Ref. Ares(2017)2227198 - 30/04/2017

D5.5
A Model Architecture for new RIs to Adopt and to act as a Guide for Existing RIs in their Development

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
Leonardo Candela	CNR
Daniele Bailo	INGV
Thomas Loubrieu	IFREMER

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

Barbara Magagna	EAA
Helen Glaves	NERC

Accepted by: Paola Grosso (WP 5 leader)

Deliverable type: [REPORT]

Dissemination level: PUBLIC

Deliverable due date: 30.04.2017/M24

Actual Date of Submission: 30.04.2017/M24

ABSTRACT

The major objective of ENVRIplus is to facilitate research in environmental science by encouraging movement towards a consistent and integrated view of data, processing and resources to meet emerging domain-specific and interoperation research needs. The adoption of common and cross-cutting ICT services by RIs (Research Infrastructures) reduces cost (re-use) and increases interoperation (standardisation). A key aspect of ENVRIplus is the reference architecture to be adopted by new RIs and towards which existing RIs should aim to align. Based on the ENVRI Reference Model, the architecture brings together all the aspects of the ICT (Theme 2) activities of ENVRIplus into a coherent framework to achieve those objectives. The architecture must sit within some constraints. ICT best practice is mandatory. Parallel initiatives in other ESFRI RIs and global consortia must be respected. Developments in e-Is (e-Infrastructures) provide opportunities for alternative deployment of applications. An appropriate interfacing mechanism between RIs and e-Is will provide for evolution of both RIs and e-Is while maintaining provision of service. Similarly, developments in VREs (Virtual Research Environments) offer improved opportunities for researchers (and other users) to access multiple RIs while appropriate interfacing will allow evolution of both RIs and VREs to sustain the consistent and integrated facilities built on the resources delivered by collaborating RIs. The degree of alignment with the architecture by RIs will improve their ability to present a research environment that supports research campaigns that need resources and capabilities from multiple RIs. The development of the ENVRIplus architecture is therefore continuous, and this deliverable (D5.5) presents the current state of progress at this point in the project. Further work on the RM (Reference Model) will provide specifications based on engineering and technology viewpoints at which time a conventional architectural design document can be produced.

Project internal reviewer(s):

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

Document history:

Date	**Version**
24.02.2017	Outline for comments
04.03.2017	Integrated comments from M Atkinson, A Hardisty;
31.03.2017	Consolidated version for further comments
03.04.2017	Project coordinator urged people to comment, consolidated version sent out

13.04.2017	Integrating P. Martin, A. Hardisty, L. Candela contributions
20.04.2017	Further integration of comment
26.04.2017	Integrating and final(?) insertions of material
27.04.2017	Version to Theme2 leader
27.04.2017	Minor corrections and improvements
27.04.2017	Version to internal reviewers
29.04.2017	Modificaions based on comments from interal reviewers
	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

1 INTRODUCTION

1.1 Setting the Scene

The ENVRI Reference architecture will promote the support for environmental scientists to investigate topics that require information and services from multiple environmental Research Infrastructures (RI). Where these scientists are engaged in research campaigns that need such resources, a federation of RIs will collaborate to meet their needs with a coherent and easily accessed integration of the facilities they each contribute. This will be worthwhile because the research campaigns that can be supported by one federation means there will be sufficient users to warrant construction and maintenance of the well organised multi-source facilities. The longevity of anticipated use amortises the costs of creation and management over many years. The productivity of researchers using the delivered working environment improves by one or two orders of magnitude, because they no longer have to manually manage the assembly of data from multiple sources, it is *automatically* transformed into consistent forms, the data is *automatically* migrated between computational resources and intermediate results and work in progress are managed *automatically* across the federated resources.

This step-change in productivity only becomes possible when the relevant RIs and other resource providers, such as researchers' institutions and e-Infrastructures, can deliver such integration while still meeting their other commitments and priorities. Here, the ENVRI Reference Architecture helps by providing both an intellectual framework and an evolutionary development path that RIs and others can incrementally adopt to deliver their contributions to the federation in a form that enables automated assembly and integration of resources; mapping automatically to the target coherent model. As some organisations find that they are expected to sustain engagement with several federations for decades, it is critical that their participation in delivering to them is well supported by the architecture and the reference implementations of its critical components. Many of these architectural components are being developed by ENVRIplus Theme 2, e.g. data cataloguing solutions, data analytics solutions (see subsequent sections). Others are well established from contemporary and prior work, e.g. AAAI solutions. The Reference Architecture organises their composition. It can be tailored to meet any required target configuration that is a composition of ENVRI RI resources, e-Infrastructures, private facilities and other established data and resource providers that are prepared to collaborate. Here, "any required target configuration" is defined as any conceptually coherent integration target that researchers can specify and for which mappings from available resources are logically understood.

1.2 Purpose of the Reference Architecture

Research infrastructures' ICT systems adopting the proposed ENVRIplus reference architecture will enable:

1. Clarification and progressive development of information models that meet the requirements of and are understood by the research communities supported;
2. Federation of resources, services and effort to deliver, sustain and support those information models, including conduct of scientific methods and processes required by those communities;
3. Providers of those resources to meet those requirements with an appropriate balance of stability for the majority of the work while skilfully supporting the incorporation of new requirements as the science advances while exploiting new technological opportunities.

Systems constructed in compliance with the proposed architecture will support long-running research campaigns. We note that many of the RIs themselves can be viewed in that way, so they may choose to use this architecture internally as well, when this matches their priorities and stages of development. However, individual RIs or federations of RIs will need to support research campaigns, as described in the next section. Consistently using this architecture will simplify their work and better amortise costs of construction, maintenance and support.

1.3 Envisaged user experience

A research campaign is typically long running. Even a short campaign will likely need to use datasets from previous campaigns not least to establish changes in environmental phenomena over time. It is therefore worthwhile investing thought and effort into supporting such campaigns well. Each research campaign will have a focus, the phenomena or challenges it seeks to address. It will often have leaders who develop strategy and gather resources to achieve those goals. These people will lead teams and be supported by institutions and funding bodies. Many RI are closely aligned with research campaigns to tackle scientific challenges. We give two topical examples from outside environmental sciences: (a) the hunt for the Higgs Boson (50 years); (b) The detection of gravitational waves (100 years)

A cluster of research campaigns has related goals and therefore may share information needs and collaborate on acquiring required resources and on developing methods. They may therefore have a greater duration and larger communities, and therefore be even more deserving of support. We introduce three example clusters of research campaigns that might be supported by groups of ENVRIplus RIs as well as drawing on other sources of information and services. We then examine how major roles within each of those research campaigns would exploit the benefits of the architecture. In the following section, we look at the same scenario from the providers point of view. In each section, we conclude by considering the implications for today's RIs.

1.3.1 Mitigating the impact of geo-hazard events

Initial work provides hazard maps indicating the risk of particular geo-hazards: earthquakes, tsunami, volcanic eruptions, extreme weather and floods. That can be followed by risk analysis to identify regions and structures that deserve specific attention because of anticipated impact. When such events occur, a rapid response is required, including advice to responders and follow up advice to communities and authorities for build-back-better campaigns. This draws on natural hazard models, local topological, geological, hydrological and land-use data, including the distribution of vulnerable people. It may draw on citizen data sources, social media, satellite images and rapidly deployed field instruments. Initial urgent response is followed by several years of support actions. The federation holds commonly required data, and has methods for assembling and revising the data needed for each event. NGOs and others would analyse the effectiveness of response and communication strategies using the data from multiple events.

Such a federation would draw on RIs such as EPOS and those concerned with climate impact prediction, such as IS-ENES[1] and portal, Climate4Impact (C4I), coupled with rescaling services. It would draw on metadata compliant with the INSPIRE directive[2] in the European context, but its

[1] https://is.enes.org/

[2] Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE), Official Journal of the European Union 50 (L108).

scope would extend beyond Europe, and much of the required information would come from other sources not integrated into this framework. For example, satellite, aircraft and drone imaging data, volunteer observations and inferences from public media streams.

1.3.2 Linking land-use management with climate change

Advising governments and agricultural organisations on appropriate crop choices, effects of global warming, current local farming practices, land exposure, slope orientation, altitude and surface geo-chemistry. These data include similar details describing topological, geological, hydrological, agronomy and land use. The federation will use these data with models to predict effects of climate change on pests and require climate model re-scaling services. Here again, there will be sensitive data concerning current land use and farm-management practices.

Many of the RIs engaged in the first example will be involved here, particularly EPOS and those modelling and monitoring the effects of global warming (e.g. ICOS) and anthropogenic effects. However, those involved with the biosphere and ecological models will also need to contribute data and expertise including optimal processing services. Additional feedback through mobile observation and citizen volunteers may again be important, but unpredictable urgency is not a concern here.

1.3.3 Environmental impact of mineral extraction

Predicting and measuring the environmental impact of mineral resource exploitation requires models of the Earth's surface and systems linked with local data and combined with the commercially sensitive mineral extraction plans, e.g., to increase lithium extraction from Cornwall in UK to support growing demand for smart devices and green-energy storage. This would require both atmospheric and coastal marine chemical and ecological time series, as well as the commercial-in-confidence operational data from companies proposing or conducting the extraction.

It would once again involve EPOS, but it would require more input from the environmental RIs monitoring atmospheric and marine conditions. A system that could be operated over decades would be needed. Multiple sites would deserve similar scrutiny, but existing observational and analytic infrastructure varies greatly.

1.3.4 Supporting leadership and innovation roles

Research leaders who recognise the need for a federation of information and resources and those that steer research campaigns perform crucial roles including:

1. Recognising the need for and refining the focus of a federation or campaign with clarity about its purpose and scope. This focus manifests as a *common conceptual core* that pervades external negotiation, internal discussions, computational and informational interworking and becomes an underpinning part of the culture. Giving it an explicit form that tracks developments becomes essential when inter-organisation automation is needed.
2. Publicising their federation or campaign to gather engagement and rally support.
3. Negotiating with stakeholders to agree constraints, priorities and identifiable aims.
4. Forming sustainability models and acquisition of resources and commitments for support.

5. Recruiting, motivating and leading teams matching their federation's or campaign's requirements for its current phase of development.
6. Establishing governance.
7. Engaging with user communities.
8. Developing, managing and revising plans for the construction of each phase of the federation or research campaign, and for the operation of these phases with resource allocation reflecting priorities.

These roles are all helped by an architecture that provides appropriate structures so that significant subdivisions of the overall requirement can be safely considered and developed independently. The ENVRI Reference Model (ENVRI RM) is designed to encourage the design of such a reference architecture. The ENVRI RM partitions the conceptual and technical space into substantial subdivisions in which existing standard solutions are already understood and can be imported and tailored to meet requirements in a given context.

For example, a key decision for these leaders is to agree on a consistent common conceptual model, including its concepts, relationships, terminology and attributes. In this scenario, the ENVRI RM provides a common terminology and structure to promote technology and solution sharing, improving interoperability. Additional critical attributes such as quality, consistency, precision, and coverage may be specified as well. This enables an effective strategy for importing established conceptual packages, e.g., GeoDCAT-AP[3] might be imported for the federation examples above and by contributing RIs, such as EPOS. Such packages contain substantial data and service bundles accessible through APIs. The incorporation of such packages into the target conceptual space with a transition to operations is relatively straightforward. In this case, establishing and sustaining operational use becomes a technical and engineering issue.

Establishing the common conceptual core, particularly by importing conceptual packages accelerates the path to operation and reduces operation and production costs, probably meeting major part (often more than 90%) of the user communities' needs – see section 1.3.5. However, this may inhibit innovation for two reasons:

1. The difference in effort and cost between doing 'standard' operations on 'standard' data and on doing new operations on data with new properties is now substantial, and
2. The agreements on the common conceptual core and its implementation are hard to conduct – stability is needed to help the 95% be productive and to avoid distracting negotiation.

Consequently, there are always small portion of the community are innovators, who create and develop new ideas and prove their value can be easily deterred. The architecture needs to nurture such innovation as it is key to progress in science and to addressing new challenges. This means the architecture or the framework shaped by it needs to permit a sub-campaign developing and testing a new idea, that uses new concepts and forms of data, i.e., outside of the common conceptual core, often from new sources, with new methods and forms of result. If the federation hosts and supports such innovation, delivering the normal environment, resources and data to the innovation team but allowing them to explore extensions, then they stay in its context and their innovation is easier to incorporate if it proves a success[4]. Adoption of such

[3] https://joinup.ec.europa.eu/node/148281
[4] Nascent leadership and innovation will disappear into skunk works if the federation's framework imposes awkward impediments to experiment and exploration.

innovation by the wider communities requires the leadership team to agree to endorse an extended version of the common conceptual core.

Flexibility in the architecture avoids treating the user community as homogenous. Sub-communities can develop specialised extensions of the common conceptual core without increasing the complexity seen by others. Sub-communities dynamically couple their extensions while enjoying the full support of the services and culture built on the core.

