D5.4
A Development Plan for Common Operations and Cross-Cutting Services based on a Network of Data Managers and Developers

WORK PACKAGE 5 – REFERENCE MODEL GUIDED RI DESIGN

LEADING BENEFICIARY: NERC

Author(s):	Beneficiary/Institution
Keith G Jeffery	NERC
Malcolm Atkinson	University of Edinburgh
Zhiming Zhao	UvA
Yin Chen	EGI
Abraham Nieva de la Hidalga	Cardiff University
Alex Hardisty	Cardiff University
Yannick Legre	EGI
Leonardo Candela	CNR

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654182

Daniele Bailo	INGV
Thomas Loubrieu	IFREMER
Barbara Magagna	EAA

Accepted by: Paola Grosso (UvA) (WP 5 leader)

Deliverable type: [REPORT]

Dissemination level: PUBLIC

Deliverable due date: 31.10.2016/M18

Actual Date of Submission: 31.01.2017 (M21)

ABSTRACT

ENVRIplus is – by its very nature – a heterogeneous distributed network of Research Infrastructures (RI) for providing the advanced supporting environments for environmental scientists. Thus, a key feature of any recommended conceptual architecture for RIs – for their own beneficial utilisation and also for RI interoperation – requires the recommendation of common operations and cross-cutting services to allow the researchers to perform their work effectively and efficiently and to allow access to RIs other than the one to which they are usually attached in order to encourage – where appropriate – multidisciplinary research. Identification of computational objects in the RM (Reference Model) of the ENVRI project provides a basis; the purpose of WP5 in ENVRIplus is 'providing a novel ENVRIPLUS Reference Model which should be developed not only based on the existing ENVRI RM but should also include the latest development insights from other successful RIs'[1]. Thus, a re-examination of the requirements from D5.1 within ENVRIplus [Atkinson et al. 2016] is the start point, wherever possible, for a proper matching with the (developing) ENVRI RM.

These common aspects emerge from two directions: (1) the state of the art, which provides opportunities for utilisation in ENVRIplus and (2) the requirements, which provide the specifications of the services and operations needed by the users of the ENVRIplus RIs.

The common aspects form a key basis to achieve the distributed, interoperating architecture recommended for ENVRIplus providing the RIs with an evolutionary direction for the individual RIs to adopt best practice and for them to become interoperable.

The development plan provides a stepwise approach to achieve the architecture recommended for ENVRIplus.

Project internal reviewer(s):

Project internal reviewer(s):	**Beneficiary/Institution**
Antti Pursula	CSC
Robert Huber	University of Bremen

Document history:

Date	**Version**
12.10.2016	Outline for comments
24.10.2016	Version to WP5 for comment (2 comments received)
11.11.2016	To internal reviewers for comment to be refined during

[1] ENVRIplus DoW (Description of Work) p31

	ENVRI week
16.11.2016	Version to colleagues who commented during ENVRI week with responses to comments
17.11.2016	New version initiated with different emphasis to accommodate changes suggested by 3 organisations
09.12.2016	New version for internal approval
22.12.2016	New version after internal discussions to representatives of other WPs dependent on this deliverable
10.01.2017	Revised version to internal reviewers
15.01.2017	Revisions following second internal review and additional comments from WP5 colleagues, from one organisation, alignment with the recently produced D5.2 and discussions with Theme 2 leader
17.01.2017	Further check and corrections by Malcolm Atkinson
18.01.2017	Incorporation of comments from Alex Hardisty
28.01.2017	Final agreed version submitted

DOCUMENT AMENDMENT PROCEDURE

Amendments, comments and suggestions should be sent to the authors (Keith Jeffery keith.jeffery@keithjefferyconsultant.co.uk)

TERMINOLOGY

A complete project glossary is provided online here:
https://envriplus.manageprojects.com/s/text-documents/LFCMXHHCwS5hh

PROJECT SUMMARY

ENVRIplus is a Horizon 2020 project bringing together Environmental and Earth System Research Infrastructures, projects and networks together with technical specialist partners to create a more coherent, interdisciplinary and interoperable cluster of Environmental Research Infrastructures across Europe. It is driven by three overarching goals: 1) promoting cross-fertilisation between infrastructures, 2) implementing innovative concepts and devices across RIs, and 3) facilitating research and innovation in the field of environment for an increasing number of users outside the RIs.

ENVRIplus aligns its activities to a core strategic plan where sharing multi-disciplinary expertise will be most effective. The project aims to improve Earth observation monitoring systems and

strategies, including actions to improve harmonisation and innovation, and generate common solutions to many shared information technology and data related challenges. It also seeks to harmonise policies for access and provide strategies for knowledge transfer amongst RIs. ENVRIplus develops guidelines to enhance transdisciplinary use of data and data-products supported by applied use-cases involving RIs from different domains. The project coordinates actions to improve communication and cooperation, addressing Environmental RIs at all levels, from management to end-users, implementing RI-staff exchange programs, generating material for RI personnel, and proposing common strategic developments and actions for enhancing services to users and evaluating the socio-economic impacts.

ENVRIplus is expected to facilitate structuration and improve quality of services offered both within single RIs and at the pan-RI level. It promotes efficient and multi-disciplinary research offering new opportunities to users, new tools to RI managers and new communication strategies for environmental RI communities. The resulting solutions, services and other project outcomes are made available to all environmental RI initiatives, thus contributing to the development of a coherent European RI ecosystem.

TABLE OF CONTENTS

Deleted:

1 INTRODUCTION

1.1 Setting the Scene

In environmental and earth sciences, data-centric approaches play an increasing role. To study the development of earthquakes or volcanoes for example, one needs continuous observation of the surrounding geographic regions and their underlying strata in order to obtain the data necessary to model various seismological processes and their interactions. Depending on the problem scale and geographical focus, these observations can only be provided by sources distributed across different countries, institutions and data centres. Moreover, such research activities also often require advanced computing and storage infrastructure in order to analyse, process, model and simulate the data. Advanced infrastructural environments to support research (*research support environments)* are clearly needed to enable researchers to access data, software tools and services from different sources, and to integrate them into cohesive experimental investigations with well-defined, replicable workflows for processing data and tracking the provenance of results.

In a recent publication [Zhao et. al. 2016] identified several kinds of support environments that must work together to support data-centric research: 1) computing, storage and network infrastructures, e.g., provided via [EGI], [EUDAT] and [GEANT], also called *e-Infrastructures (e-Is)*; 2) services for accessing, searching and processing research data within different scientific domains, called *Research Infrastructures* (RIs), e.g., [ICOS], [EPOS] and [EURO-ARGO] for the atmospheric, earth and marine sciences; and 3) environments for providing user-centred support for discovering and selecting data and software services from different sources, and composing and executing application workflows based on them, called *Virtual Research Environments* (VREs) [JISC 2010], *Virtual Laboratories (VLs)* [Belloum et al. 2011] or *Science Gateways* (SGs) [Miller et al. 2011]. An early example is myExperiment [De Roure and Goble 2007] and later D4Science [Candela et al. 2014] and the current trend to general VREs/VLs/SGs is exemplified by [VRE4EIC]. In many cases, these different types of supporting environments often overlap with each other, as shown in Fig. 1. A VRE can be deployed on private infrastructures, public clouds, or e-Infrastructures as services; it can be operated based on its own information catalogs or the resources catalogs provided by research infrastructures. In some cases, research communities can also directly conduct experiments based on resources provided by RIs. A virtual laboratory for ecology is described in [Hardisty et al 2016]

In the context of Theme 2 of ENVRIplus, we specifically focus on the three types of research support environment identified earlier: VREs (including the concepts of VLs and SGs), RIs (and their electronic representation as e-RIs) and e-Infrastructures (e-Is). Based on the specific foci of those different environments, an abstract logical relationship among them can be seen in the layers of Figure 1. This is a rough picture of the landscape: different kinds of research supporting environments and the role they play in ICT activities initiated by user communities. In some contexts, e.g., VRE4EIC, the digital representation of a RI is also called an e-RI. This is to distinguish the RI (the organisation, equipment, assets) from its digital representation (e-RI).

