transfer tools must be used (such as PRACE data services).
- If the data collection is not on a EUDAT node, but the HPC system is, then likewise remote data transfer tools must be used (such as EUDAT safe replication tools).
- This same scenario occurs if both data collection and HPC system are outwith EUDAT -- but in this case, EPOS cannot use any in-place EUDAT services.

Possible technologies for selecting data collections include iRODS clients, Cyberduck (`http`), GridFTP clients, Globus Online, Webdav clients and UNICORE FTP. Currently PRACE does not recommend use of Globus Online for data security reasons, nor does it permit `http` access to its storage areas. UNICORE FTP is under evaluation at PRACE. EUDAT have formed a number of 'test islands' in order to test different services / tools.

11 D4SCIENCE

D4Science.org is an infrastructure resulting from the efforts and contributions of a series of European Projects: DILIGENT (FP6-2003-IST-2), D4Science (FP7-INFRA-2007-1.2.2), D4Science-II (FP7-INFRA-2008-1.2.2), iMarine (FP7-INFRASTRUCTURES-2011-2), and EUBrazilOpenBio (FP7-ICT-2011-EU-Brazil).

The infrastructure realises the notion of a Hybrid Data Infrastructure and promotes the development of Virtual Research Environments:

- A Hybrid Data Infrastructure (HDI) is an innovative approach that promotes the integration of a number of technologies and resources to realise a flexible and elastic environment capable of dealing with the evolving requirements arising in data-intensive science for data and data management. Such an infrastructure is expected to be equipped with a number of "mediator" services for interfacing with existing data sources and repositories by relying, where possible, on standards. The goal is to enable a data-management-capability delivery model in which computing, storage, data and software are made available by the infrastructure *as-a-Service*.
- A Virtual Research Environment is an application that might be built by relying on an HDI and which is requested to have the following distinguishing features: *(i)* it is a Web-based working environment; *(ii)* it is tailored to serve the needs of a Community of Practice; *(iii)* it is expected to provide a community of practice with the whole array of commodities needed to accomplish the community's goal(s); *(iv)* it is open and flexible with respect to the overall service offering and lifetime; and *(v)* it promotes fine-grained controlled sharing of both intermediate and final research results by guaranteeing ownership, provenance, and attribution.

The D4Science underlying technology gCube[53] is a software system designed and implemented to enable the building and operation of such a kind of infrastructure. It can be reuse by any community of practice to deploy and operate an infrastructure with such characteristics.

11.1 Activity

D4Science is not tailored to serve the needs of a specific Community of Practice. It is currently serving two very large communities in the marine living resource management and biodiversity domain respectively.

By using such an infrastructure Communities of Practice can be supported while performing the activities including:

- Community-of-Practice-specific collaboration activity by building and operating a Virtual Research Environment serving end users;

[53] http://www.gcube-system.org/

- Resources registration, monitoring and management for a large class of resource typologies including, computational and storage resources, data sources and archives, services;
- Resources discovery and sharing across Community of Practice boundaries;
- Large-scale data processing tasks by relying on computational resources and models including, Grid, private and public Clouds;
- Exploitation of a rich array of off-the-shelf services offering both context-agnostic (*e.g.* efficient and scalable storage of files, support for scalable and uniform access to tree-based structured data) and context-specific (*e.g.* efficient and uniform access to biodiversity data stored in heterogeneous repositories) facilities;
- Users registration and management.

11.2 Data

D4Science is an infrastructure that has no proprietary data, *i.e.* the data it may give access to are actually part of the infrastructure because (a) they are dynamically acquired from recognised repositories and archives via dedicated mediator services and/or (b) the members of a Community of Practice have outsourced their management (*e.g.* by storing them on storage facilities offered by the infrastructure) to the infrastructure.

All the datasets that are somehow a resource of the infrastructure are characterised by a rich set of metadata and policies characterising its usage.

11.3 Infrastructure

The D4Science infrastructure[54] consists of a number of resources including hosting nodes, computing and storage resources, software artefacts, datasets and running instances of services. Such resources can be either (a) physically-born in the context of D4Science.org, *i.e.* they have been deployed in the context of this infrastructure, or (b) virtual entities, *i.e.* they pre-exist the infrastructure and a placeholder representing them is actually part of the infrastructure.

A living picture of the set of resources forming the infrastructure can be acquired by using a dedicated service at `http://www.d4science.org/web/guest/monitor`.

[54] `www.d4science.org`

12 DISCUSSION

Based on discussion with the six ESFRI research infrastructures, a number of common issues which affect most if not all RIs have been identified. The following problems were identified during a meeting in Vienna in April 2012:

- Federation over existing infrastructures or services.
- Quality control of data.
- Data staging or moving computation to the data.
- Computation requiring High Performance Computing (HPC).
- Metadata definition and assignment.
- Persistent identifiers mechanism.
- (Near) real-time data handling.
- Provenance, preservation and reproducibility.
- Attribution of data to data source; crediting the originator.
- Archiving versus regeneration of data and results on demand.
- Delegation of various parts of the infrastructure to multiple projects.
- Providing single sign-on to integrated services; delegated authorisation.
- Constructing and operating implementations of complex models.
- Integrated data discovery across various data centres and catalogues.