1.3.5 Supporting professional production and communication

Professional production is to support the activity of the major part of the community referred to above. It includes every step of the data lifecycle, i.e., through the five phases: Data Acquisition, Data Curation, Data Publishing, Data Processing and Data (re-)Use. It includes oversight and quality control of all the processes needed and response to requests, e.g., a rapid response to a request for a video-sequence showing impact against time from a recent intense rainfall event. The workflows need to be ready to run on resources that will be allocated promptly. All possible tasks in data management including a full provenance record should be fully automated, to save time, to reduce errors and inconsistencies and to allow professionals to focus on steering the work and evaluating the results. As the nature of events, their location and the available data is very varied the professionals will often need to quickly compose new workflows out of substantial pre-formed components, with tools that help them achieve a quality of evidence well-mapped to visualisations that will enable the targeted decision makers to understand. The runs of such workflows will cross organisational boundaries, draw on multiple sources and services, and require proper accounting and provenance record keeping according to pre-established agreements – such things cannot be set up, implanted or managed at the time by professionals delivering a responsive service. Consequently, the architecture needs to ensure that the operational framework has sufficient pre-agreements, information and automation to make such flexible production possible. For example, that framework should have provision for managing and using all the information in the common conceptual core *plus* all of the technical and administrative information needed to sustain the automated services.

1.3.6 Potential impact on RIs

To support its own community, the RI DevOps team needs to build the necessary services, preferably by importing and tailoring (usually extending) existing solutions. They also use external services, many of which have rules and APIs, over which they have little influence, these need mapping to local requirements. RI DevOps also implement algorithms and tools (which are not available elsewhere), workflows that match routine operations, and provide VREs or science gateways to support their community using their resources. As maintenance dominates lifetime cost, DevOps teams aim to minimise the software for which they are solely responsible. The ENVRI RM and the architecture derived from it should both help in identifying externally sourced subsystems and characterise those for which the R&D, maintenance and operational support can be shared or procured. The composite e-Infrastructure will grow in sophistication and capacity, and become more complex and harder to change. The managed change and maintenance of the RI's systems are required to sustain working practices as the external digital environment and available platforms change[5], to meet new requirements as their science advances. The architecture should shape the RI's e-Infrastructure to plan flexibility that will accommodate

[5] RIs and federations need to continue for decades; consider the extent of the change in ICT technologies, business practices and available subsystems in the last 25 years.

change and to have as many of the external changes as possible handled by others or automatically. This requires definition of clear interfaces between the application environment and the e-infrastructure (both within each RI and external e-Is) and also between the application environment and user interface (including VRE) systems as well as social networks in the widest sense.

We now consider how this changes as the RI decides to support a new federation. The RI's DevOps team are already fully committed to the work just described. The RI's leadership needs to decide whether to change priorities or to bring in extra staff. It then needs to negotiate with the federation's leadership over its requirements and how far they can be met by the RI. Such negotiations then need to be mapped to operational procedures, such as ways of requesting data, ways of sending data, mapping from source to target representations, where tasks are performed, mechanisms for accounting and mechanisms for detecting and handling failures. Thus, both leadership and technical staff become involved – a drain on scarce resources. As the RI improves its available information over time the federation will want to benefit. As the federation's requirements evolve it will seek improved contributions from the RI. Therefore, the call on expert time *is not a one off*.

If the RI is engaged in more than one federation, as EPOS is in the three federations envisaged above, then this management and technical effort is called on repeatedly. If the requests cannot be made consistent they lead to multiplied long-term commitments that combine to have a serious impact on productivity, agility and reliability.

The architecture should introduce patterns of consistency that are adopted by multiple RIs and multiple federations. This will greatly simplify the negotiation, agreements and technical planning. Much of it will already be available as easily adopted and operated packages that are maintained by a large and diverse community. In this framework, it will establish tools and automation that greatly help the DevOps team meet their local goals and offer external services with very little additional effort. For example, the tooling and automation can draw on comprehensive catalogues that organise the common conceptual core and the required technical and administrative information. This will include metadata that – among other things - will shape automated translation services to deliver data in agreed forms. Such automated translation has to be two-way, as incoming requests and data need translation into the locally established forms, and outgoing responses need to match external requirements.

1.4 Envisaged asset provider experience

Providers are any organisation or part of an organisation (including of the RI itself) that enable RIs and federations to operate so they can deliver basic capabilities, such as data transport, data storage, and computation, or more sophisticated services, such as identity management, catalogue management, file services, database services, VRE support, provenance management, workflow coordination, or administration of long-running campaigns. Architectural considerations and long-term cost benefit analysis pushes the operational frameworks towards using these sophisticated services. Relevant standards bodies such as W3C (Wold Wide Web Consortium[6]) and working groups in the RDA (Research Data Alliance[7]), will have informed the creation of those services, so that they can be easily combined and tailored to meet particular

[6] http://www.w3.org
[7] https://www.rd-alliance.org/

needs. The architecture needs to take into account such prior work. It then needs to extend consistency and compatibility to meet needs discovered through co-design and co-pioneering with each RI and federation. The architects should accumulate that experience and distil it into wisdom that guides future solution strategies. They should continuously engage with providers and those standardisation agents to improve long-term solutions which will be adopted widely.

The architectural agency here offers three benefits to providers:

1. Provide access to already agreed APIs and modes of use, encouraging their uptake by explaining how they map into the framework being developed or already operational.
2. Consider and merge new requirements, so that there are fewer external negotiations required for providers, and these are more technical, consistent and compatible[8].
3. Contribute a wider vision and experience when issues are encountered, resolving incompatibilities or developing a new group of services.

The Engineering Viewpoint of the ENVRI RM - currently ongoing work - partitions this complex space into tractable and related parts, and provides a means of recording precisely the agreements, e.g., as sophisticated service APIs and behaviours that are consistent with their definition. The interplay between architectural discussions and the ENVRI RM is the foundation for such definitions, as both cross organisational and discipline boundaries. The ENVRI RM provides a uniform framework with well-defined subsystems of components specified from different complementary viewpoints developed by considering the data lifecycle in multiple RIs. It thereby delivers a precise vocabulary and structural framework for RI architectures. The architects develop more detail and interpretations as they propose solutions in multiple RI contexts.

When providers support RIs and federations they prefer to do so through consistent multi-purpose APIs and well-rehearsed patterns of use. In most cases, i.e., for most working practices and operational loads within RIs and federations, this will perform sufficiently well, and will be tailored and composed by workflows to meet requirements. Very occasionally a special arrangement will be needed to reach a required capability. Providers need to provide:

1. Training for a range of categories of users, including novices and various kinds of expert.
2. Supporting the selection of providers and setting up mechanisms for using the provided facilities, for the RI DevOps teams.
3. Handling diagnostic and performance issues and difficult queries from users, helping the RI DevOps Team.
4. Supporting standard record keeping and accounting sufficient for provenance and campaign management tools.
5. Providing metadata to describe the services and data they handle with automated production of metadata wherever possible.
6. Supporting distributed workflow enactment.
7. Supporting automated handling of changes in the provided service – as far as possible.
8. Sustaining reliable delivery of the service with continuous maintenance.

Inevitably, for many valid reasons, software and systems that do not comply with agreed standards or have such support, will be actively used, e.g., because of software and practices developed earlier being brought into the RI or federation. The associated information representation, methods or working practices may be deeply ingrained in a community's culture,

[8] Of course, major consumer RIs or federations with a substantial load, e.g., EISCAT-3D for computation and data storage, may still engage in direct negotiation.

e.g., through education, training, documentation and publications. Changing important and frequently used elements of the research environment has a high cost. A cost-benefit analysis, conducted in consultation with stakeholders and users must precede change. The architecture provides a context for such analyses, making it easier to localise change and to estimate disruption. If a decision to go ahead is agreed then the architecture should reduce the change's implementation cost and help define a path where a succession of such changes is feasible if, and when, a community commits to such changes. Just as architects would be consulted when a major change to a building is contemplated, the ENVRI architects may be consulted when an RI or federation contemplates a significant change. One strategy they may be able to devise, is to move to new underpinning services that are standard and supported with an automated mapping to and from the old forms until the culture evolves (i.e. backward compatibility). When consistently developed by architects engaging with multiple RI contexts and guided by the generic RM, this should mean that delivery of such user-accommodating and research-sustaining change draws on common solution strategies and software.

1.5 Recap and overview

We have outlined above the 'philosophy' behind the development of the proposed architecture. It envisages RIs cooperating with each other and a central architectural team (at least partly drawn from RI ICT staff) for ENVRIplus to maximise benefits and minimise costs. It envisages use of e-Infrastructure within each RI and also external e-Is as appropriate. Indeed, for cost-reduction (and in some environments privacy and security) it is recommended to do as much processing as possible within a RI e-infrastructure utilising existing investments in infrastructure and scale out to using external e-Is when resources become insufficient for the task or if so doing is demonstrably more economic. It addresses sustainability aspects including curation and provenance. It brings together not only technical architectural solutions meeting the requirements of D5.1, but also references the interplay between architecture and governance, financial and legal aspects. This approach to ICT architecture is expected to provide an appropriate architectural basis – to be further developed as the project progresses - for ENVRIplus.

1.6 Technical summary

The key features needed to achieve such ICT-supported research federations are:

1. A *metadata notation* that is tractable to the federation architects and which supports automated maintenance of the relationships that are needed as properties of the target, as constraints by the providers or for comprehensible machine-to-machine communication about what is available and how it can be used.
2. For any target integration, a *specification of its canonical conceptual (technology type independent) and logical (technology type independent but implementation dependent) schema* and its mapping to practical representations in the above notation. The canonical schema is an agreed and adopted definition of the universe of discourse that enables communication between the participants. Each participant will normally have additional information. Mechanisms for agreeing and revising this canonical schema are crucial. It will start with an essential core that grows. It needs to be comprehensible to the researchers who shape it and to those who use it. This will depend on simplicity and on presentation. As it grows, diversity in its community of users may be supported by views that are relevant to identified subgroups.
3. Mechanisms to allow innovators and evaluators to use *extensions to the canonical schema*, that that are localised to their own context, to allow specialised subgroups to

thrive and still connect with and use their federation's resources and model. One source of extension to the canonical schema will be ingest of successful extensions.

4. *Descriptions in the agreed notation of resources, services and data*. These will be presented in comprehensible form to humans, and support search, automatic use and automatic mappings to and from targets.
5. *Tools and computational support* for creating, preserving and revising all the above descriptions.
6. *Systems for managing these descriptions*, for finding them, and for transporting such descriptions to components or users who use them. This will be organised around catalogues, and include description movement, description interpretation, description version identification and description preservation.
7. Provision of *well-engineered work environments tuned for each category of participant*.
8. *Actions and processes* initiated, overseen and organised by practitioners including – from the canonical schema – generating representations to meet the standards of external federations of which a RI is a part.
9. *Mechanisms implementing those actions and processes* by distributing them to the resources provided by a federation that are needed to perform all necessary controls, mappings, accounting and recovery. One such mechanism is a workflow. They will need to accept work in terms of the target schema.
10. T*ools for managing all such actions and processes*, to examine them for validity, diagnosis and cost, to organise user-controlled data classification, and all steps in a data management life cycle. These tools will be drawing on the integrated platform and present all interactions in terms of the target model.

2 METHOD USED TO DEVELOP THE REFERENCE ARCHITECTURE

2.1 Introduction

The targets for the ENVRI architecture approach presented in Section **Error! Reference source not found.** include diversity and competing demands that need to be resolved by a well-structured dialogue. This dialogue needs to include:

1. Representatives knowledgeable about the RIs' research priorities, their resources and the constraints and commitments under which they and their communities operate.
2. Leaders of system and software engineering teams who understand costs and can commit resources.
3. An architectural group that develops a comprehensive understanding of the technical and operational requirements and options, sustains the discussion on a productive path and facilitates identification of solutions and development paths.

This proposed Information Systems Strategy and Engineering Group (ISSEG) has to balance representing those three viewpoints well with the focus to develop a practical, useful architecture quickly, which is then widely adopted. This requires the ISSEG to have authority. It will develop this by:

1. Integrating and distilling requirements to maintain a prioritised list of architectural issues.
2. Surveying and analysing the evolving digital-technology context to recognise options that will have sustained and supported value.
3. Gathering and interpreting current experience from operational practice and agile development campaigns.
4. Periodically assimilating an integration of the ENVRI RM, current technological options and proposed operational behaviour, called *the reference architecture,* that can be recommended to the RIs and those using compositions of RIs as it is currently the best way of proceeding. It will therefore take into account previous reference architectures to avoid unnecessary disruption and balance long-term benefits with transition costs.

This report presents a first iteration of that reference architecture. The ISSEG should normally issue a revision yearly. This will require effective information gathering via steps 1, 2 and 3. That in turn will require effective communication with those developing e-Infrastructure, developing RIs and researchers in each RI, particularly in the cases where new aspects of the architecture are being adopted. It will also require breadth of coverage and experience in the architectural and research issues. The ISSEG will develop this experience and provide an effective forum for pooling ideas and sharing solutions across RIs.