Deleted: F

ENVRIplus is – by its very nature – a heterogeneous distributed network of Research Infrastructures (RIs) for providing the advanced supporting environments for environmental scientists. Thus, a key feature of any recommended conceptual architecture for RIs – for their own beneficial utilisation and also for RI interoperation – requires the recommendation of common operations and cross-cutting services to allow the researchers to perform their work effectively and efficiently and to allow access to RIs other than the one to which they are usually

attached in order to encourage – where appropriate – multidisciplinary research. Identification of computational objects in the RM (Reference Model) of the ENVRI project provides a basis; the purpose of WP5 in ENVRIplus is in 'providing a novel ENVRIPLUS Reference Model which should be developed not only based on the existing ENVRI RM but should also include the latest development insights from other successful RIs'. Thus, a re-examination of the requirements from D5.1 developed with the RIs within ENVRIplus is the start point, wherever possible, for a proper matching with the (developing) ENVRI RM.

These common aspects emerge from two directions: (1) the state of the art which provides opportunities for utilisation in ENVRIplus and (2) the requirements which provide the specifications of the services and operations needed by the users of the ENVRIplus RIs.

The common aspects form a key basis for the route to a distributed, interoperating architecture recommended for ENVRIplus providing the RIs with an evolutionary direction for the individual RIs to adopt best practice and for them to become interoperable.

The development plan provides a co-design (with the RIs) stepwise approach to achieve the architecture recommended for ENVRIplus.

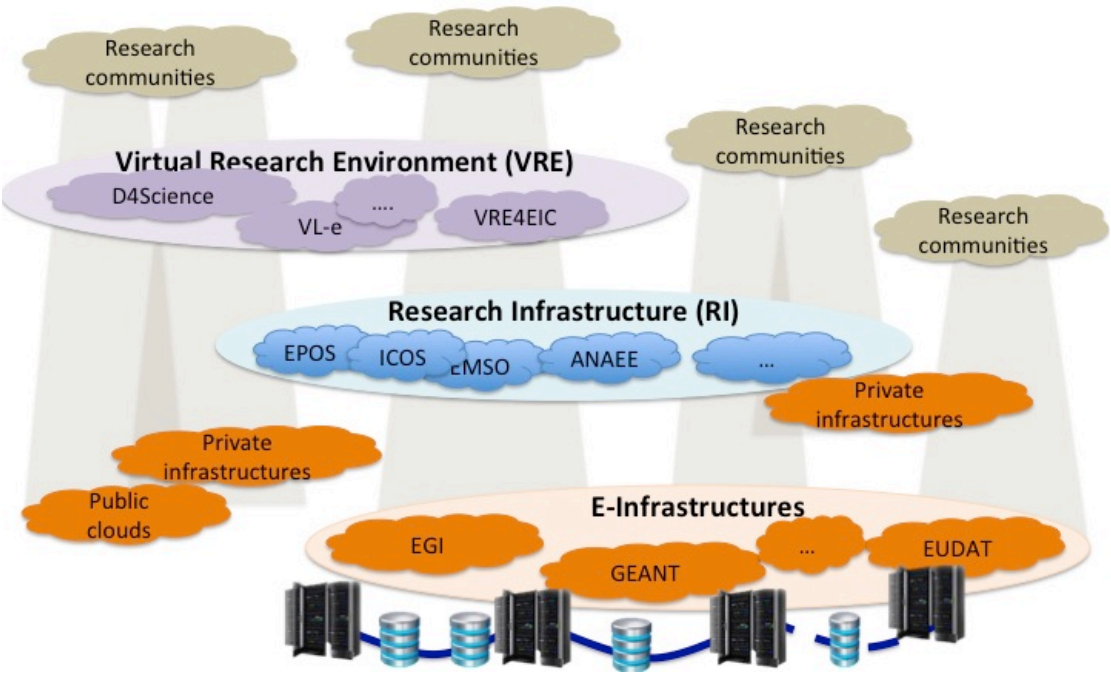

FIGURE 1: A VIEW OF THE DIFFERENT KINDS OF RESEARCH SUPPORT ENVIRONMENTS AND THE ROLE THEY PLAY IN ICT ACTIVITIES INITIATED BY USER COMMUNITIES.

1.2 Method

This deliverable relies on input from D5.1 defining the relevant state of the art and the requirements of ENVRI Plus RIs [Atkinson et al 2016]. However, it also relies on discussions with

colleagues in all WPs within Theme 2 (WP5 Reference model guided RI design, WP6 Inter RI data identification and citation services, WP7 Data processing and analysis, WP8 Data curation and cataloguing, WP9 Service validation and deployment) and a re-examination of the ENVRI RM for already defined operations that can be used in ENVRIplus. In parallel with the work for this deliverable, work on extending the RM in the light of D5.1 is produced as D5.2 and is related to this work in the tables (Sections 2.4 and 3.5). It also relies heavily on information from colleagues representing the RIs in the various WPs.

The deliverable is aimed at the specification of common operations (i.e. those operations which are common to several RIs and which could – with benefit for interoperation and maintenance – be standardised) and cross-cutting services (i.e. those services which can act across many or all RIs and which could – with benefit in increased range of services and reduced maintenance – be adopted by some, most or all RIs).

The deliverable assumes that – as a key part of the architecture and following best practice in interoperating distributed systems – there is a conceptually rich metadata catalog or consistently-represented interoperable catalogs available to all component services. All RIs agree that a catalog is necessary although RIs have varying standards and practices for their catalogs.

This catalog or catalogs (however implemented – whether physically realised or acting as a reference specification) has to interoperate with – and therefore be a superset of – the service catalogs (existing or implicit) of the individual RIs to provide the interoperability required. This technique has been used for many years in various domains. Building on earlier work, the prime reference is [Sheth and Larsen 1990] although there has been much subsequent development and elaboration for assets beyond databases.

The RIs within ENVRIplus are at varying stages of maturity with more-or-less developed ICT support (the e-RI) and in particular with independent evolutionary paths to date. A fundamental principle of ENVRIplus is that if RIs can share expertise in the form of common operations and cross-cutting services (of which the first would be concerned with catalog interoperation) then (a) the research communities benefit from better systems and interoperation; (b) the cost of ICT systems maintenance for each RI is reduced; (c) it is possible to interoperate across the RIs so encouraging new research based on multidisciplinary science.

The method outlined below follows well-known engineering principles of identification, characterisation and proposition leading to integration into the evolving proposed architecture and description in the RM (WP5).

The method is as follows:

1. Identify common operations that exist or are planned in all or a significant number of RIs. This is aided by the agile use cases work and their progressive description and characterisation in the RM (Reference Model) refined further in D5.2 using information from D5.1;
2. Characterise them to understand different features, inputs, outputs and parameters;
3. Propose common operations – based on best practice deduced from state of the art – that could replace or augment the existing ones and thus (a) reduce cost in shared

maintenance; (b) increase ease of use because a researcher using >1 RI finds the same operations available;

4. Document the common operations as a community recommendation after extensive consultation;
5. Identify from use cases and requirements the required cross-cutting services (i.e. interoperation mechanisms);
6. Characterise the metadata- needed by the identified cross-cutting services - describing data, software components, services, users and resources (such as sensors) at each RI (commonly in one or more catalogs);
7. Define a canonical superset as a conceptual metadata scheme[2] to allow mapping and conversion between these metadata formats via the canonical 'exchange' conceptual metadata catalog;
8. Propose the cross-cutting services to be supported to provide the interoperation using the canonical metadata catalog;
9. Document the canonical metadata scheme and the mappings/conversions as a recommendation.

The aspects concerned with the catalog and metadata overlap with work done in WP8, which covers T8.1 curation, T8.2 catalog and T8.3 provenance. These aspects are related through the rich canonical metadata catalog (see steps 7-9 above) that interoperates with the RIs' separate catalogs and also provides information to drive the services for interoperation. The relevant common operations and cross-cutting services were documented initially in the ENVRI-RM (ENVRI Reference Model) and – in parallel with the work producing this deliverable – updated, refined and improved in D5.2.

[2] Here we distinguish scheme from schema with the latter in the sense of database schema with entities, attributes, constraints; scheme is less restrictive and includes semantics (including lists of allowed values) with crosswalks between different semantic sets

1.3 Conceptual Interoperation in the context of ENVRIplus

There is increasing understanding [Sheth and Larsen 1990] of how to design the architecture for interoperation in heterogeneous distributed systems and how to plan developments within that context. A key feature is trying to satisfy simultaneously the requirement for improved services within individual RIs and the requirement for cross-RI (interoperable) services. ENVRIplus covers many heterogeneous ICT systems in RIs and by its nature is a distributed system. Typically, a user of one RI will do most of her research at that RI using the data and resources delivered by the ICT system of that RI but increasing numbers of researchers also require access to other RIs. Thus, in addition to common operations providing reduced cost at each RI (by standardisation and re-use) the second benefit is that the user (or a workflow generated by the user) finds the same operations at other RIs when interoperating.