The priorities of different RIs vary, but of the above problem areas, the following have been designated as priorities for ENVRI at large:

- **Federation over existing infrastructures or services**: entailing the challenge of bringing together existing infrastructure components and services as contributions to the construction of an RI, as well as the challenge of bringing about interoperability between separately owned and operated facilities that each contribute to the RI.
- **(Near) real-time data handling**: entailing the challenge of being able to:
    - collect, store and catalogue data as it arrives in real-time from sensors;
    - process data into derived data products in real-time;
    - analyse data in real-time.
- **Integrated data discovery across various data centres and catalogues**: entailing the challenge of being able to easily discover data which are heterogeneous (in format, content and metadata) and which are stored at different places.

As such, these areas will likely receive the greatest initial attention by work packages 3 and 4 in the next year, subject to the availability of solutions and the judgement of expert personnel acting in the various incoming project tasks.

12.1 ENVRI Reference Model

One of the overarching goals for ENVRI is the development of the ENVRI Reference Model (RM-ENVRI), based on International Standards for Open Distributed Processing (ODP). The creation of this reference model is an exercise in trying to describe the archetypical research infrastructure and its interactions with external services. In doing so, it is hoped that new insights can be obtained into the construction of real research infrastructures.

SEVENTH FRAMEWORK PROGRAMME

The information gathered on the ESFRI RIs allows us to make some initial observations as to how the archetypical RI might decompose with regard to the five ODP viewpoints: enterprise, information, computation, engineering and technology. It should be noted that each viewpoint ultimately draws on all aspects of a system insomuch as each viewpoint describes it entirely from a given perspective, rather than each viewpoint being some arbitrary partition of that system. With that in mind, we make certain observations which feed into Task 3.3 of the ENVRI project (Common Reference Model for the environmental infrastructures).

12.1.1 Enterprise viewpoint

The enterprise viewpoint is concerned with the roles present in a system and the capabilities and responsibilities inherent in those roles, as well as any generic (non-computational) policies and constraints on the system as a whole.

The roles and communities extant in the enterprise viewpoints of RIs should be closely aligned across RIs. This is because whilst each RI primarily offers its services to a particular scientific community, the non-domain specific aspects of these communities should generally be common to all of them.

In particular, the classes of user of interest to an RI are the same. As well as the core expert scientist community, there is the 'citizen scientist' community, as well as journalists and policy makers, and the engineers which contribute code and resources to an RI. In addition, there is one 'user community' which will become particularly vital to the ESFRI RI. This is the community of other services and infrastructures which consume and contribute the RI's data products.

There is an increasing need for integration of data from many RIs, and an increasing reliance between RIs to achieve their more advanced goals. Thus, how external resources interact with an RI must be understood as an important use-case as well as a technical question, which concerns the enterprise viewpoint of ODP.

12.1.2 Information viewpoint

The information viewpoint is concerned with the data objects used within a system, which includes all metadata and process information necessary for the smooth running of that system.

The clearest requirement for any information model is compatibility with the myriad (meta)data standards being applied by the various research infrastructures. These things need to be at the very least acknowledged by any model, though it is less clear whether to provide an abstraction over these services or to treat them individually as classes of data object.

One question is how to represent continuous data products, particular those which are subsampled by other processes. Do these things represent an information object, or a computational object from which information objects are extracted? Questions of how to assign persistent identifiers to objects also needs addressing. Aside from the actual process of assignment, is there need for a global service for assigning identifiers, or can separate RIs

assign their own, with additional foreign identifiers added to metadata of a data object if transferred into a different context (RI)?

12.1.3 Computational viewpoint

The computational viewpoint is concerned with the functions and processes engaged in by a system, and with the communication links between computational objects.

Each RI has its own set of high-level computations which are specific to its particular data products. However, there are also lower-level computations based on certain standard data management tasks (assigning of persistent identifiers, provenance recording, data transport, discovery, handling access and authorisation, *etc.*) which should be present in all RIs, and may be generically definable in a manner applicable to all RIs despite differences in underlying engineering.

For some of the RIs which are more focused on basic data provision rather than data integration or computational services, there is little consideration for providing a platform for workflow composition and enactment. Nonetheless, some consideration should be given as to how the RI presents its data such that it can be staged for computation, even if actually performed externally to the RI.

12.1.4 Engineering viewpoint

The engineering viewpoint is concerned with the topology of an infrastructure and the interfaces between different physical nodes of a system.

On the one hand, the engineering viewpoint is the most distinctive for each individual RI, being dependent on the composition of their various sensor networks and the distribution of their primary data collection sites. On the other hand, the engineering viewpoint may prove fruitful for ENVRI to consider across RIs by considering the interfaces between RIs (and thus the ability to propagate and integrate data products).

12.1.5 Technology viewpoint

The technology viewpoint concerns standards and specific tools and programs. Given this, it is a (relatively) easy viewpoint from which to identify commonalities between RIs, since the use of the same or equivalent technologies for equivalent aspects of different RIs can be easily noted.

It is also quite simple, given the other viewpoints, to identify lack of technology. This may present an opportunity for ENVRI to identify the requirements and recommend a technology (or even provide a prototype technology if sufficiently common and aligned to the expertise of the ENVRI consortium).