This deliverable relies on input from D5.1 defining the relevant state of the art and the requirements of ENVRIplus RIs [Atkinson et al 2016] and D5.4 [Jeffery et al 2017a] providing the approach to the architecture. However, it also relies on deliverables from and 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. To generate the reference architecture the RM needs to be enhanced with details of the engineering and technology viewpoints. This work is proceeding in parallel with the production of this deliverable and is being positively affected and reinforced by it.

The requirements in D5.1, of course, are derived from inputs elicited from all RIs in ENVRIplus. This deliverable represents all those condensed views, after some rationalisation of differences and within the constraints of best architectural practice.

The deliverable is aimed at the specification of a reference architecture[9] for new RIs and as a target with which existing RIs may align over a period compatible with their commitments and priorities.

2.2 Requirements of the Reference Architecture

D5.1 and D5.4 conclude that the key elements required in the architecture are:

1. Catalogs of assets to be used by software services in providing offerings for end-users including within a RI community and interoperating across RIs;
2. A set of software services of two kinds: (a) common services that are recommended to exist and be used in all RIs with consequent cost-savings by re-use of services developed once; (b) cross-cutting services that provide the interoperation capability between and across RIs;
3. Capabilities to scale out/in and up/down as any RI requires increased computing and/or storage capacity and resources beyond that within the RI – this implies use of e-Is or external services;
4. Capabilities to interface to a VRE (Virtual Research Environment) providing the end-user not only with portal access across RIs but also appropriate communication and research management capabilities.

All RIs agree that a catalog of assets (with exposure of relevant assets to any federation or end-user request) is necessary although RIs have varying standards and practices for their catalogs. 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. This catalog or catalogs (however implemented – whether physically realised or acting as a reference specification) must 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 ICT support and in particular with independent evolutionary paths to date. A fundamental objective of ENVRIplus is that if RIs can agree upon the architecture for – and share expertise and the cost of development of – common operations and cross-cutting services (including those concerned with catalog interoperation) then (a) the research communities benefit from better systems and interoperation; (b) the cost of ICT systems support and maintenance for each RI is reduced; (c) it is possible to interoperate across the RIs so encouraging new research based on multidisciplinary science. There is of

[9] The reference architecture is the architecture recommended for use in and between RIs within ENVRIplus in order to achieve the cost-benefit associated with use of common and cross-cutting services. It is based on the information and computational viewpoints of the RM reified by the engineering and technology viewpoints.

course a cost in this development and in subsequent sustainability, so a business plan is required. This also relates to curation of the RI assets and the subsystems to permit interoperation using the rich canonical metadata catalog.

The reference architecture may be implemented:

1. At all RIs to gain advantage from common and cross-cutting services but at the expense of maintaining at each RI a catalog representing all ENVRIplus RIs and not just the local RI;
2. At a central (to be defined) organisation providing common and cross-cutting services which individual RIs may download and use but with the ENVRIplus catalog and associated services centrally maintained;
3. Any of a set of possibilities between these two extremes. The reference architecture is designed to permit any of these configurations since the choice depends as much on governance, financial and legal aspects as technical.

2.3 The Overall Relevant ICT Environment

ENVRIplus has considerable complexity. In order to design an architecture, it is first necessary to understand the environment (or ecosystem) of organisations, their objectives / competencies and offerings. Theme2 of ENVRIplus has proposed (D5.4 [Jeffery et al 2017a]) a view of the overall environment (Figure 1) based on the ENVRIplus RI requirements, the ENVRIplus partner competences and taking into account current and proposed developments in ICT architecture.

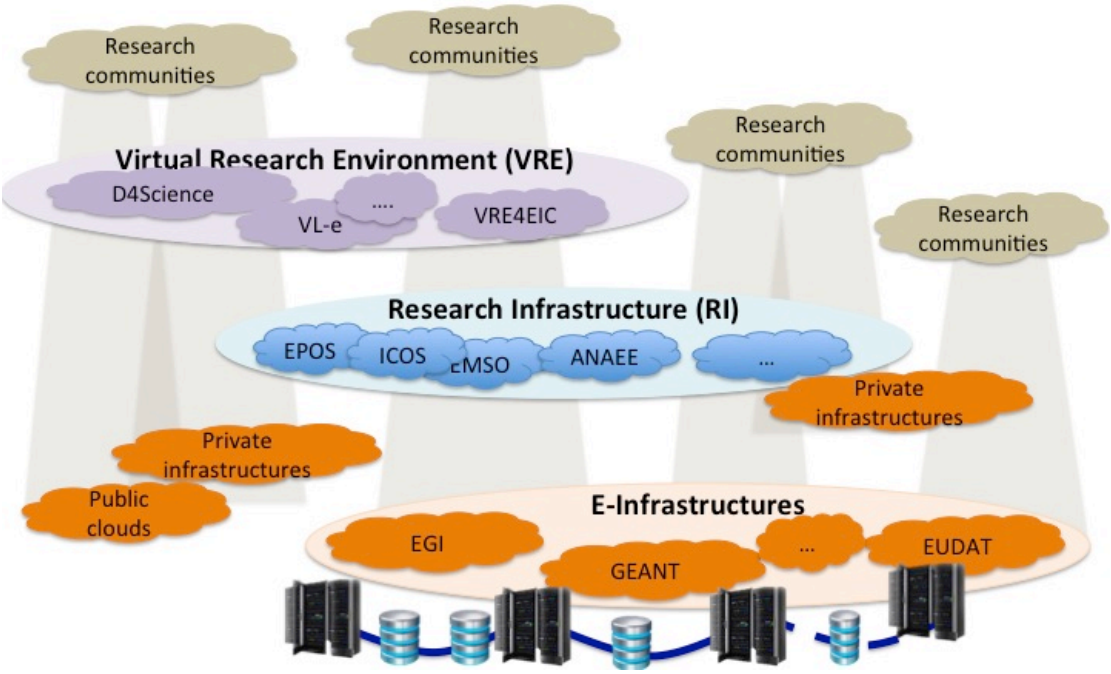

FIGURE 1 : A VIEW OF THE DIFFERENT KINDS OF RESEARCH SUPPORT ENVIRONMENTS AND THE ROLE THEY PLAY IN ICT ACTIVITIES REQUIRED BY USER COMMUNITIES.

A few RIs rely to a greater or lesser extent – depending on their own provision of ICT facilities - on underlying e-Infrastructures providing basic services of networking, computing platforms, data storage facilities and open access to research publication outputs. Examples are GEANT, PRACE, EGI, EUDAT, OpenAIRE and the emerging EOSC (European Open Science Cloud). The EOSCpilot[10] project (starting January 2017) may attempt to cover some of the wider issues beyond e-Infrastructure including some form of integration. 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**Error! Reference source not found.**):

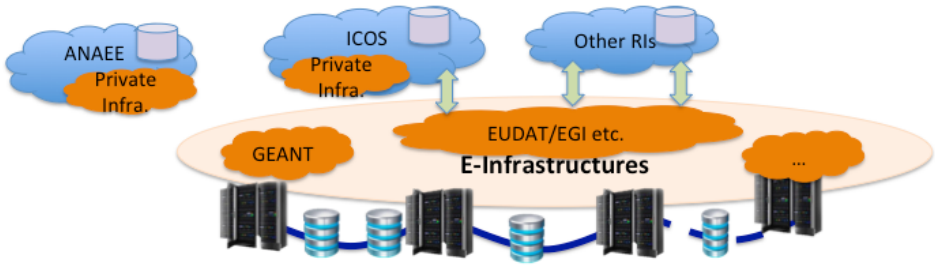

FIGURE 2 SIMPLIFIED VIEW OF TYPES OF RIS IN ENVRIPLUS

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 with 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 convert (convertor pairs) with every other RI (the well-known n^2 problem described in D5.4) or, alternatively, each RI converts to/from a conceptual canonical superset metadata catalog (or limited set of catalogs) reducing the convertor pairs (i.e. the software services to be developed) 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 convert between the local metadata standard and the canonical standard and furthermore to be able to convert the RI assets (especially datasets). The effort required to achieve this, once the scope and form of the canonical target has been agreed,

[10] https://eoscpilot.eu/

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**Error! Reference source not found.**).

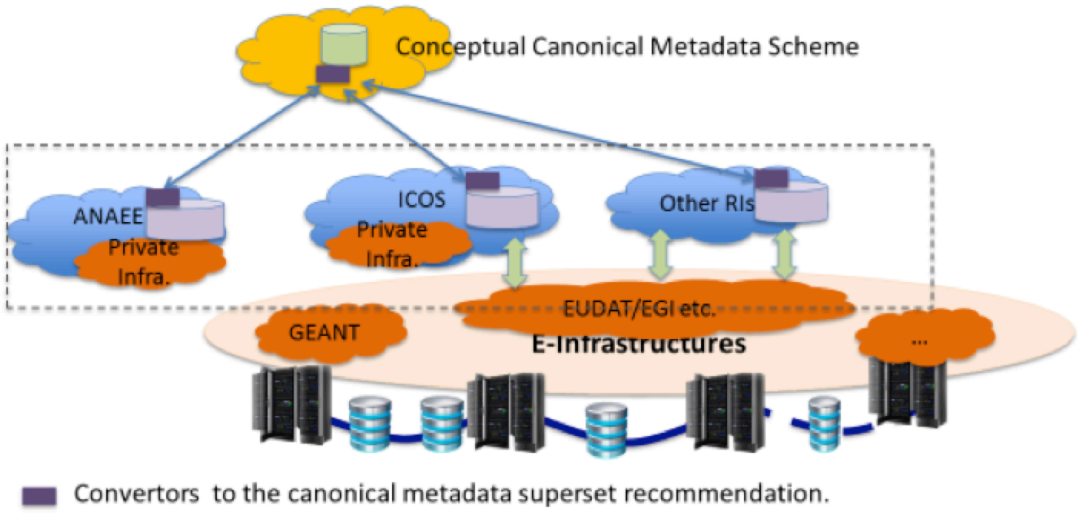

FIGURE 3 USER AT ONE RI USING ASSETS OF ANOTHER RI

The conceptual canonical schema may be used in two ways. It may be used so that conversion is always at the RIs. The advantage is that the associated local catalog (from which conversion is done to the canonical catalog for interoperation) is current. This is the 'distributed query' approach. Alternatively, the canonical catalog may be realised physically – either at each RI or at some external location. The advantage of this approach is that each RI has a complete picture of the assets and access conditions at all RIs – thus allowing for optimal deployments. The disadvantage is maintenance of the physical canonical catalog.

For truly interoperable access one further architectural step is needed making the conceptual canonical superset catalog(s) physical which can be provided by each RI, or by an external environment of an independent provider. In addition to RI-RI interoperation, third party (i.e. users not belonging to any particular specific RI - for example a citizen scientist or policymakers) require homogeneous access to one or several RIs via, e.g., VREs (Virtual Research Environments), some portal systems, or dedicated services. Existing RI user 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). 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 far 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.

Recommending a superset (possibly incrementally and evolving) 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.

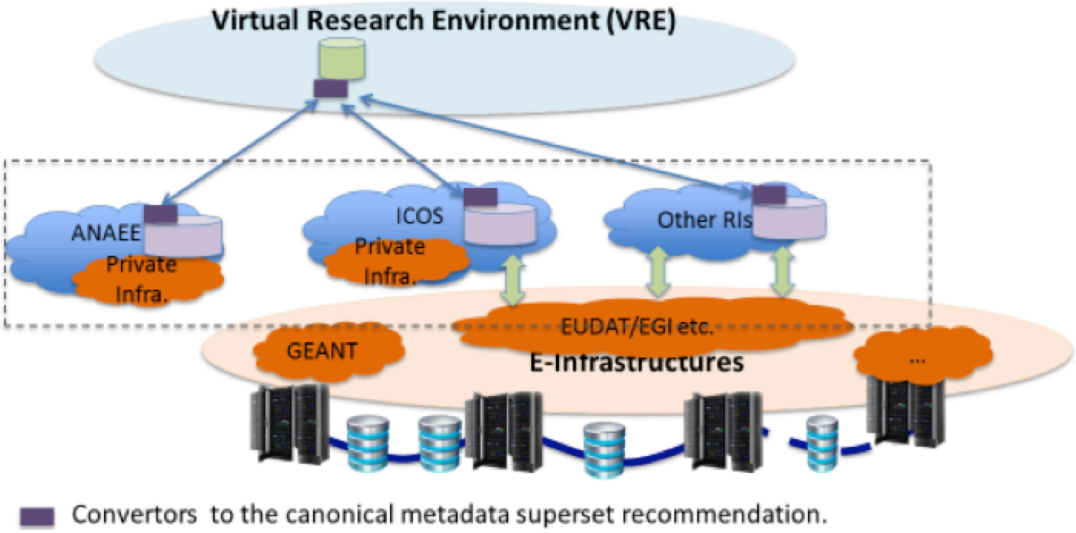

FIGURE 4 EXTERNAL USER ACCESSING MULTIPLE RIS VIA A GENERIC VRE

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/respected. 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 – or federal agreements affecting he architecture - emerge in their own disciplines. Of course, this is mitigated by good architectural design – allowing evolution with new services while maintaining backward compatibility for interoperation.