Services provide an end-user with a wrapped set of assets such as data, software and access to computing resources or sensors. Beyond services there exist the data assets themselves: the same asset may be composed into multiple different services. A key feature of interoperation is access to the assets (e.g. datasets or software) through metadata as well as to the services – also through the metadata.

These processes for interoperation are:

1. Discovery of assets: datasets, software, workflows, services, persons as experts, organisations, facilities, equipment, publications (white and grey) etc.;
2. Contextualisation of assets: ensuring relevance and quality for the purpose of the user;
3. Action: support of (and later autonomic) construction of a workflow to execute the user request.

Over time it is expected (the purpose of ENVRIplus) that interfaces to one or more RIs and their assets will become standardised so encouraging not only increased utilisation but also interoperability. This standardisation extends to common operations available at all RIs and also cross-cutting services to allow the end-user to initiate the same common operations across multiple RIs so facilitating multidisciplinary research. Once that state is reached, the heterogeneity lies particularly in:

(a) The datasets (both syntactic and semantic);
(b) Particular software required for processing e.g. data collected by instrumentation due to the nature of the sensors/instruments and particular data formats;
(c) Particular software required for particular analyses, simulations or visualisations due to the research being conducted and the multiple data formats.

Most RIs focus their attention on domain-specific metadata that may be governed by their discipline's collaborative agreements. They will often mix this with more generic data as the result of standards emerging from W3C and OGC, as well as compliance with governmental and stakeholder requirements, such as compliance with the INSPIRE directive. Such metadata choices are far from universal and pervasive across the ENVRIplus RIs. When accessed from another RI (or from a VRE) a conceptual canonical metadata format is required, which is understandable both to the RI holding the dataset (via convertors) and to the end-user working from her particular RI environment. An example of success in another domain (health) is UMLS[3]. Of course, if all ENVRI RIs agreed on the same metadata format (syntax and semantics)

[3] https://www.nlm.nih.gov/research/umls/new_users/online_learning/OVR_001.html

interoperation becomes relatively trivial (there are still governance issues and non-functional requirements to consider) and this has been done among a few RIs.

However, most ENVRI RIs have heterogeneous metadata and assets. Thus, for any assets (not just datasets) conversion is necessary. One-off conversion between any pair may be done using transformation techniques e.g. XSLT[4] with manual re-scripting as metadata standards evolve. More generally, pairs of convertors – usually implemented as brokers – are necessary between metadata formats. Given *n* metadata formats then *n*(n-1)* (i.e. almost *n* squared) pairs of convertors would be needed using typical programmatic brokering technology where the conversion is 'hard-wired' into the software. This technique requires much effort and cost in reprogramming as metadata standards evolve. An alternative approach is to identify a canonical superset metadata scheme [Jeffery and Koskela 2015]. Building convertor pairs only between the required metadata standards of each RI and the canonical metadata scheme reduces to *n* the convertor pairs to be built as metadata-driven brokers. This is the approach recommended by RDA (Research Data Alliance) [Nativi et al 2015]. Moreover, this approach can also promote the cross-RI support for the high-level supporting environment, such as Virtual Research Environments, e.g., VRE4EIC.

Given the metadata mappings, then convertors can be generated using the mapping as a specification. For data, this demonstrated success quite early in geoscience [Sutterlin et al 1977] with manual construction of the convertors from the specifications derived from the mappings. Some attempts to automate this more generally have been partially successful [Skoupy et al 1999]. Convertors for software have – in general – been less successful although automated re-writing of software from one language to another has been demonstrated using cross-compilers and/or interpreters[5].

The RIs rely to a greater or lesser extent on underlying e-Infrastructures providing basic services of networking, computing platforms, data storage facilities and open access to research publications. Examples are GEANT, PRACE, EGI, EUDAT, OpenAIRE and the emerging EOSC (European Open Science Cloud). The eponymous project (starting January 2017) may attempt to cover some of the wider issues beyond e-Infrastructure. Most RIs also have their own computing platforms, which provide some or all of the services outlined above. Most also have access to networks of equipment/sensors/detectors with appropriate processing. Currently many RIs within ENVRIplus have an existing or planned user access portal within the ICT system of the RI. Some just have a simple UI (user interface) such as a web page displaying the basic metadata and URL for access to assets. A few RIs are placed in an integrated 'silo' with user interface/portal/VRE [Candela et al 2013] at the user facing end and tightly integrated e-I facilities (e.g. access to cloud computing) at the infrastructure end. This has advantages of integration and potential cost-savings for one RI but (a) reduces choice and therefore the ability of the RI to obtain the best 'deals' from e-I suppliers; (b) limits scalability because of the choice of e-I; (c) inhibits interoperation beyond the group in the silo because of silo 'lock-in'; (d) makes it more difficult to have a fully featured VRE spanning across RIs beyond the silo to allow wider interdisciplinary research. These types of RI e-Infrastructures are illustrated (Figure 2):

Deleted: F

[4] XSLT: Extensible Stylesheet Language Transformations
[5] https://en.wikipedia.org/wiki/Source-to-source_compiler

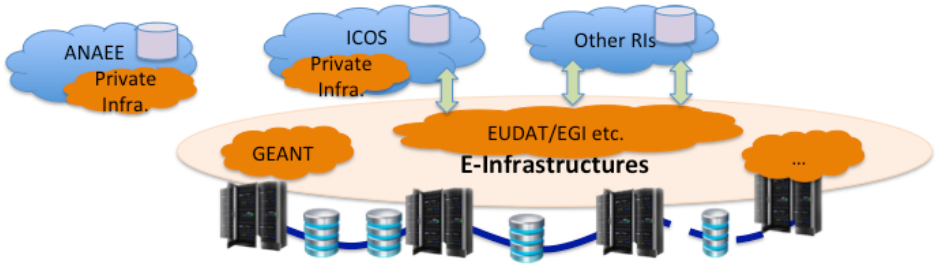

FIGURE 2: SIMPLIFIED VIEW OF TYPES OF RIS IN ENVRIPLUS

Thus, a user accessing multiple RIs is faced with heterogeneity. The portals are of different designs with varying offerings and possibilities for the end-user, and the UI in other cases may be simple commands or a web page of hyperlinks. There may or may not be API (programmatic) access to RI services.

An architecture for RI to RI interoperation would provide an end-user at one RI access to all other RIs required as if the other RIs were part of her RI. To achieve this, it is necessary either for each RI to be able to interconvert (convertor pairs) with every other RI (the n^2 problem described above) or, alternatively, each RI interconverts with a conceptual canonical superset metadata catalog (or limited set of catalogs) reducing the convertor pairs to n. The conceptual canonical superset catalog provides the reference local standard for interoperation and – by matching and mapping – the specification for the convertors required at each RI to interconvert between the local metadata standard and the canonical standard and furthermore to be able to interconvert the RI assets (especially datasets). The effort required to achieve this, once the scope and form of the canonical target has been agreed, is considerable, but requires significantly less technical effort and maintenance than that for pairwise conversion. This architecture may be seen as a step away from silos and towards interoperation between RIs (Figure 3):

Deleted: F

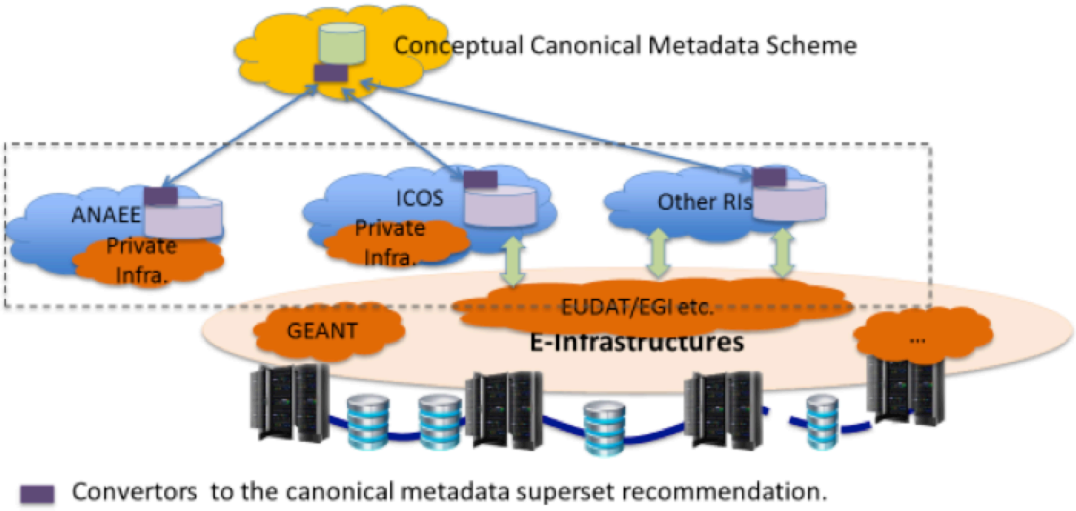

FIGURE 3:USER AT ONE RI USING ASSETS OF ANOTHER RI

However, for truly interoperative access a further architectural step is needed making the conceptual canonical superset catalog(s) physical which can be provided by a set of RIs, or by an external environment. In addition to RI-RI interoperation, third party (i.e. users not belonging to any particular RI for example a citizen scientist or policymakers) require homogeneous access via a VRE (Virtual Research Environment) or some portal system spanning multiple heterogeneous

RIs. Existing RI users could also use such a facility. Some ENVRIplus beneficiaries are participating in [VRE4EIC] which may be visualised as operating with ENVRIplus RIs as in (Figure 4). The VRE4EIC project aims to develop a reference architecture and toolset for VREs and is working closely with the other EC-funded VRE projects in the cluster as well as VLs (Virtual Laboratories) in Australasia and SGs (Science Gateways) in North America. Each RI would provide common operations (services) as extensive as possible, linked together by cross-cutting services whilst maintaining local analytical, simulation and visualisation facilities appropriate to that RI, together with the domain-specific datasets.

Deleted: F

Recommending a superset of the existing metadata standards for ENVRIPLUS RIs and providing mappings between individual RIs and this superset will promote the achievement of the homogeneous view over heterogeneity; such a solution (a) interoperates with the catalogs of each RI; (b) has a superset canonical homogeneous representation of the heterogeneity of the existing or planned catalogs of the RIs; (c) has appropriate content to support the processes required as defined in D5.1. Of course, if the RIs all used the same canonical catalog format as the superset catalog (but with content partitioned for their own domain) then interoperation would be much easier. It should be noted that the canonical conceptual catalog does not preclude – and indeed encourages and facilitates – RI to RI interoperation using already agreed metadata standards and interoperation processes. However, the proposed architecture includes the VRE option for wider end-user access e.g. for citizen science.

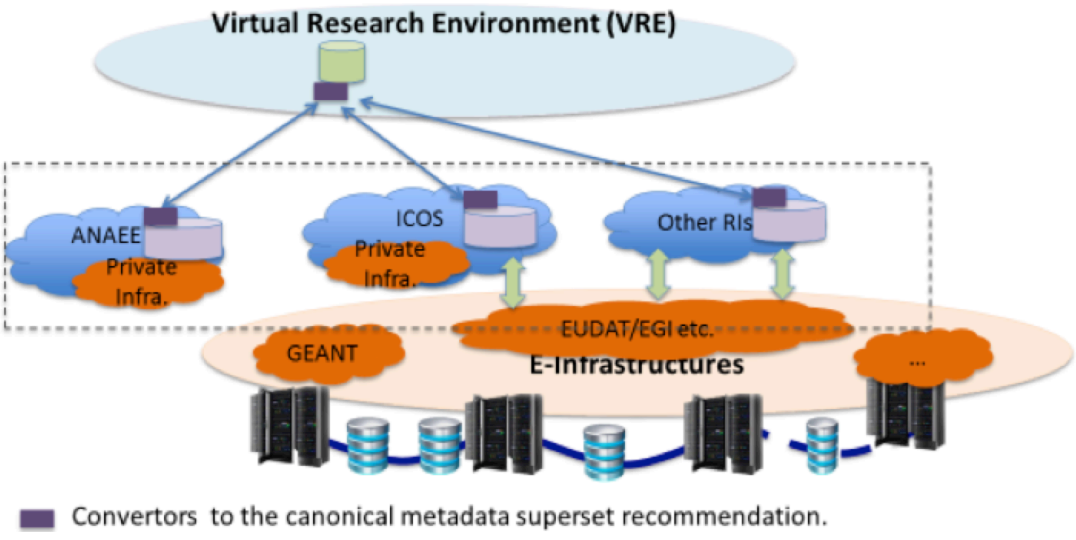

FIGURE 4: EXTERNAL USER ACCESSING MULTIPLE RIS VIA A GENERIC VRE

Each RI will need to judge the benefits of providing interoperation and using common or cross-cutting services in the context of its own commitments and priorities, in order to decide when and how far to engage. They may also be concerned about retaining identity or ensuring that the projects they support have independence and identity. Some RIs have international agreements which require to be honoured. Their investment, particularly their community's culture and working practices, will need to be preserved or nurtured through any transition. They may also feel that committing to multi-RI conventions may inhibit their ability to innovate when new opportunities emerge in their own disciplines.

It is necessary also to track developments in interoperation of environmental science RIs in other continents. This provides not only state of the art but also a model (or models) for comparison. As one example, DataONE in North America provides essentially a portal to datasets but also

provides interoperation capabilities [Cook et al 2012]. It also stresses data management planning and provides extensive education facilities as well as encouraging user exchange of best practices. It is following essentially the same approach as ENVRIplus and the metadata activity in RDA is currently co-led by representatives of DataONE and ENVRIplus[6].

It should be noted that a simple level of interoperation can be achieved with OpenSearch[7]. If each RI generates simple metadata to the specification required, then OpenSearch can select assets described by that metadata from multiple RIs which expose that metadata and provide the appropriate interface. However, the simplicity of the metadata (even with the OpenSearch optional extensions) precludes very precise relevance and recall and there is no attempt at integration of the assets selected. Thus, this technique requires considerable manual effort by the researcher when attempting multidisciplinary research.

[6] https://www.rd-alliance.org/groups/metadata-ig.html
[7] http://www.opensearch.org/Home

2 COMMON OPERATIONS

2.1 Introduction

The provision and adoption of commonly required operations across the multiple RIs in ENVRIplus will (a) produce cost savings in development and maintenance; (b) provide the basis for interoperation since the end-user and her workflow would have available the same operations in another RI compared with those at her usual RI. Ideally the operations are implemented as services (with advantages for software development including functional and non-functional parameters) i.e. within a SOA (Service-Oriented Architecture). Since 2010 (but mainly in 2014) experimental attempts have been made to construct systems with services self-organising based on local metadata associated with the service and metadata describing the architectural constraints although an initial proposal was made earlier [Giorgiadis et al 2002]. Generally, it is found necessary to orchestrate services (e.g. using BPEL[8]) or integrate services (e.g. using ESB[9]) both of which impose layering. Providing the services within a layered architecture as illustrated above ensures clear separation of functionality of concerns (as discussed and defined in the ENVRI RM)[10] and allows RIs to choose appropriate e-I and VRE services through clearly defined interfaces. There are 6 major topics of operations that have been identified in D5.1: curation, identification/citation, cataloguing, processing, optimisation, provenance (within which there are operations at service level) and 3 cross-cutting aspects to ensure harmonisation across RIs: architecture design for RI development, linking model for meta-information linking and RM for the common vocabulary. The common operations and cross-cutting services lie in the intersections of these topics and aspects.

2.2 Identification of Common Operations

The work reported in D5.1 included an attempt to identify common operations. The heterogeneity found among the various RIs was significant and some of the software appeared to be somewhat monolithic rather than structured and modular (which is necessary to allow re-use and interchange). This made deducing the underlying common operations somewhat difficult.

2.3 Characterisation of Common Operations

The work in D5.1 also attempted to characterise the operations that existed in existing RIs that could be considered as common operations. The analysis (both of characterisation of RIs and requirements) was based on the topics of ENVRIplus linked to the work of the various WPs as indicated below:

1. Identification and citation (WP6);
2. Curation (WP8);
3. Cataloguing (WP8);
4. Processing (WP7);
5. Provenance (WP8);

[8] Business Process Execution Language

[9] Enterprise Service Bus

[10] ENVRI RM Section 6.1: Environmental Research Infrastructures are supported by ICT systems that should be developed following state-of-the-art software engineering methods and architectures; which currently include the application of Layered, Service Oriented Architecture (SOA) and Cloud Architectural models.

6. Optimisation (WP7);
7. Community Support (WP15, WP14);

There was some divergence emerging from D5.1 between consideration of (a) ICT systems for managing data collection through sensors and other equipment; (b) ICT systems for managing data in a data centre including curation and provenance; (c) ICT systems for data analytics, visualisation, data mining and simulation; (d) ICT systems for managing the whole research lifecycle including management of projects, project proposals, funding, CVs, bibliographies etc.