It is necessary (to ensure utilisation of the best concepts) 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 in various formats and with some common metadata 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[11]. Similarly, ongoing work in Australia (virtual laboratories) also provides portal services linked with a toolkit of generic processing services. However, for effective integration (a) a rich canonical metadata standard is required with convertors to the commonly used metadata standards; (b) the data should be convertible to a minimalist set of data formats to reduce the number of coversions required when interoperating.

[11] https://www.rd-alliance.org/groups/metadata-ig.html

It should be noted that a simple level of interoperation can be achieved with OpenSearch[12]. 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 particularly for Geo aspects) 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.

2.4 Method

All the above leads to a stepwise (but agile and spiral) method to adoption of the ENVRIplus architecture by the RIs.

1. Define the catalog (D8.3) to encompass the requirements from D5.1 and the design work in T8.2 and decide if it is conceptual or realised physically;
2. Define the common and cross-cutting services; this is derived from the RM modified and updated by considering the requirements of D5.1 including (a) interfaces to support pre-existing federations of which ENVRIplus RIs are members; (b) interfaces to external computing and storage capacity (e-Is and external platforms); (c) interfaces to VREs; and further developed in the engineering and technical viewpoints of the RM;
3. Test the components individually;
4. Integrate into a platform and test;
5. Implement in minimally 2 RIs and test;
6. Release for adoption widely among ENVRI RIs.

Note that any of the steps may be (and should be) repeated in an agile, spiral development method.

The role of the semantic linking activity in ENVRIplus is to explore the landscape of controlled vocabularies used for structuring (meta)data across all environmental science research infrastructures and provide tools and methods for browsing that landscape and creating mappings between overlapping vocabularies. The Reference Model for research infrastructures is seen as being a key contributor to the activity insomuch as the modelling of RIs' architecture and information flows permits the contextualisation of the different vocabularies as well as the research entities that use them or are described by them.

With regard to the architectural recommendations provided by this document, the semantic linking framework should assist in the creation of adapters from RIs' canonical conceptual and logical schemas to the metadata notations required for building the federated catalogues used for cross-RI research asset discovery and retrieval (see Section 2.3). It should also provide material support for analysing the current state of the art regarding RI metadata and modelling by providing a high-level schema for encoding that information (i.e. it should support the uploading of RI descriptions built using the Reference Model into a knowledge base to permit query and analysis). An additional contribution is to the optimisation of the use of e-Infrastructure by RIs or by users working through VREs, by providing a vocabulary for describing requirements (e.g. for quality of service or access control) on user-defined workflows involving e-RI or e-I services that can be translated into constraints on e-I platform provisioning.

[12] http://www.opensearch.org/Home

3 REFERENCE ARCHITECTURE

3.1 Introduction

This deliverable does not provide a conventional architectural specification. ENVRIplus architecture development is continuous; this deliverable documents the current state. Requirements have been collected (D5.1) and an approach to the architecture defined (D5.4). The scientific (enterprise), information and computation viewpoints of the RM have been defined and are in parallel undergoing improvements. The engineering and technology viewpoints are now being tackled in parallel with the production of this deliverable. Hence, D5.5 sets the scene and records the current progress. It also indicates the direction of travel including integration of the engineering and technology viewpoints of the RM and provides a plan for the future architectural development, adoption and implementation closely linked with the activity of WP9.

The architecture depends on several 'pillars'. One is the defined requirements collected from the RIs (as D5.1). One is governance and policies in the RIs and hence in ENVRIplus: in data management, use of unique identifiers, security and privacy, metadata standards. One is current best practice in the ICT profession: the best systems development methods, the optimal languages to use; the recommended interfaces (usually based on standards) supported by international experience. A successful architecture has to blend the demands of these 'pillars' to produce an acceptable and useable design. ENVRIplus does not have WPs or Tasks relating to some of the architectural components required as indicated under the various 'aspects' below.

3.1.1 Data Management / Lifecycle Aspects

The architecture needs to support data structures and operations dealing with all phases of the data lifecycle, including curation and publishing (to ensure availability) and provenance (to understand the derivation and to assess relevance and quality for a purpose). This relates to tasks T8.1 and T8.3

3.1.2 User Interface Aspects

The way in which the wealth of assets in ENVRIplus RIs is presented to end-users is of critical importance if uptake and utilisation is to be achieved. Currently most RIs have user portals – (varying from simple lists of URLs of assets to portals with some VRE capability) but the ENVRIplus project should provide a user interface facility either to be adopted by all RIs to allow interoperation or to be hosted at some external facility providing a view over all RIs. The required capabilities of such a user interface facility are not dissimilar from those offered by online retail marketplaces such as Amazon, where various merchants' products can be browsed, items selected into a basket, and checkout to pay for (download) them.

3.1.3 Identification and Citation aspects

Acknowledgement and attribution of research contributions is vital and citations, in addition, may be used as a measure of quality and impact. Citation depends critically on identification of the digital object being cited (which should be atomic or a collection of atomic objects and with appropriate versioning). This relates to WP6 work.

3.1.4 Catalog Aspects

The catalog provides the 'view' of ENVRIplus RIs and their assets as decided by the RIs themselves by the way in which they expose their own catalog information to the canonical catalog (with appropriate conversion). This is not unlike the way hotels expose their information

to a hotel booking portal system (such as booking.com[13]), or airlines their flights to an airline booking system (such as Opodo[14]). The recommendation from T8.2 (D8.3) is to construct the canonical (interoperation) catalog in both CKAN[15] (as used in EUDAT[16]) and CERIF[17]. This implies construction of convertors from the local metadata standards in the individual RI catalogs to these standards.

3.1.5 Data Processing/Analytics/Simulation/Visualisation Aspects

Processing/analytics, simulation and visualisation services are required by all RIs. If all RIs used the same software services there are advantages in interoperation and costs. However, RIs are likely to also have specific requirements in this area, which are outside of any common services.

The large heterogeneity characterising data analytics both in terms of existing technologies and solutions as well as in terms of understandings and expectations from environmental RIs has been largely discussed in D7.1 [Candela et al. 2017]. When devising data processing solutions for ENVRIplus, the following aspects emerged: (i) technology should be "ease of (re-)use", i.e., it should not distract effort from the pure processing task; (ii) the "as-a-Service" provision mode should be preferred to the "do-it-yourself", i.e., scientists should be provided with an easy to use working environment where they can simply inject and execute their processing pipelines without spending effort in operating the enabling technology. This makes it possible to rely on economies of scale and keep the costs low; (iii) solutions should be "hybrid", i.e., it is neither suitable nor possible to implement one single solution that can take care of any scientific data processing need; (iv) support for research developers who produce and refine the code and workflows that underpin many established practices, scientific methods and services should be provided; (v) support for operations teams who need to keep the complex systems within and between RIs running efficiently as platforms change and communities' expectations rise while funders become more miserly should be provided; (vi) support for scientific innovators (playing with ideas, working on samples in their own favorite R&D environment, and then testing their ideas at moderate and growing scale) should be provided; (vii) as much care as possible must be invested in protecting researchers working practices by smoothly injecting novel approaches enacting them to perform their daily tasks.

It is also possible to make use of more generic distributed services provided by e-Infrastructure, but only if such services provide sufficiently standard interfaces and are adequately catalogued such that it is possible to insert them into an application workflow with minimal engineering effort. In particular, there is considerable scope for the use of real-time 'big data' analytics frameworks such as Apache Storm or Spark for a range of data processing applications on the Cloud (or Cloud-like scalable virtual environments), but only if the effort of setting up and running these frameworks is minimised. Thus, the architecture for interoperable RIs should involve publication of service specifications for services hosted on e-I platforms in order to facilitate the construction of workflows involving invocations of those services (either as part of the workflow or as part of the deployment of the workflow onto e-Infrastructure) whether those workflows are produced internally by the e-RI or directly via a VRE.

[13] www.booking.com
[14] www.opodo.com
[15] https://ckan.org/
[16] https://www.eudat.eu/
[17] http://www.eurocris.org/cerif/main-features-cerif

3.1.6 Workflow Aspects

In general, a user request to ENVRIplus may be regarded as a workflow. Constructing a workflow to deal with a user request is not simple. For example:

The end-user request must be understood and validated. This uses the canonical metadata catalog to ensure availability of appropriate assets (under appropriate conditions of re-use), their relevance and quality. The workflow to meet the user requirement is then composed (ideally automatically but in practice by the user with progressive automation assistance). The workflow can be extremely simple (such as download a dataset to my computer) or complex (requiring access to 5 e-RIs (e-Research Infrastructures, the digital representation of a RI) for various selections of services and data, those and another 3 e-RIs for 'best of breed' open source software components, harmonising the data to a canonical standard, composition of the software components, deployment of the workflow in distributed and parallel fashion cross 3 e-I platforms and if necessary further data collection from instruments/detectors). The workflow is deployed physically and monitored. The deployment might be distributed, partitioned and parallel depending not only on platform availability, performance and cost but also data locality and data transport costs, all constrained by the user requested parameters. Non-functional requirements including rights, privacy and security also influence the (re-)deployment, not least in maintaining European citizen privacy by utilising only European platforms for personal data and of partitioned datasets for privacy and security. The deployed workflow's products are catalogued with metadata (including provenance and version information) and curated for future re-use. The workflow as a physical deployment may be modified during execution to maintain the SLA (service level agreement), including QoS (Quality of Service) parameters. Parts of the workflow may execute asynchronously in different places. The workflow is also stored for later re-use and re-deployment.

The efficient use of e-I platforms requires that the workflow request be analysed and decomposed such that user-level QoS requirements and RI-level data dependencies are recognised and translated into infrastructure-level constraints on the selection, provisioning and runtime use of resources provided across e-Is. There should be a mechanism by which requirements are 'filtrated' through the layers of VRE, e-RI and e-I in order to ensure that e-I services are able to optimise their behaviour in the best possible way.

Note that such services are part of the e-I platform, and therefore should be provided to the RIs in a way that does not require intimate knowledge of them, but rather lets the RIs (or in some cases the VRE) pass on requirements and let the e-I platform determine the best way to service them. Thus, the main requirement for RIs is not to explicitly build compatibility with specific optimisation services, but rather to be able to deliver their requirements according to some standard interpretation and to be able to interact with a single standard broker for facilitating the final workflow. That broker should be able to accept a generic workflow specification from a VRE or e-RI, and return it with additional properties describing the specific services and resources (to be) used to realise the workflow on the provided platform.

3.1.7 Deployment Aspects

As indicated in the examples above there are aspects of deployment that are not necessarily simple. The need for dynamic scaling (out/in and up/down) during execution and the need for possible redeployment to maintain the SLA and QoS parameters are challenging; as is the ability to handle coordination of different parts of the workflow executing asynchronously

The (re-)deployment of application/experiment workflows by or via RIs depends inherently on the underlying e-I platform, and whether that platform is public or private to the given RI. The use of public/leased e-I resources by RIs is useful for easing pressure on individual RIs and permitting those RIs to offer additional services (particularly for data analytics/processing), but the efficiency and efficacy by which those resources are used determines whether such additional services will be ultimately exploited by more than a few technical enthusiasts. The cost model also bears impact on the future sustainability of such services; after project funds have dried up, only services deemed especially valuable or essential by research communities will be given more than passing attention.

An increasing body of development exists to make efficient use of Cloud services offered by e-Infrastructures. The resulting tools and services are potentially at the disposal of RIs, but require integration into the RI's operational practices, though as mentioned previously this can be done 'at arm's length' by using intermediary brokers capable of matching QoS requirements to infrastructure properties and behaviour patterns (e.g. trigger conditions for dynamic scaling). As mentioned previously, the filtration of requirements so that they are accessible by such tools at the infrastructure level is also essential.

From the above, it is clear that aspects concerning user interface, workflow construction and deployment are outside the current workplan of ENVRIplus. Similarly, trust, security and privacy are not the subject of a given work package or task (although it is a relevant consideration in many). However, the catalog formats and the common and cross-cutting operations are defined. Together these should also cover curation and provenance (WP8), the parameters for trust, security and provenance (to be used by an AAAI (Authentication, Authorisation, Accounting Infrastructure) and processing (WP7).

3.2 RI Capabilities Required

This section defines the capabilities the RIs expect to have. The ENVRIplus architecture will specify them, and their development – influenced by the RM engineering design patterns – will be a joint activity between the RIs (i.e. staff from the RIs do development work ideally jointly with other RIs and the ENVRIplus Theme2 team) and the Theme 2 ICT team.