2.3.1 ICT systems for managing data collection through sensors and other equipment

This aspect is covered by WPs 1,2,3 and 4. At present there is relatively little commonality among systems collecting data from sensors or detectors whether stand-alone or in networks. Various standards exist for particular kinds of sensor and OGC (Open Geospatial consortium) – within its SWE (Sensor Web Enablement) programme – has recommended sensorML[11] to describe sensors but its XML representation has been criticised since it is large although comprehensive. For data collection from sensors OGC SOS[12] is the commonly used standard and there is also the Semantic Sensor Network Ontology from W3C[13].

2.3.2 ICT systems for managing data in a RI including curation and provenance

This section relates closely to T8.2 and D8.3 (catalog) but also to T8.1 and D8.1 (curation) and the upcoming work in T8.3 (provenance), which has actually started already to ensure the catalog definition is sufficient for the purposes of provenance. The recommendation (from D8.3 of T8.2 and also from the improvements to the ENVRI RM in D5.2 following on from the analysis of requirements and state of the art in D5.1) is to have a superset canonical conceptual catalog, which interoperates with existing and planned RI catalogs. The canonical conceptual catalog may be realised physically in several ways; by a superstructure or by local structures at each RI. The semantic linking framework (T5.3) should provide the specification of the matching and mapping necessary to power the convertor pairs between the individual RI metadata standards and that of the canonical conceptual scheme. This mechanism allows the RIs to evolve towards interoperation progressively. The RI catalogs use a large variety of metadata standards including variants or dialects of ISO, W3C, European directive or European recommended standards. The most commonly used (or planned to be used) relevant metadata standards are:

(a) [CKAN]: Comprehensive Knowledge Access Network in several dialects but most commonly that of EUDAT B2SAVE/B2FIND;

(b) [ISO19115]/[INSPIRE]: an ISO standard and an EU directive for geospatial data discovery commonly through CSW (Catalog services for the web) of OGC (Open Geospatial consortium) as the XML encoded version ISO19139. There are many domain-specific named standards (especially in oceanography for example) that are based on INSPIRE e.g. [SEADATANET];

(c) [NetCDF]: widely used as a data standard (with metadata) in environmental science;

[11] http://en.wikipedia.org/wiki/SensorML
[12] http://www.opengeospatial.org/standards/sos
[13] https://www.w3.org/TR/vocab-ssn/

(d) [SensorML]: used – as the name indicates – especially for metadata associated with sensor data;

(e) SSNO (Semantic Sensor Network Ontology) from W3C mentioned above;

(f) [DC]: Dublin Core from W3C a simple metadata format of 15 elements but with various dialects and semantics. Several other metadata standards incorporate the DC 15 elements;

(g) [DCAT] (Data Catalog Vocabulary) from W3C based on DC but extended to manage catalogs. A recent workshop[14] was held (under the auspices of VRE4EIC with W3C) to plan extensions to DCAT to overcome some perceived shortcomings;

(h) [CERIF] (Common European Research Information Format, an EU recommendation to Member States used widely for research information and within ENVRIplus as the catalog format in EPOS [Bailo et al 2016] including interoperation via CERIF-XML. CERIF can interoperate with most of the others (mapping has been done and convertors are available) and also with OIL-E, the language used to express the ENVRI RM (ongoing work in the VRE4EIC project) so it can be validated by the ENVRI RM (Reference Model) since the formal mapping has been done.

It should be noted that in other domains outside of ENVRIplus there are many more standards. An example from social science is [DDI]: Data Documentation Initiative. Should in future RIs from such domains need to interoperate with ENVRIplus RIs (for multidisciplinary research) then their metadata characteristics will have to be mapped to the canonical metadata catalog.

The canonical superset catalog (or catalogs) has to support discovery, contextualisation and action, interoperation with multiple metadata standards and be able to represent temporal information for curation, provenance and versioning.

The common data management operations on a catalog are:

1. Input or edit metadata directly;
2. Input metadata via a convertor from another standard directly or by harvesting;
3. Output metadata directly;
4. Output metadata via a convertor to another standard which may be requested through harvesting;
5. Validate metadata (by software including logical constraints using computer processing or by user manual checking);
6. Display metadata in tabular or graphical form;
7. Set select, project, union, difference, Cartesian product (join) in inner and outer versions (i.e. the relational operators) acting on the metadata;

In addition, the catalog acts as the mediator for more substantive data management actions where the metadata record(s) are updated or added to record the action. These actions include:

1. Download dataset to <location> with or without conversion;
2. Download software to <location> with or without conversion;
3. Upload dataset to <location> with or without conversion;
4. Upload software to <location> with or without conversion;

The conversion is necessary should the target platform environment be unable to handle a canonical format (at present the usual case). In all these cases, the data management operation has also to ensure that NFRs (non-functional requirements) recorded in the metadata are

[14] https://www.w3.org/2016/11/sdsvoc/

satisfied; typically, these include rights, privacy and security aspects as well as performance commonly recorded as service level agreements. Furthermore, each operation needs to be recorded in the metadata to provide a provenance trace and curation records to be used as appropriate information for contextualisation in any subsequent request.

The user request(s) to the catalog result in a set of assets to be utilised to satisfy the user request. The user can manually assemble the assets into a workflow but it is anticipated that semi-automated and then fully autonomic workflow construction can be achieved, usually in the VRE. This implies several more common operations:

1. Assemble/compose workflow: this is a complex task, involving not only software and datasets but also platforms for computing and perhaps access to sensors; furthermore, the NFRs relevant to the whole workflow have to be respected. The workflow may well include the creation of a 'working set' of data assembled by data management operations on the individual (distributed) datasets and this 'working set' requires metadata so that it may be used as a first-class object in cataloguing, curation and provenance and any subsequent computing operations;
2. Validate workflow: confirm with the user that the request intent is satisfied by the proposed workflow;
3. Execute workflow: including reporting progress to user to allow 'steering'. This includes moving datasets and software to various distributed locations (as indicated above), ensuring the workflow serialisation (even over distributed parallel execution) is maintained and respecting the NFRs;

These operations all depend on the catalog for discovery and contextualisation and also each operation has to record in the metadata the result in order to provide provenance and curation information and to provide contextualisation metadata for future requests. New data products may be produced, new datasets and potentially new workflow specifications and these need to be recorded for curation and provenance purposes.

Of course, the workflow is not specific to data management and may include operations provided as services from the subsequent sections below.

2.3.3 ICT systems for data analytics, visualisation, data mining and simulation

This section relates to WP7: Data processing and analysis. At present the various RIs within ENVRIplus have varying levels of provision of – what might be described broadly as – analytics. The requirement – derived from a variety of questionnaire responses and use cases in D5.1 - is for a set of modular services, each of which performs an analytical function, which can be composed into workflows together with appropriate data (and any necessary convertors) to provide the end-user with the required result of her request. Monolithic analytics solutions – while useful in particular circumstances – do not accord with the interoperability and common operations philosophy of ENVRIplus. Nonetheless, established working practices require support – at least for a transition period – and so some degree of backward compatibility is required. The combination of a rich metadata catalog with curation and provenance can assure this.

The required common operations are:

1. Univariate statistics, sum, mean, standard deviation, median;
2. Bivariate statistics: correlation;
3. Bivariate statistics: analysis of variance;
4. Multivariate statistics: factor analysis, principal components analysis, discriminant analysis;

5. Time series analysis;
6. Data mining numeric;
7. Data mining textual;
8. Visualisation univariate – graph with points;
9. Visualisation bivariate – graph with points
10. Visualisation multivariate – graph (lines with points), 3-D perspective (towers, bars, lines);
11. Simulation: a library of simulation routines (representing e.g. Navier-Stokes) for composing for a particular purpose;

Of course, individual RIs have more specialised analytical services already implemented, planned or identified in the requirements of D5.1. The ENVRIplus architecture needs to be flexible enough for RIs to take advantage of the latest technologies. Again, this is an advantage of using a rich metadata catalog; achieving this flexibility through reprogramming (including changing of interface specifications) as necessary is too expensive and time-consuming.

2.3.4 ICT systems for managing the whole research lifecycle including management of projects, project proposals, funding, CVs, bibliographies etc.

From the information in D5.1, this area has – in general – not been given much consideration by the RIs within ENVRIplus but user requirements for these services exist. The provision of systems providing these services is widespread in universities, research laboratories and funding agencies. In the research-performing organisations such systems provide external-facing information about research activity including projects, persons (as experts), publications and other results. This information is used to create new partnerships with other academic institutions and with industry for innovation and wealth-creation and/or improvement in the quality of life. The systems also provide automated management reporting within the organisation and to external stakeholders including funders. The systems are used to measure performance and compare with other organisations – both by the research-performing organisations and by funders.

The information in such systems is therefore useful to manage the research lifecycle and reduce the burden on the researcher – especially continually reporting in different ways to various stakeholders about the research.

Typical operations are:

1. Input directly metadata on organisations (including an organisational structure with a network of units), persons, projects, publications, patents, products and the relationships between them;
2. Ingest via convertors from catalogs in other metadata standards metadata on organisations, persons, projects, publications, patents, products and the relationships between them;
3. Export via convertors to catalogs in other metadata standards metadata on organisations, persons, projects, publications, patents, products and the relationships between them;
4. Select organisations (including an organisational structure with a network of units), projects, persons, publications etc. using certain criteria;
5. Display in tabular or graphical form the results of the selection;

Various standards exist and can be utilised. For researcher identification, there is ORCID[15], for documents DOI[16], for cross-referencing of documents CrossRef[17] and for referencing to funding FundRef[18]. CERIF is used widely in this domain by funders (such as the EC for the European Research Council Grant system and some national funders) and by research organisations such as universities and research laboratories. Commonly commercial implementations from Elsevier[19] and Thomson-Reuters[20] are used.

[15] https://orcid.org/
[16] https://www.doi.org/
[17] http://www.crossref.org/
[18] http://www.crossref.org/fundingdata/
[19] https://www.elsevier.com/about/press-releases/science-and-technology/elsevier-acquires-atira,-a-provider-of-research-management-solutions
[20] http://converis.thomsonreuters.com/

2.4 Proposed ENVRIplus Common Operations and relationship to ENVRI RM

The common operations are those listed in Section 2.3 and derived from D5.1 questionnaires, use cases and analysis. The following table presents these common operation and their relationships to the ENVRI RM as currently expressed in D5.2.

	Operation Identified for ENVRIplus	**Operation in ENVRI RM**
1	Input or edit metadata directly;	Catalog service update catalog
2	Input metadata via a convertor from another standard;	Setup mapping rules Perform mapping
3	Output metadata directly	Part of Catalog Service
4	Output metadata via a convertor to another standard;	Setup mapping rules Perform mapping
5	Validate metadata (by software including logical constraints or by user manual checking);	<not defined at appropriate detail> but could be done by setup mapping rules and perform mapping
6	Display metadata in tabular or graphical form;	Part of catalog service query resource
7	Set select, project, union, difference, Cartesian product (join) in inner and outer versions (i.e. the relational operators) acting on the metadata;	Part of catalog service query resource
8	Download dataset to <location> with or without conversion;	Data transfer service including transporter; conversion not defined unless within process data (where it not defined in detail)
9	Download software to <location> with or without conversion;	<not defined at appropriate detail> But could use Data transfer service including transporter; conversion not defined unless within process data (where it not defined in detail)
10	Upload dataset to <location> with or without conversion;	Data transfer service including transporter; conversion not defined unless within process data (where it not defined in detail)

11	Upload software to <location> with or without conversion;	<not defined at appropriate detail> But could use Data transfer service including transporter; conversion not defined unless within process data (where it not defined in detail)
11	Assemble/compose workflow: this is complex involving not only software and datasets but also platforms for computing and perhaps access to sensors; furthermore, the NFRs relevant to the whole workflow have to be respected;	<not defined at appropriate detail> part of Coordination service/processing service
12	Validate workflow: confirm with the user that the request intent is satisfied by the proposed workflow;	<not defined at appropriate detail> part of Coordination service/processing service
13	Execute workflow: including reporting progress to user to allow 'steering'. This includes moving datasets and software to various distributed locations (as indicated above), ensuring the workflow serialisation (even over distributed parallel execution) is maintained and respecting the NFRs;	<not defined at appropriate detail> Could be partly covered by Data Use and/or part of Coordination service/processing service
14	Univariate statistics, sum, mean, standard deviation, median;	<not defined at appropriate detail> Could be within Data Processing
15	Bivariate statistics: correlation;	<not defined at appropriate detail> Could be within Data Processing
16	Bivariate statistics: analysis of variance;	<not defined at appropriate detail> Could be within Data Processing
17	Multivariate statistics: factor analysis, principal components analysis, discriminant analysis;	<not defined at appropriate detail> Could be within Data Processing
18	Time series analysis;	<not defined at appropriate detail> Could be within Data Processing
19	Data mining numeric;	<not defined at appropriate detail> Could be within Data Processing
20	Data mining textual;	<not defined at appropriate detail> Could be within Data Processing
21	Visualisation Univariate – graph with points;	<not defined at appropriate detail> Could be within Data Processing
22	Visualisation bivariate – graph with points	<not defined at appropriate detail> Could

		be within Data Processing
23	Visualisation multivariate – graph (lines with points), 3-D perspective (towers, bars, lines);	<not defined at appropriate detail> Could be within Data Processing
24	Simulation: a library of simulation routines (representing e.g. Navier-Stokes) for composing for a particular purpose;	<not defined at appropriate detail> Could be within Data Processing
25	Input directly metadata on organisations (including an organisational structure with a network of units), persons, projects, publications, patents, products and the relationships between them;	Catalog service update catalog unless within process data (where it not defined in detail)
26	Ingest via convertors from catalogs in other metadata standards metadata on organisations, persons, projects, publications, patents, products and the relationships between them;	<not defined at appropriate detail> unless within process data (where it not defined in detail)
27	Export via convertors to catalogs in other metadata standards metadata on organisations, persons, projects, publications, patents, products and the relationships between them;	<not defined at appropriate detail> unless within process data (where it not defined in detail)
28	Select organisations (including an organisational structure with a network of units), projects, persons, publications etc. using certain criteria;	Part of catalog service query resource
29	Display in tabular or graphical form the results of the selection;	Part of catalog service query resource

It is clear from the above that the ENVRI RM is working at a level of abstraction higher than the operations identified from D5.1 and here condensed to common operations. Thus, while many of the identified operations can be imagined to fit within the ENVRI RM defined components ,that remains to be validated by more detailed descriptions of the components and their properties. On the other hand, several of the identified common operations have no obvious relationship to current ENVRI RM components. Recent and parallel work to produce D5.2 has addressed some of these issues, the planned developments of the ENVRI RM (particularly in the area of the canonical metadata catalog and related operations) and the planned work on the Engineering Viewpoint will address more.

3 CROSS-CUTTING SERVICES

3.1 Introduction

Cross-cutting services are intended to make interoperation among RIs easier; that is to enable a user of one RI to be able to access other RIs as if they were the same as her usual RI. There are 3 cross-cutting aspects to ensure harmonisation across RIs: architecture design for RI development, linking model for meta-information linking and RM for the common vocabulary. There are 6 major topics of operations: curation, identification/citation, cataloguing, processing, optimisation, provenance (within which there are operations at service level). The common operations and cross-cutting services lie in the intersections of these topics and aspects.

There are several kinds of cross-cutting service. The basic ones are:

1. Move metadata – with any necessary conversion – from one location to another;
2. Move data – with any necessary conversion – from one location to another;
3. Move a software module – with any necessary conversion – from one location to another;
4. Move a workflow specification – with any necessary conversion – from one location to another;

More advanced services are:

1. From RI A (or a VRE over one or more RIs) initiate one or more processes (a workflow) on one or more datasets on RI B;
2. From RI A (or a VRE over one or more RIs) initiate one or more processes (a workflow) on one or more datasets on RI B, RI C, RI D…;

Eventually, inter-operation and a holistic information model delivered via the canonical metadata scheme will enable (mostly automated) workflows to work across RIs and e-Is. Ultimately, users should not have to be concerned with which infrastructures are enacting which parts of their workflows, and the optimisers should be choosing platforms to minimise costs that are identified by researchers or resource providers.

It is assumed that the target RIs manage the interaction with the underlying e-Is although the characteristics of the e-Is must be known to the RI catalog and – if initiated from a VRE over multiple RIs – to the VRE catalog.