3.2.1 Common Services

The common operations are specified to cover the RI identified requirements for services operations that should be implemented in all RIs to provide efficient and effective processing in each RI. Whilst recognising that RIs potentially can support a wide range of functions (as illustrated in Figure 5) those services highlighted in Table 1 represent the minimum subset that are initially considered as essential to provide operations to support the main aspects of the data lifecycle.

FIGURE 5 COMMON OPERATIONS OF RIS, INDICATING MINIMAL ESSENTIAL SET

TABLE 1 COMMON SERVICES TO BE SUPPORTED BY RIS

	Service	***Description***	***ENVRIplus wp.task***
1	PID	Provides globally-readable persistent identifiers (PIDs) to infrastructure entities, mainly datasets, that may be cited by the community.	6.n
2	Catalogue	Concerned with cataloguing of metadata and other characteristic data associated with datasets and other assets stored within the RI ICT environment.	8.n
	Annotation	Concerned with updating of records (such as datasets) and catalogues in response to user annotation requests.	Note
Note: Annotation service is lesser priority than the other services listed in this table and in Table 2.			

Ideally, each service listed in Table 1 must be fully specified in terms of the service offered to the user (stage 1, the overall service description); its logical parts and information flows (stage 2, the functional capabilities and information flows); and the operations / responses syntax needed to support those functional capabilities and information flows (stage 3, exchange formats and APIs technical specifications). This maps to RM components and operational interfaces (APIs in the Engineering Viewpoint) presently being specified in parallel work as the engineering design patterns.

Note: Adopting a community approach to service specification (like the 3-stage approach outlined above) engages the relevant stakeholders and provides the best available means to reach agreement and consensus on the scope and content of needed services. At present, there is no such coordinated approach with the ENV Community, leading to technical conflict, duplication of effort, re-invention and overlap. In the long-run, this precludes achievement of interoperability and reduces maintainability and sustainability.

3.2.2 Cross-Cutting Services

The cross-cutting operations services are specified to cover the RI identified requirements for services that should be implemented in all RIs to provide efficient and affective effective interoperation across RIs. Table 2 lists those considered as essential minimum needed.

TABLE 2 CROSS-CUTTING SERVICES TO BE SUPPORTED BY RIS

	Service	***Description***	***ENVRIplus wp.task***
	AAAI	Handles authorisation requests and authentication of users before they can proceed with privileged activities.	Note
3	Data transfer	Concerned with movement of data into and out of the RI.	7.n
4	Coordination	Delegates processing tasks sent to specific execution resources; coordinates multi-stage workflows and initiates execution.	7.n
Note: AAAI service is of wider interest than just the ENVRI Community and is probably out of the direct scope of the ENVRIplus project. ENVRI Community should adopt the standard mechanisms that emerge, such as FIM4R and should contribute to ensure ENVRI Community needs are met.			

3.2.3 Identification and Citation

Persistent unique identification of digital objects is essential to permit accurate (de-)referencing and access, but also to permit construction of linkages between digital objects. These linkages may have various semantics (such as 'is-part-of') but also more complex relationships such as citation. Citation may simply be a linkage with semantics <object identified by UIDa is cited by object identified by UIDb> but in practice it is more complex. The citation may be to a specific part of the object – even down to an individual sentence in a document or data value in a dataset. Worse, documents and datasets evolve (hopefully tracked by provenance) and in so doing it becomes necessary to reconstruct the state of the digital objects involved at the time the citation link was made. Work in RDA has proposed a mechanism using queries but this – in turn – requires datasets to have – or be enhanced with – temporal information.

There are many forms of persistent unique identification. In the document world DOIs are used. In semantic web environment URIs (or permanent URIs) are used. In the database world keys are used (and their unique integrity maintained by the software). There are numerous other systems and environments for creating and managing persistent unique identifiers such as the handle system[18].

[18] https://en.wikipedia.org/wiki/Handle_System

There are two not-fully-solved problems with object identification: (a) atomicity (i.e. what is the granularity of the object to be identified) and associated (de-)composition; (b) versioning. Ideally (b) is handled by provenance information (although practically such metadata rarely is provided). (a) can only be handled with reference to the practices of the community which may prove problematic for cross-community (interdisciplinary) research.

3.2.4 Catalog of Metadata

T8.2 has defined the formats for the superset canonical catalog as CKAN (as used in EUDAT) and CERIF. RIs are expected to provide services to convert to/from these formats from their own local metadata formats (cross-cutting services). The Theme2 team will assist in specifying the matching/mapping and developing the convertors.

3.2.5 Analytics/Simulation/Visualisation

D7.1 contains a specification of a data analytics platform [Candela et al. 2017] with certain characteristics. In particular, that platform was conceived to (i) *be extensible*, i.e., the platform is "open" with respect to the analytics techniques it offers / support and the computing infrastructures and solutions it relies on to enact the processing tasks; (ii) *promote distributed processing*, i.e. the platform executes processing tasks by relying on "local engines" / "workers" that can be deployed in multiple instances and execute tasks in parallel and seamlessly; (iii) *be offered by multiple interfaces*, i.e., the platform offers its facilities by both a (web-based) graphical user interface and a (web-based) programmatic interface (aka API) in OGC WPS; (iv) *cater for scientific workflows*, i.e., the platform is exploitable by existing WFMS as well as should support the execution of a processing task captured by a workflow specification; (v) *be easy to use*, i.e., the platform is easy to use for both algorithms / method providers and algorithms / method users; (vi) be *open science friendly, i.e.*, the platform transparently inject open science practices (provenance recording, repeatability) in the processing tasks executed through it.

ENVRIplus is not planning to offer specific facilities for data visualization. However, it is possible to implement certain typologies of data visualization by specific methods of the data analytics platform.

3.2.6 APIs

The specification of the common and cross-cutting services includes also a specification of each API. Current best practice guidelines suggest that such APIs are simple, uniform and deterministic, with low atomicity i.e., they each represent one (micro)service that does one thing and one thing only.

In the case of the data analytics platform every analytics method integrated into the platform is automatically exposed by OGC WPS. By relying on this, standard-based clients have the possibility to discover the method, be informed on the peculiarities (e.g. input and output), execute the method, and get back the results as well as any issue occurred during the execution.

3.2.7 User Interface / Interface to VREs

ENVRIplus has not specified data structures and operations concerning user interfaces. It is suggested (see below) that ENVRIplus should define an interface to VREs such that the RIs may choose which VRE offering to use.

The data analytics platform resulting from WP7 is actually equipped with a web-based graphical user interface. This GUI is based on portlets technology thus it is suitable to be exploited by any portlets engine. The major challenges are related with AAA, e.g. how to make it possible for VRE

users to be recognized by the instance of the platform operated by a certain RI goes well beyond the GUI domain.

3.2.8 Interface to e-Infrastructures

RIs wish to use e-Infrastructures in two ways: (a) where they outsource (part of) their data processing and management to e-Is; (b) where the end-user initiates a workflow which is executed using the RI's own infrastructure until such a time as additional resource is required at which time the workflow scales out/up in a seamless manner using e-Infrastructures for the additional resource required. Similarly, if a citizen scientist using a VRE initiates a request across RIs then each fragment of the workflow specific to each RI may, itself, need to scale out and use e-Infrastructures.

3.2.9 Trust, security, privacy

ENVRIplus does not have specific work on this topic although its influence is pervasive. The catalog (if in CERIF) can provide the parameters necessary to drive an AAAI, while this provides access and authorisation protection at workflow composition time it does not provide pervasive security to avoid – for example – code injection or unauthorised access during deployment and execution of the workflow. This would require each of the common and cross-cutting services either (a) to refer to the catalog for this information or (b) to carry the information through the API across each interface between services. It is likely that a system of certificates could be used by the ENV RIplus RIs although compatibility with the emerging recommendations of the EC-funded project AARC2[19] will be a constraint.

3.2.10 Curation and provenance

Curation and provenance depend on the common and cross-cutting services as specified in the RM implemented within each RI and possibly also centrally and the local RI catalog providing the requisite information. For interoperation, the information is required in the canonical rich metadata catalog. The curation information includes linkages to distributed duplicate copies, distributed partitions of datasets and relevant metadata for dataset recovery or reconstruction. The provenance information provides both lifetime information on digital object evolution and temporal snapshots of the state of one or more digital objects at a particular time. The W3C PROV set of recommendations are relevant and there is ongoing work on PROV/CERIF mappings. CKAN does not natively provide such information.

3.2.11 Agile, dynamic redeployment

The deployed workflow needs to be executed. If during execution (or even at deployment time using estimates) it is clear that SLA/QoS criteria will be violated an agile redeployment is required utilising different e-Is (whether within the RI or external). A new deployment of the workflow is calculated and deployed, monitored and the appropriate performance and provenance information recorded. This is (re-)used at later stages by the same or other users to optimise the choice of assets to be used and the proposed deployment.

3.3 ENVRIplus Capabilities Required

This section defines the provision across all RIs that should be done by ENVRIplus as a project. It may be deployed by some central organisation or at each RI.

[19] https://aarc-project.eu/

3.3.1 Monitoring and Logging

This section defines the capabilities: the ability to monitor and adapt the use of e-infrastructure during application/experiment execution; and the ability to record and analyse historical performance in order to make better deployment decisions in future.

The optimisation microservices being developed in Task 7.2 provide a number of functionalities for helping deploy applications on Cloud. Among those is the ability to select the optimal topology of available resources (e.g. VMs) from a pool to meet the deadline requirements of an application workflow, the ability to provision VMs across multiple sites (Clouds or Cloud domains) and the ability to deploy application components from a remote repository onto provisioned VMs automatically. These services are being extended to be able to self-diagnose when faults occur and re-provision accordingly; in conjunction with a suitable monitoring framework, e-infrastructure resources can be adapted to violations of QoS criteria. Standard interfaces (mainly REST-based) and the use of the microservice architecture ensure that these services can be built upon and extended in order to offer up additional deployment agility as necessary.

In addition, these services are being developed to rely not only on their immediate operating context, but also on information stored in separate knowledge bases. The use of suitable standards/vocabularies for describing infrastructure or applications (e.g. TOSCA or INDL) in principle allows substantive information about applications, QoS constraints, infrastructure and provenance to be stored and accessed by any number of services; should new variants of services be developed that make better use of this information to provide more sophisticated functionality, then further optimisation of workflow deployments will become available to with the RIs and RI users 'for free', insomuch as no additional effort at e-RI or VRE levels, e-Is and VREs. The platform is required to integrate the new services provided that they are sufficiently well-described and catalogued.

3.3.2 Common catalog of metadata and associated services

A rich canonical conceptual catalog is required for interoperation. Whether it is implemented as a physical catalog or not is discussed above. T8.2 has defined the catalog as using CKAN (as used in EUDAT) and CERIF.

3.3.3 Interface to e-Infrastructures

ENVRIplus needs to be able to deploy the workflow across one or more e-Infrastructures either within the RI(s) involved in the assets of the workflow or external to the RIs. In order to achieve this a clean interface needs to be defined. Various EC-funded projects have attempted this: PaaSage[20] has demonstrated a solution and the ongoing MELODIC[21] project should elaborate the work of PaaSage including especially dealing better with data locality and avoiding latency.

3.3.4 Interface to VREs

The interface to VREs is being defined in conjunction with the VRE4EIC project where UvA represents ENVRIplus as a partner.

[20] https://www.paasage.eu/
[21] http://cordis.europa.eu/project/rcn/206028_en.html

4 ACHIEVING THE REFERENCE ARCHITECTURE

4.1 Introduction

D5.4 proposed a plan:

4.1.1 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.

4.1.2 Development: M19-M30

The list of operations (common and cross-cutting services) 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. Reference model guided architecture model can be found in ANNEX1.

4.1.3 Deployment as prototype: M30-M33

The common and cross-cutting services will be deployed first as a prototype at two testbed RIs 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.

4.1.4 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. These steps are now considered in more detail for the various aspects of Theme2:

The focus of the optimisation task 7.2 is to complement the integrated VRE provided by Task 7.1 by providing a loosely-coupled set of microservices for improving data delivery and processing based on specific use-cases identified within the ENVRIplus project as being of interest to environmental scientists or RI engineers, which the RIs can then choose to make use of as they please. These services will be promoted to the RIs via the same channels as the other ENVRIplus common services, and will also be made available publicly via e.g. the EGI marketplace.

4.2 Development to Prototype

In order to make most efficient use of available resources (and people), the optimisation microservices will be implemented based on existing open source components adapted to capture the work already being done in the ENVRIplus technical use-cases. Specifically, for all use-cases that involve some technical investigation, it will be asked if that investigation can be generalised and a recipe derived which can be partly or wholly updated as an optimisation service. For example, if a data analytics framework is found to be useful for providing tailored data subscription services to a large group of users, then a service could be developed for automatically deploying such a framework on e-infrastructure in order to allow more RIs to more efficiently set up data subscription pipelines. A base set of services for generic selection and provisioning of e-infrastructure resources and then deploying arbitrary application components

on those resources have already been prototyped; the focus of the task is therefore simply to add more specialised deployment agents to support a range of generically useful application workflows.