3.2 Identification of Cross-Cutting Services

At present cross-cutting services beyond the metadata level between heterogeneous RIs in ENVRIplus do not really exist, although domain-specific metadata formats are common in some groups of RIs (e.g. in oceanography) permitting interoperation. Within RIs there are some examples of services overcoming heterogeneity among facilities and services of the included organisations. Within ENVRIplus RIs there is relatively limited expertise in simple interoperation (metadata matching and mapping, provision of convertors) although conformance with the EUDAT standards and services allows some interoperation among closely-related RIs. There is almost no experience in complex interoperation (distributed parallel execution of workflows over multiple RIs).

3.3 Characterisation of Cross-Cutting Services

Since few if any cross-cutting services exist they cannot be characterised. However, the requirement for such services is recorded and understood and the services are listed above. The key aspect is that they all require a superset canonical metadata catalog.

3.4 Proposed Canonical Metadata Scheme and Mappings

It is unrealistic to expect all ENVRIplus RIs to have or adopt a single metadata scheme. Each has chosen (or is planning to choose) its metadata scheme for its own particular purposes. The ENVRIplus canonical conceptual metadata scheme therefore has to be:

(a) A superset of existing and planned metadata standards used in RI catalogs (although for interoperation a subset would be used);
(b) Have mappings and convertors from and to those RI metadata standards;
(c) Have formal syntax to assure integrity and correct machine processing;
(d) Have declared (multilingual) semantics to allow validation and to provide cross-walking between term sets in ontologies, thesauri or dictionaries;
(e) Be capable of describing not only datasets but also software, workflows, services and computing platforms (including sensors/detectors/equipment) related to organisations and persons;

Meeting these requirements would require considerable development of metadata standards such as DC, DCAT, CKAN or INSPIRE. As indicated above, CERIF can meet these requirements [Jeffery et al 2014] and of course other metadata standards could be extended to do so. D8.3 (from T8.2) recommends using both CERIF and CKAN (as used in EUDAT).

3.5 Proposed ENVRIplus Cross-Cutting Services and relationship to ENVRI RM

The following table presents these common operation and their relationships to the ENVRI RM as currently expressed in D5.2.

The cross-cutting services are those listed in Section 3.1 and derived from D5.1 questionnaires, use cases and analysis. The operations relate solely to cross-cutting services. Services that could be common to RIs are listed under common operations in section 2.

	Operation Identified for ENVRIplus	**Operation in ENVRI RM**
1	Move metadata – with any necessary conversion – from one location to another;	Data transfer service and data transporter but needs also convertors unless within process data (where it not defined in detail)
2	Move data – with any necessary conversion – from one location to another	Data transfer service and data transporter but needs also convertors unless within process data (where it not defined in detail)

3	Move a software module – with any necessary conversion – from one location to another	Data transfer service and data transporter but needs also convertors unless within process data (where it not defined in detail)
4	Move a workflow specification – with any necessary conversion – from one location to another	Data transfer service and data transporter but needs also convertors unless within process data (where it not defined in detail)
5	From RI A (or a VRE over one or more RIs) initiate one or more processes (a workflow) on one or more datasets on RI B;	Coordination service/processing service
6	From RI A (or a VRE over one or more RIs) initiate one or more processes (a workflow) on one or more datasets on RI B, RI C, RI D…	Coordination service/processing service

Similarly to what we had seen in regard to common operations, it is clear from the above that also in this case the ENVRI RM is working at a level of abstraction higher than the operations identified from D5.1 and here condensed to common operations. Thus, while many of the identified operations can be imagined to fit within ENVRI RM defined components that remains to be validated by more detailed descriptions of the components and their properties. On the other hand, several of the identified cross-cutting operations have no obvious relationship with current ENVRI RM components. Recent and parallel work to produce D5.2 has addressed some of these issues, the planned developments of the ENVRI RM (particularly in the area of the canonical metadata catalog and related operations) and the planned work on the Engineering Viewpoint will address more.

4 RECOMMENDATIONS

4.1 Introduction

The following recommendations define the actions that should be pursued in the remainder of the ENVRIplus project and beyond in order to facilitate interoperability.

4.2 Further Development of ENVRI RM

It is clear from the requirements and state of the art analysis (D5.1) and the analyses above that:

(a) Many RIs do not have the full range of services required by their users;
(b) Many RIs do not have the services required for interoperation (cross-cutting services);
(c) Many of the services were already specified – at a rather abstract level – in the ENVRI RM;
(d) Many of the services required were not specified or not specified in sufficient detail in the ENVRI RM and subsequent work is planned for their incorporation;

This implies that the ENVRI RM needs to be developed to a further level of detail, as the Engineering and Technology Viewpoints are developed, to match with the common and cross-cutting operations emerging from the requirements and state of the art analysis of D5.1.

4.3 Metadata

It is clear that the canonical metadata catalog or catalogs – to assist interoperation across RIs – is an essential component. It has to be able not only to describe (for discovery, contextualisation and action) the RI assets (such as datasets, software components workflows, persons (experts), organisations) but also the NFRs associated with them. The choice of (or development of) an appropriate superset canonical metadata format or formats is critical to the success of ENVRIplus. D8.3 from T8.2 recommends CERIF and CKAN (as used in EUDAT) as candidates. It may be necessary for BEERI to appoint a technical group to oversee the decisions on the catalog syntax and semantics, the content and the related operations.

4.4 Network of Data Managers and Developers

ENVRIplus has already a de facto network of data managers and developers at the various RIs. However, at present each RI has its own philosophy of purpose, its own processes and procedures, its own governance, its own characteristic datasets and software and its own computing platforms with or without sensors/detectors/instruments. The agile use case activities are highlighting the aspects that are common and those specialised to one RI. Progressively, these are being described in the RM in a formal way (OIL-E is encoded in RDF) allowing logic operations (e.g. deduction and induction) to aid reasoning such as comparison of the characterisation of RIs. Also, there is some turnover of staff and so the way in which a particular RI operates has to be communicated thorough induction and training activities, and novel ideas from the new staff have to be assessed and if supported, carried forward. This means that the plan is recommended to include:

1. Familiarisation with an agreement on the common superset catalog, operations and cross-cutting services;
2. Their development;
3. Their deployment including maintenance and upgrades;
4. A mechanism for dealing with new proposals for upgrading the components of the ENVRIplus environment;

4.5 Proposed Development Plan

The development plan was developed to guide the activities in Theme 2 but co-design with the IT representatives of the RIs has always been envisaged. Clearly, RIs' requirements will evolve and the co-design approach should ensure the Theme 2 activities run parallel to those changing requirements. The plan was developed with assistance from the RM, which focused views on the common problems and guided the discussion on the architectural design required. However, the plan needs to be sufficiently flexible to accommodate the evolving requirements of the RIs.

4.5.1 Familiarisation: M19-M24

A set of training activities will be developed and disseminated to assist data managers and developers familiarise with ENVRIplus concepts, architecture and development requirements. The training will be based on this deliverable and will be highly interactive to ensure engagement with the RIs.

4.5.2 Development: M19-M30

The list of operations that are common and to be developed will be prioritised (not least because there are some dependencies). The development activity will follow the agile methodology with short sprints and small teams – with IT people drawn from various RIs – working together.

4.5.3 Deployment as prototype: M30-M33

The common services will be deployed first as a prototype at a testbed RI and – once demonstrated, RIs will be invited to evolve their existing architecture and operations to adopt and utilise the common ENVRIplus set of common and cross-cutting services. For some RIs this activity will extend beyond the end of the project.

4.5.4 Upgrading mechanism: M30-M36

In order to ensure the software supporting the operations is current, an upgrading mechanism will be developed and implemented. This will involve (a) identification of new common / cross-cutting operations from novel requirements; (b) prioritisation and approval for development; (c) software development to support the operation followed by testing; (d) implementation in the testbed RI then adopted across RIs in ENVRIplus.

5 CONCLUSIONS

The deliverable is aimed at the specification of common operations (i.e. those operations which are common to several RIs and which could – with benefit for interoperation and maintenance – be standardised) and cross-cutting services (i.e. those services which can act across many or all RIs and which could – with benefit in increased range of services and reduced maintenance – be adopted by some, most or all RIs). Following an analysis of state of the art and requirements (D5.1) using agile groups there has been – at and between ENVRI meetings – discussions with other WPs especially within Theme 2. The characterisation of the RIs using the RM has provided a formal basis for description and analysis. The key features to emerge are (a) the importance of the conceptual canonical rich metadata catalog(s); (b) the definition of common operations (and specialised operations to particular RIs) to be encoded within a SOA as services and (c) the definition of cross-cutting services including convertors for metadata and RI assets such as datasets and software components – again to be encoded within a SOA as services.