4.3 Upgrading mechanism

The microservice model assumed for the optimisation services will be supported by standard metadata that can be harvested and catalogued by a VRE. The specification of such metadata falls to the semantic linking task (5.3) in ENVRIplus, with support from the cataloguing task (8.2). The use of such cataloguable metadata is to allow the easy specification of additional and alternative microservices that can be deployed on e-infrastructure, registered and then freely used to support application workflows without any additional integration effort needed at e-RI or VRE levels. In principle, a sufficiently clever VRE could automatically select the best services to provide the required functionality (select VMs, provision them, deploy components, configure specific data analytics frameworks, etc.) without user intervention, simply based on their requirements. So, the upgrading mechanism is kept as simple as possible.

4.4 Production

The optimisation services will be provided via public repositories for future use (e.g. the EGI Application Database) and will be open source to encourage uptake and maintenance by other developers who wish to build upon the specific functionalities provided. The scope of the ENVRIplus project however is to provide proofs-of-concept to the e-RI and e-I communities; thus, maintenance of the services cannot be guaranteed beyond the lifespan of the project.

4.5 Familiarisation

The data processing solution envisaged by D7.1 is offered as-a-Service. In fact, it is integrated into a Virtual Research Environment[22] operated by the D4Science.org infrastructure for each domain. ENVRIplus practitioners can go and play with it, i.e. execute the available analytics methods or integrate their own methods. This assists each RI in defining requirements more precisely. There is a real and identified need for each RI to be familiar with the techniques of other RIs to gain experience and ideas.

5 RECOMMENDATIONS

5.1 Introduction

The recommendations define the actions that should be pursued in the remainder of the ENVRIplus project and beyond in order to harmonise the information structures (particularly the metadata) of the RIs where appropriate and to facilitate interoperability.

5.2 Evolution by Existing RIs and evaluation

Existing RIs have a large investment already. Thus, the services developed and tested should be adopted within an evolutionary plan including also any upgrading of the RI catalog required to provide information necessary for the services.

The common and cross-cutting services need final specification (based on the RM) including engineering and technology viewpoints and either agreement on using existing services or development of new ENVRIplus services. This requires the RIs to cooperate under the

[22] https://services.d4science.org/group/envriplus

architectural umbrella. Once implemented and tested in one or a few RIs, the software should be rolled out and adopted by any new RI together with an appropriate metadata catalog to provide information required by the services.

The architectural model and services designs described above would be valid only when they are positively evaluated by ENVRIplus participating RIs. To support this, current work in WP9 endeavors the use case identification and service validation.

Use cases are collected from ENVRIplus community. In order to facilitate interoperability, the use cases need to involve at least 2 ENVRIplus RIs, with the purpose of testing ENVRIplus services for real-world scientific research. 14 use cases are identified[23]. One Agile group is formed for each use case. Such self-organized agile teams are typically led by scientific RIs, with members of Theme 2 technical experts. The agile activities are motivated by own interests, flexible and efficient, more importantly they enable the cross collaborations among different RIs, organizations and WPs. For example, the use case "IC_14 SOS & SSN ontology based Data Acquisition and Near Real Time Data Quality checking services" involves 8 RIs (EMSO, FIXO3, ANAEE, EPOS, SIOS, SeaDataNet, EuroArgo, and EGI) and over 10 organisations. It investigates standardization for data acquisition from observation sensors and quality control issues[24].

The agile teams test the design principles and services developed by Theme 2, and bring back evaluation results also new requirements from RIs. These will feed back to the design. The architectural model described above captures the snapshot of current requirements from ENVRIplus RIs. It would be needed to be extended in order to cope with the new requirements resulting from the evolutions of RIs.

The agile activities also help promote the ENVRIplus products to ENVRIplus RIs. During ENVRI weeks, sessions are organized for agile teams to give demos and presentations. They are specially requested to show how their approaches can be generic and benefit other RIs. Other RIs are welcome to join the agile investigations and apply the service approaches to own community usage scenarios.

5.3 Adoption by new RIs and evaluation

Design and service products of ENVRIplus can be introduced to new environmental research RIs beyond ENVRIplus. Such new RIs can be reached by ENVRIplus participating RIs who have established collaborations through previous and existing research projects. Theme 2 currently organises a serial set of workshops to visit individual RIs, part of the efforts facilitate the engagement with new communities. For example, a EuroArgo/SeaDataNet site-visiting workshop is host by Ifremer, 6-7 Apr 2017, Brest. The local hosting RI invited related marine research communities and projects including, SeaDataCloud, AtlantOS, and CMEMS (that are not ENVRIplus RIs). During the workshop, Theme 2 used ENVRI RM to analyze the requirements of these systems, also presented ENVRIplus services and approaches. In future workshops (for example, the upcoming EPOS visiting workshop in September), we plan to show developed demonstrators, and collect new use cases from new communities.

Many ENVRIplus Theme 2 developed services are generic that can be introduced to other domain areas. Using networking platforms such as, EGU, RDA, EGI and EUDAT conferences,

[23] Use Case list: https://confluence.egi.eu/display/EC/Use+Case+List
[24] IC_14 use case: https://confluence.egi.eu/pages/viewpage.action?pageId=7840619

ENVRIplus actively presents its project results, seeking collaboration opportunities beyond environmental science, and identifying use cases that can test the services results.

5.4 Maintaining Architectural Currency in the World

User requirements change as research advances making new demands on the RIs and requiring updating of the services and metadata catalog. This flexibility is essential to the success of ICT support within the RIs and such flexibility is built into the architectural design. Similarly new ICT offerings arise and – if appropriate – should be available to RIs and their communities for their advantage. Again the key aspect – given the defined interfaces of the architecture – is the catalog.

5.5 Customisation for specific RIs

Specific RIs may wish to add to the 'core services' of the architecture proposed with domain-specific extensions or additional services including where necessary additional metadata information in the catalog. However, the 'core' must remain unaltered to allow for interoperability and backward compatibility. The architecture is designed to accommodate this.

5.6 Positioning ENVRIplus in the European Open Science Cloud

As part of the European Commission Digital Single Market strategy[25], the European Open Science Cloud (EOSC) initiative was officially launched in April 2016 by the European Commission. EOSC promotes not only scientific excellence and data reuse but also job growth and increased competitiveness in Europe, and drives Europe-wide cost efficiencies in scientific infrastructures through the promotion of interoperability on an unprecedented scale.

The vision of ENVRIplus community is well aligned with EOSC. The community should consider how can better support the initiative and how to position ENVRIplus into the landscape of EOSC. Several ENVRI RIs are actively involved in EOSC discussions, as partners of the ongoing EOSC-pilot project and participants of the EOSC-hub proposal.

At the technical level, the architecture model and service design have anticipated the requirements to fit into the open architecture of EOSC. In particularly, there is a need for work to define more details to the interfaces to European major e-Infrastructures (such as EGI, EUDAT, GEANT, and PRACE). Usages of e-Infrastructure services and resources could be exploited, for example in the areas of:

- Compute-intensive processing
- Data-intensive storage and processing
- Integration of e-Infrastructures' services to provide new functionality
- Publishing and sharing data and services using ready-to-use e-Infrastructure services
- Shipping data and services across countries – especially useful for those RIs do not have the capacity
- Enabling Open Science by federating data and services for all disciplinary
- Replication and backup of data and services
- Host of web services
- Improving accessibility by providing nearer access sites
- Combinations of all above

[25] European Commission (2015), Open Science at the Competitiveness Council. http://ec.europa.eu/digital-agenda/en/news/open-science-competitiveness-council-28-29-may-2015

WP9 has already headed in this direction, working towards integration with the *European Open Science Cloud*. With the involvements of technical experts from EGI and EUDAT, several agile teams are testing services and technology from these e-Infrastructures. Successful experience will help connect ENVRIplus to EOSC easily.

6 CONCLUSIONS

The development of the ENVRIplus architecture is continuous, and this deliverable (D5.5) presents the current state of progress at this point in the project. Further work on the RM (Reference Model) will provide specifications based on engineering and technology viewpoints at which time a conventional architectural design document can be produced.

A key aspect of ENVRIplus is the reference architecture to be adopted by new RIs and towards which existing RIs should aim to align. Based on the ENVRI Reference Model, the architecture brings together all the aspects of the ICT (Theme 2) activities of ENVRIplus into a coherent framework to achieve those objectives. The architecture must sit within some constraints. ICT best practice is mandatory. Parallel initiatives in other ESFRI RIs and global consortia must be respected. Developments in e-Is (e-Infrastructures) provide opportunities for alternative deployment of applications. An appropriate interfacing mechanism between RIs and e-Is will provide for evolution of both RIs and e-Is while maintaining provision of service. Similarly, developments in VREs (Virtual Research Environments) offer improved opportunities for researchers (and other users) to access multiple RIs while appropriate interfacing will allow evolution of both RIs and VREs to sustain the consistent and integrated facilities built on the resources delivered by collaborating RIs. The degree of alignment with the architecture by RIs will improve their ability to present a research environment that supports research campaigns that need resources and capabilities from multiple RIs.

The major objective of ENVRIplus is to facilitate research in environmental science by encouraging movement towards a consistent and integrated view of data, processing and resources to meet emerging domain-specific and interoperation research needs. The adoption of common and cross-cutting ICT services by RIs (Research Infrastructures) reduces cost (re-use) and increases interoperation (standardisation).

The reference architecture is the basis for this achievement.

7 IMPACT ON THE PROJECT

The main motivation of ENVRIplus is to enable researchers to access, utilise and interoperate across individual and multiple RIs in the environmental domain. The provision of a reference architecture for adoption by new RIs and as an evolutionary target for existing RIs is fundamental.

Work documented in this document has identified the preferred architecture and approach. Recommendations including an adoption plan have been provided.

8 IMPACT ON STAKEHOLDERS

The adoption of the architecture will enable stakeholders to meet their requirements for interoperation across individual and 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 stakeholders outside the research domain (such as policymakers and citizens) may also benefit.

9 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)

[Candela et al. 2017] L. Candela, G. Coro, P. Pagano, M. Atkinson, R. Filgueira, D. Bailo, C.-F. Enell, M. Fiebig, F. Haslinger, M. Hellström, A. Vermeulen, H. Lankreijer, R. Huber, S. Joussaume, F. Guglielmo, V. Mendez (2017) Interoperable data processing for environmental RI projects: system design. ENVRIplus Project Deliverable D7.1

[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/

[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 et al. 2017 a] K. G Jeffery, M. Atkinson, Z. Zhao, Y. Chen, A. Nieva de la Hidalga, A. Hardisty, Y. Legre, L. Candela, D. Bailo, T. Loubrieu, B. Magagna (2017) A Development Plan for Common Operations and Cross-Cutting Services based on a Network of Data Managers and Developers. ENVRIplus Project Deliverable D5.4, January 2017

[Jeffery et al. 2017 b] K. G Jeffery, Z. Zhao, B. Magagna, A. Nieva de la Hidalga, L. Candela, C.-F. Enell, M. Hellström, A. Hardisty, C. Paxton, F. Toussaint (2017) Data Curation in System Level Sciences: Initial Design. ENVRIplus Project Deliverable D7.1, January 2017

[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

[Loubrieu et al. 2017] T. Loubrieu, F. Merceur, A. Chanzy, C. Pichot, D. Boulanger, F. André, M. Hellström, B. Magnana, A. Nieva De La Hidalga, Z. Zhao, P. Martin (2017) Interoperable cataloging and harmonization for environmental RI projects: system design. ENVRIplus Project Deliverable D8.3, January 2017

[Martin et al. 2016] P. Martin, Z. Zhao, M. Stocker, R. Huber, J. Heikkinen, A. Kallio (2016) Performance optimization for environmental RI projects: System Design. ENVRIplus Project Deliverable D7.3

[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

ANNEX1 Model Architecture based on the RM

This section illustrates how the RM is being used to develop the reference architecture.

10 Model Architecture

10.1 Intent

Provide suitable mechanisms to facilitate the automated/seamless assembly and integration of resources of RIs and service providers (e.g. e-Is and VREs).

10.2 Also known as

- Integration Architecture
- ENVRI Architecture

10.3 Motivation (Forces)

A set of RIs requires integrating their information [data and metadata] and services to enable interdisciplinary research or to support of long term research campaigns.

An independent RI wants to integrate its data products and services into an existing federation of RIs.

RIs need to take advantage of the shared services and assets provided by existing e-Infrastructures, such as PREACE, GÉANT, EUDAT and EGI.

10.4 Applicability

In the current landscape, many e-Is provide services which allow the cost-effective integration of RIs and RI federations. There is also a strong need for streamlining common operations that allow building automated and semi-automated data processing pipelines for processing and interpreting large quantities of data.

10.5 Structure

The integration of RIs and infrastructure providers can be aligned to the phases of the research data lifecycle. There are three specific communities which can be identified: Research Infrastructures (RIs), electronic infrastructures (e-Is) and virtual research environments (VREs).