The document made the following recommendations to ENVRIplus in order to support cross-RI interoperability:

1. the ENVRI RM needs to be developed to a further level of detail, as the Engineering and Technology Viewpoints are developed, to match with the common and cross-cutting operations emerging from the requirements and state of the art analysis of D5.1.
2. The choice of (or development of) an appropriate superset canonical metadata format or formats is critical to the success of ENVRIplus. D8.3 from T8.2 recommends CERIF and CKAN (as used in EUDAT) as candidates. It may be necessary for BEERI to appoint a technical group to oversee the decisions on the catalog syntax and semantics, the content and the related operations.
3. The Network of data managers and developers should have a programme of work including:
    a. Familiarisation with an agreement on the common superset catalog, operations and cross-cutting services;
    b. Their development;
    c. Their deployment including maintenance and upgrades;
    d. A mechanism for dealing with new proposals for upgrading the components of the ENVRIplus environment;
4. The Development Plan should be elaborated in more detail and then pursued.

6 IMPACT ON THE PROJECT

The main motivation of ENVRIplus is to enable researchers to access, utilise and interoperate across multiple RIs in the environmental domain. The provision of common operations and cross-cutting services is vital to this objective. The use of a rich superset canonical metadata format is required not only to provide the required access and utilisation of the assets across multiple RIs, but also to ensure the NFRs and governance aspects are managed appropriately.

Work documented in this document has identified common operations and cross-cutting services. This enables further work towards interoperability. Recommendations including a development plan have been provided.

7 IMPACT ON STAKEHOLDERS

The development and deployment of common operations and cross-cutting services will enable stakeholders to meet their requirements for interoperation across multiple RIs in the environmental domain. This will benefit researchers in their work but will also benefit data managers and systems staff because of reduced costs and improved effectiveness and efficiency of services. If ENVRIplus moves towards an environment including a VRE or similar easy-to-use comprehensive environmental research interfaces, then citizens may also benefit.

8 REFERENCES

[Atkinson et al 2016] Atkinson, M.; Hardisty, A.; Filgueira, R.; Alexandru, C.; Vermeulen, A.; Jeffery, K.; Loubrieu, T., Candela, L., Magagna, B., Martin, P., Chen, Y., Hellström, M. (2016) A consistent characterisation of existing and planned RIs. ENVRIplus D5.1

[Bailo et al 2016] Daniele Bailo, Damian Ulbricht, Martin L. Nayembil, Luca Trani, Alessandro Spinuso, Keith G. Jeffery 'Mapping solid earth Data and Research Infrastructures to CERIF' Proceedings 13th International Conference on Current Research Information Systems, CRIS2016, 9-11 June 2016, Scotland, UK. Procedia Computer Science (Elsevier) available at: https://www.epos-ip.org/sites/default/files/repository/blocks/CRIS2016_paper_16_Bailo.pdf

[Belloum et al. 2011] Belloum, A.S.Z., Inda, M. A., Vasunin, D., Korkhov, V., Zhao, Z., Rauwerda, H., Breit, T., Bubak, M.T. & Hertzberger, B. (2011). Collaborative e-Science Experiments and Scientific Workflows. IEEE Internet Computing, 15(4), 39-47.

[Hardisty et al 2016] BioVeL: a virtual laboratory for data analysis and modelling in biodiversity science and ecology BMC Ecology 2016 doi: 10.1186/s12898-016-0103-y

[Candela et al 2013] Candela, L., Castelli, D. & Pagano, P., (2013). Virtual Research Environments: An Overview and a Research Agenda. Data Science Journal. 12, pp.GRDI75–GRDI81. DOI: http://doi.org/10.2481/dsj.GRDI-013

[Candela et al. 2014] L. Candela, D. Castelli, A. Manzi, P. Pagano, Realising Virtual Research Environments by Hybrid Data Infrastructures: the D4Science Experience. International Symposium on Grids and Clouds (ISGC) 2014, Proceedings of Science PoS(ISGC2014)

[Cook et al 2012] R Cook, W Michener, D Vieglais, A Budden, R Koskela. Dataone: A distributed environmental and earth science data network supporting the full data life cycle. EGU General Assembly 2012. Available at https://scholar.google.co.uk/citations?view_op=view_citation&hl=en&user=EpBp494AAAAJ&citation_for_view=EpBp494AAAAJ:_FxGoFyzp5QC

[De Roure and Goble 2007] myExperiment-a Web2.0 Virtual Research Environment; available at http://eprints.soton.ac.uk/263961/1/myExptVRE31.pdf

[CERIF] http://www.eurocris.org/cerif/main-features-cerif

[CKAN] http://ckan.org/features-1/metadata/

[DC] http://dublincore.org/documents/dces/

[DCAT] https://www.w3.org/TR/vocab-dcat/

[DDI] http://www.ddialliance.org/Specification/

[EGI] European Grid Initiative, www.egi.eu

[EPOS] EPOS, a European Research Infrastructure on earthquakes, volcanoes, surface dynamics and tectonics, http://www.eposeu.org/

[EUDAT] European Data Infrastructure, http://www.eudat.eu/

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

[GEANT] - http://www.geant.org/

[Giorgiadis et al 2002] Self-Organising Software Architctures for Distributed Systems, WOSS '02 Proceedings of the first workshop on Self-healing systems, Charleston, South Carolina — November 18 - 19, 2002, pp 33-38, Available at http://dl.acm.org/citation.cfm?id=582135

[ICOS] ICOS, a European distributed infrastructure dedicated to the monitoring of greenhouse gases (GHG) through its atmospheric, ecosystem and ocean, http://www.icos-infrastructure.eu/

[INSPIRE] http://inspire.ec.europa.eu/metadata/6541

[ISO19115] http://www.iso.org/iso/iso_catalog/catalog_ics/catalog_detail_ics.htm?csnumber=53798

[Jeffery et al 2014] K.G. Jeffery, N. Houssos, B. Jörg, A. Asserson (2014) "Research Information Management: The CERIF Approach", Int. J. Metadata, Semantics and Ontologies, Vol. 9, No. 1, pp 5-14 2014.

[Jeffery and Koskela 2015] Keith G Jeffery and Rebecca Koskela (2015) 'RDA: The Importance of Metadata' ERCIM News Issue 100 January 2015 http://ercim-news.ercim.eu/en100/special/rda-the-importance-of-metadata

[JISC 2010] https://www.jisc.ac.uk/rd/projects/virtual-research-environments

[Miller et al.2012] Mark A. Miller, Wayne Pfeiffer, and Terri Schwartz. 2012. The CIPRES science gateway: enabling high-impact science for phylogenetics researchers with limited resources. In Proceedings of the 1st Conference of the Extreme Science and Engineering Discovery Environment: Bridging from the eXtreme to the campus and beyond (XSEDE '12). ACM, New York, NY, USA [

[Nativi et al 2015] Stefano Nativi, Keith G Jeffery, Rebecca Koskela (2015) RDA; Brokering with Metadata' ERCIM News Issue 100 January 2015 http://ercim-news.ercim.eu/en100/special/rda-brokering-with-metadata

[NetCDF] https://www.unidata.ucar.edu/software/netcdf/docs_rc/

[Sheth and Larsen 1990] Amit P Sheth, James A Larsen: Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases ACM Computing Surveys v22 no 3 September 1990

[SEADATANET] http://www.seadatanet.org/Metadata

[SensorML] http://www.opengeospatial.org/standards/sensorml

[Skoupy et al 1999] Skoupy,K; Kohoutkova,J; Benesovsky,M; Jeffery,K G: 'Hypermedata Approach: A Way to Systems Integration' Proceedings Third East European Conference, ADBIS'99, Maribor, Slovenia, September 13-16, 1999, Published: Institute of Informatics, Faculty of Electrical Engineering and Computer Science, Smetanova 17, IS-2000 Maribor, Slovenia,1999, ISBN 86-435-0285-5, pp 9-15

[Sutterlin et al 1977] P G Sutterlin, K G Jeffery, E M Gill (1977) 'Filematch: A Format for the Interchange of Computer-Based Files of Structured Data'. Computers and Geosciences 3(1977) 429-468.

[VRE4EIC] www.vre4eic.eu

[Zhao et al. 2016] Zhiming Zhao, Paul Martin, Cees de Laat, Keith Jeffery, Anrew Jones, Ian Taylor, Alex Hardisty, Malcolm Atkinson, Anneke Zuiderwijk- van Eijk, Yi Yin, Yin Chen: 'Time critical requirements and technical considerations for advanced support environments for data-intensive research' Proceedings IT4RIS workshop Porto 29 November-2 December 2016