- A **RI community** focuses on providing access to research data assets for different scientific research areas. The RI systems are designed to facilitate access to data and to data processing services.
- An **e-I community** focuses on providing reusable computing assets, such as storage, processing units, or communication networks for RI communities.
- A **VRE Community** focuses on integrating research assets from the e-I and RI communities and making them accessible to users.

The type of services and products provided by RI, e-I and VRE communities can be mapped to different phases of the research data lifecycle (acquisition, curation, publishing, processing, and use), as shown in Figure 6. RIs may focus on a specific subset of the data lifecycle phases, or provide end-to-end services covering the whole data lifecycle. e-I specialise on phases of the data lifecycle providing long term storage, and processing facilities. e-Is play a particularly

important role in the integration points located between lifecycle phases. VREs focus mainly on enabling the use and processing of data.

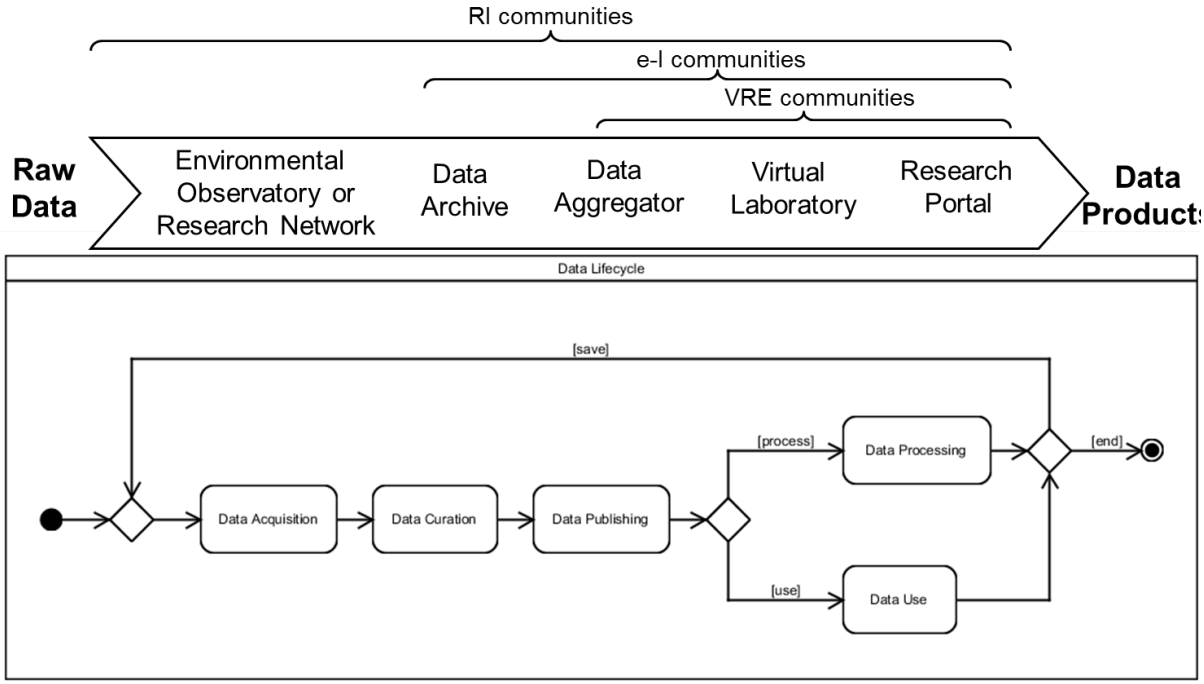

Figure 6 Alignment of communities with the research data lifecycle. Each box in this UML diagram represents a phase of the data lifecycle, the arrows indicate the flow of information, and the diamond shapes indicate fork/joint of flows.

10.6 Participants

The phases of the data lifecycle can be assigned to different roles. Figure 7 shows this alignment to describe a generic RI system which partitions the data lifecycle in five major subsystems (defined in the ENVRI RM). In turn, the data lifecycle phases themselves be further decomposed to allow a more detailed description of the roles and behaviours associated with each phase.

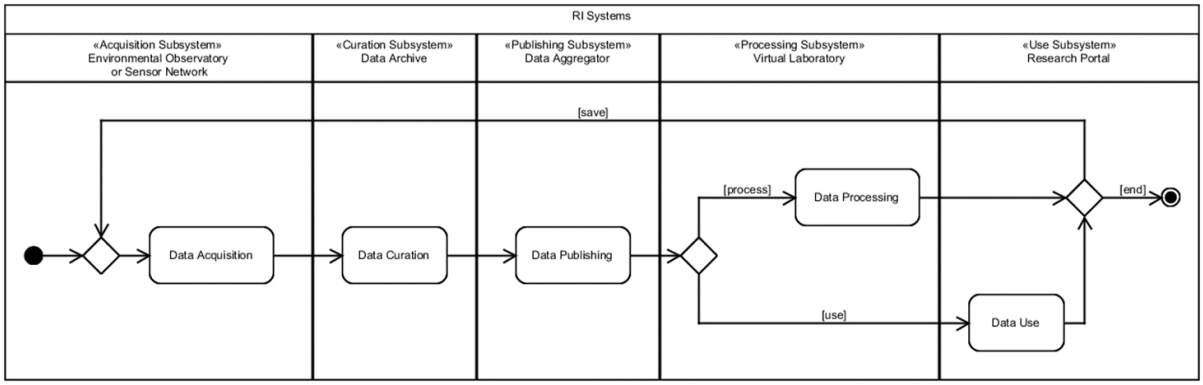

Figure 7 An RI system supporting the data lifecycle. On each swim-lane, the stereotypes (<<stereotype>>) indicate the type of subsystem according to the RM, the names assigned to the swim-lane provide common examples of naming these subsystems

Data Acquisition Subsystem (example: Environmental Observatory, Sensor Network): A subsystem that provides the assets and services that support collection of data from observations.

Data Curation Subsystem (example: Data Archive): A subsystem that provides the assets and services to curate, preserve and archive scientific data.

Data Publishing Subsystem (example: Data Aggregator, Data Manager, Publisher): A subsystem that provides the assets and services to assist data publication, discovery and access.

Data Processing Subsystem (example: Virtual Laboratory): A subsystem that supports the processing of published data, providing data processing services such as data mining, visualisation, and statistical analysis.

Data Use Subsystem (example: Research Portal, Community Portal): A subsystem that allows accessing research assets and services to the public.

10.7 Collaboration

The components mentioned previously can interact different ways, depending on the requirements and goals of individual RI. For instance, sensor networks can be implemented and integrated as RIs whose main aim is the acquisition of data. In other cases, RIs may delegate to data acquisition to service provider, and concentrate on curation or publishing of the data, while some RIs may have the capacity to support the entire data lifecycle. There are no restrictions to the number of behaviours and roles which an RI can delegate. Often, groups of RIs collaborate with each other in performing different tasks to implement a full data lifecycle. RIs can play any of the roles, VREs can play the roles of Research Portal and Virtual Laboratory, and e-Is play more specialised roles in support of specific activities such as data transfer, data processing, data identification, or data storage.

Interactions between subsystems allow composition by the RIs to support different tasks across the phases of the data lifecycle. The essential focus is on these areas of interaction between RIs (or between subsystems), named 'interaction points' on the ENVRI RM. An interaction point is defined as the pattern in which a set of services, protocols and devices collaborate to support the integration between two subsystems.

Definition of term

Interaction point: A pattern that describes the configuration of components, services and data which support the integration of two subsystems. An interaction point can appear within a single RI, between a pair of RIs, between an RI and a third-party, or between two RIs and a third-party.

Examination of the interaction of subsystems exposes a set of possible integration/collaboration scenarios. For each of these scenarios, the interfaces between the subsystems can be specified in terms of protocols that define the subsystems' behaviour in relation to one another.

The Interaction point engineering patterns can be:

1. Implemented jointly by agreement between the two RIs;
2. Offered by one RI to another as a service to which the other RI can bind and exploit;
3. Offered as a third-party service that RIs can bind to and exploit i.e., outsourcing.

ENVRIplus has identified six critical crosscutting architectural components : (1) identification and citation, (2) curation, (3) cataloguing, (4) processing, (5) provenance and (6) optimisation. Figure 8 shows the location of these components within the data lifecycle, according with the current version of the ENVRI RM. This location however does not prevent the use and/or integration of

these components within a specific phase, it only indicates which are the most likely phases to get the most use of such components, as the example on Figure 9

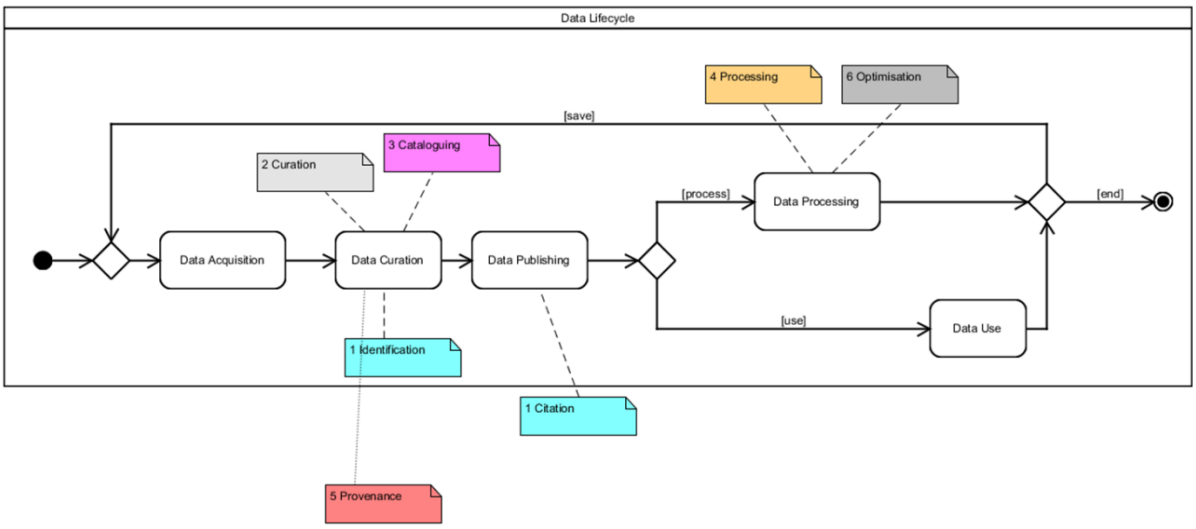

Figure 8 Location of Critical Crosscutting Components within the data lifecycle, according to the Current Version of the ENVRI RM.

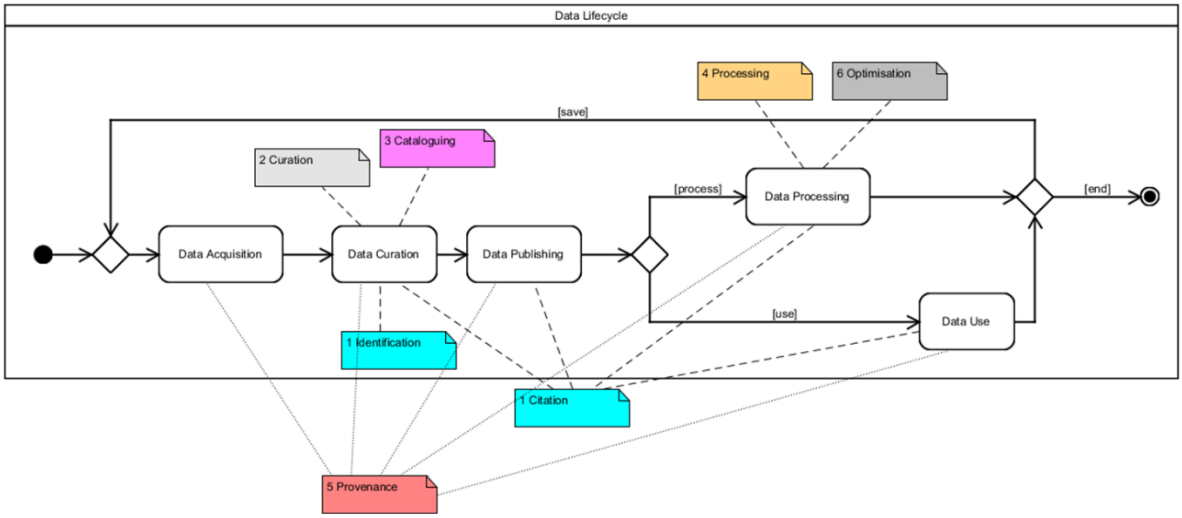

Figure 9 Extended use of citation and provenance components within the data lifecycle

10.8 Consequences

Research Infrastructure Systems developed using the ENVRI reference architecture will:

1. Achieve clear and progressive development of information models that meet the requirements of and are understood by the research communities supported.
2. Depend on a federation of resources and services to deliver, sustain, and support those information models, including conduct of scientific methods and processes required by research communities.
3. Allow providers of resources to meet those requirements with an appropriate balance of stability while supporting the incorporation of new requirements in line with scientific advances, while exploiting new technological opportunities.

10.9 Implementation

As shown in the Figure 8, the cross cutting components take part in the curation, publishing, and processing phases of the data lifecycle, similarly, the architectural components

supporting them are expected to be part of the corresponding subsystems as shown in Figure 10, Figure 11, and Figure 12.

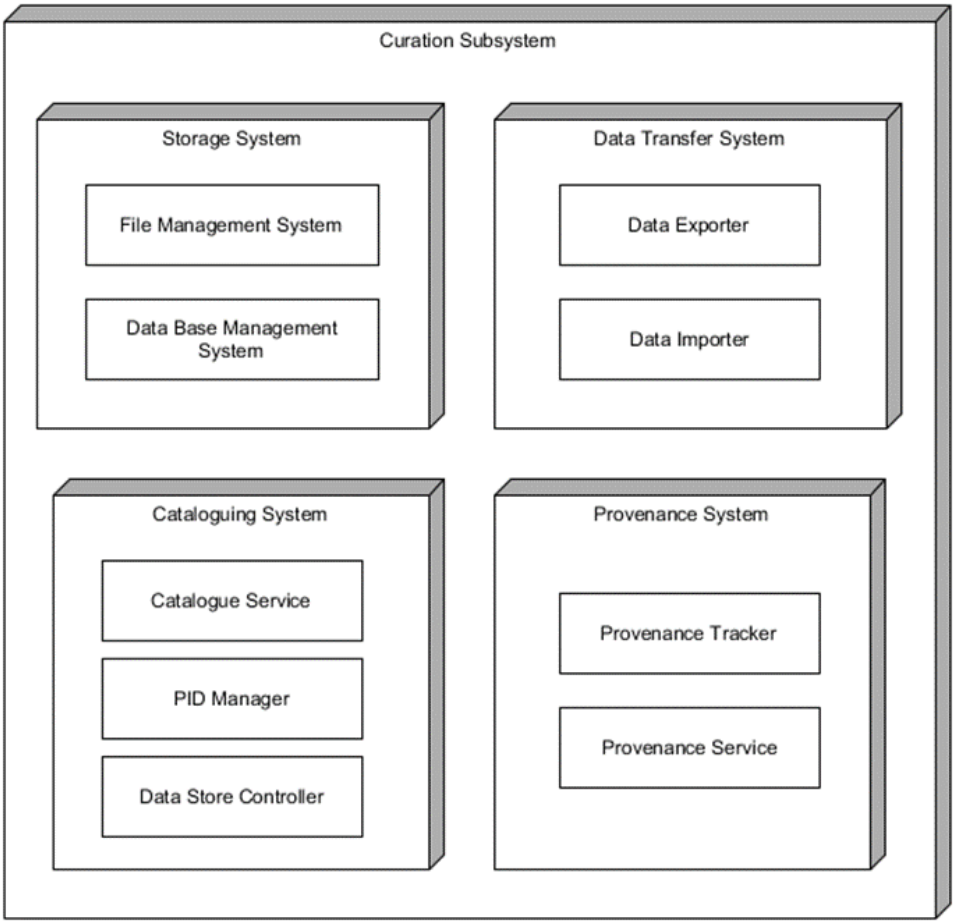

Figure 10Components of the Data Processing Subsystem

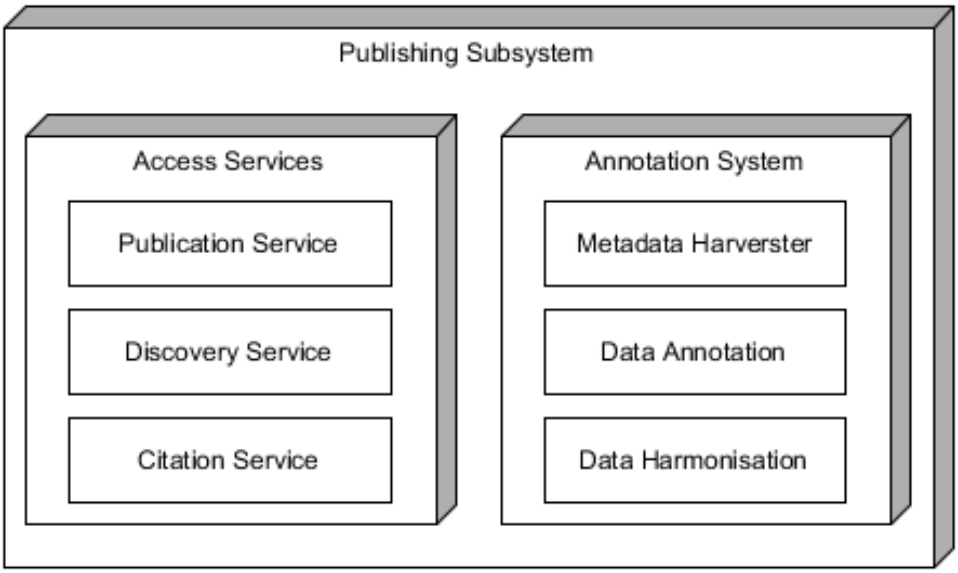

Figure 11 Components of the Data Publishing Subsystem

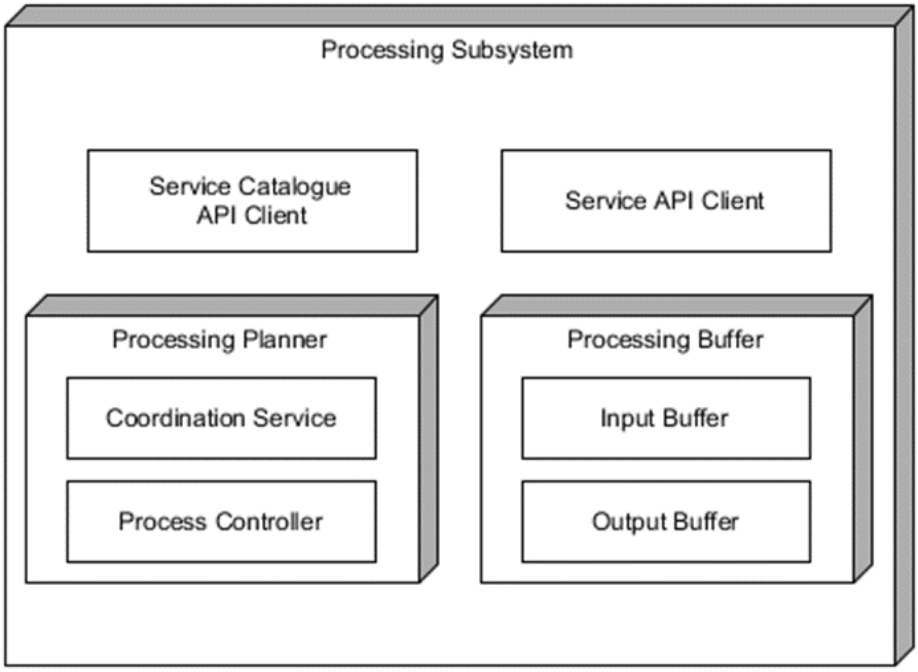

Figure 12 Components of the Data Processing Subsystem

The main component for enabling the integration of RIs is the development of a federated

10.10 Examples

The implementation of a PID Manager makes the use of a PID Registry transparent to the RI system, by handling the communication with the PID API. For instance, if the RI decides to use DOI for identification then the PID manager will need to be configured to communicate with three APIs (Figure 13). In another example, if the RI decides to use ePIC handles, the PID manager will need to be configured to communicate with one API (Figure 14).

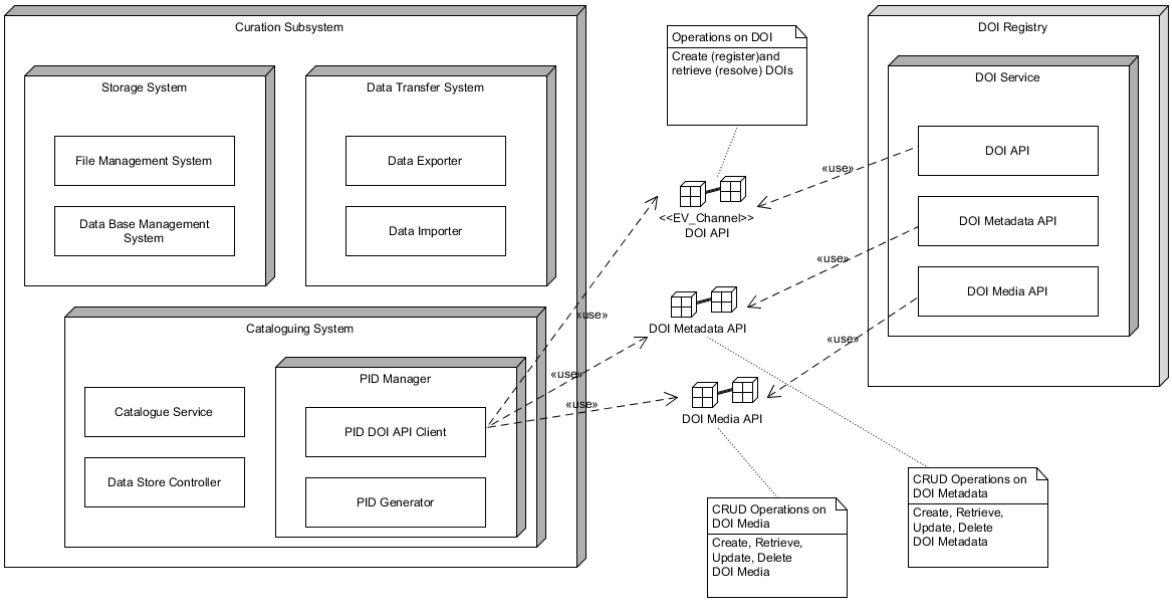

Figure 13 Integration of DOI to the Curation Subsystem

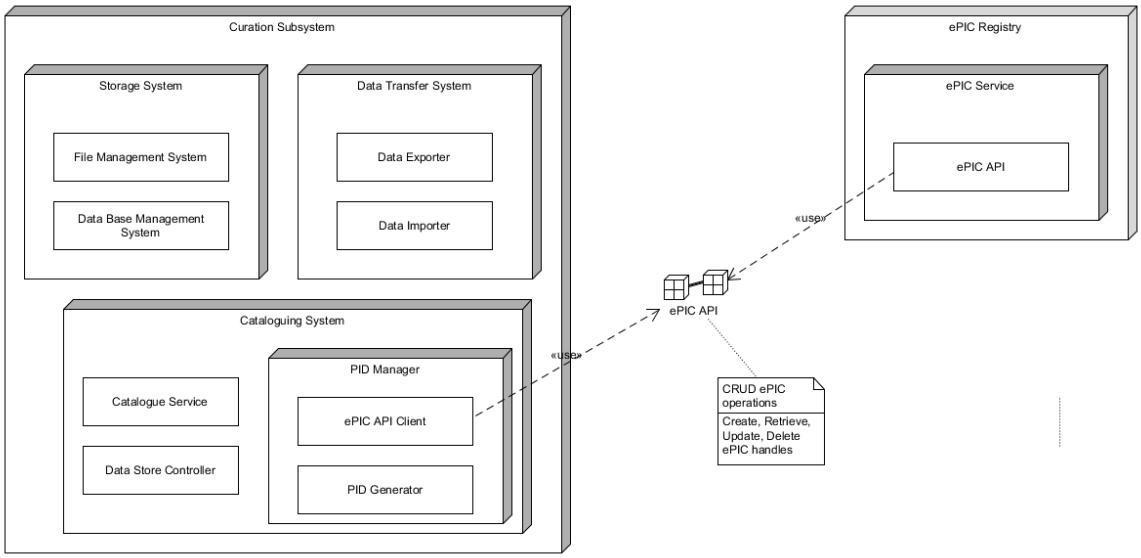

Figure 14 Integration of ePIC Handles to the Curation Subsystem

The implementation other components, such as extended cataloguing systems, can follow a similar pattern, using standardised APIs to make the integration of internal and external catalogues transparent.

10.11 Known Uses

DASSH uses DOIs. For this DASSH has stablished a minting process, DASSH acquired a prefix for the identifiers and generates the corresponding suffixes as needed.

EuroARGO also uses DOI, but the current method assigns a single DOI to different versions of a constantly evolving data set. Each new version receives a different suffix complement. In this case, the EuroARGO Data manager keeps track of the assigned suffixes for the existing versions of their public dataset.

LTER does not provide identifiers for the datasets they make public. LTER is an integrator of data from three types of providers: Expert, Advanced, and Basic. They only provide identification for Basic providers, however this is a service provided by EUDAT, as part of the B2Share service, which provides ePIC handles for the LTER datasets stored using the service.

SeaDataCloud Proposes an architecture based in at least seven distributed catalogues which different RIs can reuse to annotate data which is then shared through a planned VRE system

10.12 Relations

The RA organizes the composition of critical architectural components. The proposed RA architecture will help in the integration of the products of theme 2.